А. Леск

ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

ИЗДАТЕЛЬСТВО БИНОМ

ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

INTRODUCTION TO BIOINFORMATICS

Arthur M. Lesk
University of Cambridge

In nature's infinite book of secrecy
A little I can read.
– *Anthony and Cleopatra*

OXFORD
UNIVERSITY PRESS

А. Леск

ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Перевод с английского
под редакцией
доктора биол. наук, профессора А. А. Миронова
и доктора хим. наук, профессора В. К. Швядаса

Москва
БИНОМ. Лаборатория знаний
2009

УДК 577.3
ББК 28.071я73
Л50

Леск А.

Л50 Введение в биоинформатику / А. Леск ; пер. с англ. — М. : БИНОМ. Лаборатория знаний, 2009. — 318 с. : ил., [4] с. цв. вкл.

ISBN 978-5-94774-501-6 (русск.)
ISBN 0-19-925196-7 (англ.)

В учебном издании, написанном английским ученым — пионером в использовании приемов информатики в биологических исследованиях, ведущим преподавательскую работу в Кембриджском университете, изложены основы информационных технологий в применении к биологическим наукам. Приведены тексты некоторых программ, упражнения и задачи.

Для студентов университетов и научных работников.

УДК 577.3
ББК 28.071я73

Первый тираж издания осуществлен при финансовой поддержке Российского фонда фундаментальных исследований по проекту № 05-04-62033

Учебное издание

Леск Артур

ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Ведущий редактор канд. хим. наук *Т. И. Почкаева*
Редактор канд. хим. наук *Н. А. Аникин*
Художники *С. Инфантэ, Н. А. Новак*
Оригинал-макет подготовлен *О. Г. Лапко* в пакете LaTeX 2_ε

Подписано в печать 02.06.09. Формат $70\times100/16$.
Усл. печ. л. 26,325. Тираж 1000 экз. Заказ № 3964

Издательство «БИНОМ. Лаборатория знаний»
125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272, e-mail: binom@Lbz.ru, http://www.Lbz.ru

При участии ООО «ЭМПРЕЗА»

Отпечатано с готовых файлов заказчика в ОАО «ИПК «Ульяновский Дом печати». 432980, г. Ульяновск, ул. Гончарова, 14

ISBN 978-5-94774-501-6 (русск.)
ISBN 0-19-925196-7 (англ.)

Предисловие редакторов русского издания

Современная биоинформатика возникла в конце 70-х годов XX в. с появлением эффективных методов расшифровки последовательностей ДНК. Датой выделения биоинформатики в отдельную научную область можно считать 1980 г., когда вышел первый номер журнала Nucleic Acids Research, целиком посвященный компьютерным методам анализа последовательностей. Важной вехой в становлении и развитии биоинформатики стал проект по секвенированию генома человека. Именно с этого времени биоинформатика перестала быть только вспомогательным инструментом. При переходе к анализу полных геномов компьютерные методы информационного анализа стали играть настолько важную роль, что эти исследования вылились в самостоятельное научное направление. Геномы содержали большое количество генов, для многих из которых не было никаких экспериментальных данных. Неисследованные гены необходимо было найти в геноме и предсказать их функцию. Это должно было привести к отбору наиболее интересных участков генома с целью их последующего изучения и благодаря рациональному планированию экспериментальной работы.

Огромную роль в развитии технологии чтения генетической информации сыграло развитие компьютерной техники и вычислительных методов. Неудивительно, что интенсивное развитие биоинформатики совпало по времени с победным шествием компьютерных технологий. Это лишний раз подтвердило, что глубина научного знания чрезвычайно сильно зависит от технических возможностей. Другой важнейшей вехой в развитии биоинформатики стало возникновение и повсеместное распространение технологий Всемирной сети — Интернета. Большое число разнообразных баз данных и программных инструментов теперь доступны через Интернет. Биоинформатика, пожалуй, является одной из тех областей науки, которые в очень большой степени зависимы от Интернета и успешно развиваются благодаря Интернету. Следует подчеркнуть, что очень важное для биологии и медицины политическое решение об открытости сложнейшего биологического текста современности — генома человека — сделало эту информацию по-настоящему доступной для ученых всего мира лишь благодаря Интернету.

Сегодня мы находимся на начальном этапе использования генетической информации о живой материи, однако развитие все более эффективных методов расшифровки биологических текстов и разработка методов биоинформатики позволяет надеяться на серьезный прогресс в понимании строения, механизмов функционирования и регуляции живых систем. В результате становится возможным изучение и понимание все более сложных биологических систем, появляется возможность их системного исследования, установление эволюционных связей в живой природе, создание новых биотехнологий, лекарственных препаратов и методов лечения. Биоинформатика в существенной степени уже

способствовала развитию фундаментальных знаний в самых разных областях науки, а не только в биологии и медицине. На очереди, например, совершенно новые возможности в исследовании истории развития человечества и миграции народов. Благодаря развитию методов биоинформатики геногеография может отследить путь распространения людей по нашей планете, начиная от Адама, анализ индивидуальных особенностей генома представляет неограниченные возможности персонифицированной медицины и индивидуального развития каждого человека. Очевиден огромный прикладной потенциал биоинформатики, причем ее роль важна как для медицины, так и для самых различных технологий, где и используются элементы и принципы функционирования живых систем.

Уже сегодня, основываясь на установленных при помощи биоинформатики механизмах регуляции метаболизма микроорганизмов, созданы промышленные процессы получения аминокислот и других ценных веществ и материалов, выявлены молекулярные мишени для создания новых лекарственных препаратов и методов лечения. В то же время значительную часть генетической информации еще следует расшифровать и открыть новые возможности использования этого знания на благо человечества. Таким образом, биоинформатика — одно из наиболее актуальных направлений науки о человеке и окружающей среде.

Вниманию читателя предлагается перевод известного учебника Артура Леска «Введение в биоинформатику». Эта книга является именно введением в предмет, поскольку, как уже подчеркивалось, современная биоинформатика чрезвычайно быстро развивается, что широко освещается в многочисленных международных научных журналах и на регулярных конференциях. В книге значительное внимание уделяется развитию практических навыков и решению типовых биологически осмысленных задач. К сожалению, биоинформатика, и особенно то, что связано с Интернетом, развивается столь быстро, что в книге можно встретить уже несколько устаревшие сведения и неточности (о чем в тексте сделаны соответствующие замечания редакторов перевода). За рамками книги остался ряд важных задач биоинформатики — распознавание генов, предсказание вторичной структуры РНК, поиск мотивов, анализ экспрессии генов и др. Однако несмотря на эти издержки, книга послужит ценным пособием для самого широкого круга читателей. Она полезна не только для студентов и аспирантов, приступающих к изучению и использованию биоинформатики, но и для специалистов, работающих в различных областях наук о живой материи. Важно, что главы книги сопровождаются примерами решения реальных биологических задач и интересными упражнениями.

Первичный перевод книги был осуществлен студентами Факультета биоинженерии и биоинформатики МГУ имени М. В. Ломоносова при поддержке преподавателей кафедры английского языка. Большую помощь при переводе гл. 5 оказал канд. хим. наук Г. Г. Чилов. Существенный вклад в подготовке русской версии книги внесен редактором перевода канд. хим. наук Н. А. Аникиным. Некоторые затруднения при переводе и редактировании были связаны с необходимостью использования уже сложившейся терминологии, которая во многих случаях представляет собой «кальку» с английского (как и в ряде

других наук). Так, например, выражение «дизайн лекарств» отражает весьма сложный путь, который следует пройти при создании и оптимизации структур лекарственных препаратов, имеющих объектом воздействия конкретные молекулярные мишени (а не процесс предпродажного оформления приобретаемой субстанции). Однако мы рассчитываем, что даже в тех случаях, когда незнакомый новый термин может иметь разные значения, читатель легко установит его смысл и в дальнейшем сможет легко ориентироваться в русскоязычной научной литературе.

Надеемся, что данная книга вас заинтересует и принесет вам ощутимую пользу при путешествии в увлекательный мир биоинформатики.

А. А. Миронов,
В. К. Швядас

Предисловие

День 26 июня 2000 г. отмечен кардинальными переменами в биологии и медицине. В этот день премьер-министр Великобритании Тони Блэр и президент США Билл Клинтон провели объединенную спутниковую пресс-конференцию, чтобы объявить о завершении расшифровки генома человека. Заголовок на первой странице The New York Times гласил: «Генетический код человеческой жизни взломан учеными». Расшифровка последовательности длиной в 3 млрд пар оснований была кульминацией более чем 10-летней работы. В течение этих 10 лет цель всегда была на горизонте, вопрос состоял лишь в том, насколько быстро развиваются технологии и как гладко протекает финансирование. Некоторые исторические вехи на этом пути перечислены ниже.

В этой пресс-конференции наряду с политиками участвовали ученые. Среди них — Джон Салстон, директор Центра Сенгера в Великобритании, который стоял у истоков методов широкомасштабного секвенирования последовательностей. Он занимался проектом начиная с самых ранних стадий вплоть до нынешнего международного консорциума. Рядом с президентом США Клинтоном находились Фрэнсис Коллинз, директор Национального Института Исследований Генома Человека, представляющий общественный фонд, и Дж. Крэйг Вентер, президент и главный научный руководитель корпорации Celera Genomic Corporation — коммерческий сектор. Представляя двух последних людей, так и хочется хотя бы мысленно добавить: «В красном углу ... и в синем углу ...». Они, несомненно, были соперниками, а на поздних стадиях между ними шло напряженное соревнование, чуть ли не до драки.

Под соревнованием здесь скрывалось нечто большее, чем просто стремление быть первыми и получить кредит доверия в науке. Участников этого «забега» не проверяли на допинг, они должны были придумать НОВЫЕ ЛЕКАРСТВА. Клиническое применение научных результатов служило главным доводом для поддержки проекта «Генома человека». Как только было решено, что последовательности генов можно патентовать, что подразумевало огромные потенциальные выплаты за лекарства, — коммерческий сектор торопливо стал представлять для патентования наборы расшифрованных последовательностей. А академические группы в это же время стремились как можно быстрее опубликовать каждый новый фрагмент расшифрованной последовательности, чтобы не дать возможности корпорации Celera (или кому бы то ни было еще) обратиться за соответствующими патентами. Академические группы, дружно

выступавшие против Celera, были представлены коллективами научных лабораторий, в основном из Великобритании и США, в том числе:

Сенгеровский центр в Великобритании (The Sanger Centre in England)
Университет Вашингтона в Сент-Луисе, Миссури, США (Washington University in St. Louis, Missouri)
Уайтхед институт, Массачусетс, США (the Whitehead Institute at the Massachusetts)
Медицинский колледж Бэйлора, Хьюстон, США (Baylor College of Medicine in Houston, Texas)
Объединенный Институт Геномных Исследований при Ливерморской Национальной Лаборатории, США (The Joint Genome Institute at Lawrence Livermore National Laboratory in Livermore, California)
Центр Геномных Наук РИКЕН, Япония (RIKEN Genomic Sciences Center, now, Yokahama, Japan)

И «коммерсанты», и «академики» имели шанс обогатиться. Корпорация Celera имела исходный капитал, которым она могла бы рискнуть, и внешний источник финансирования — компанию PE Corporation, а, после того как проект стал общественным, еще и тех, кто хотел поучаствовать в сенсации. Центр Сенгера был обеспечен Британским комитетом медицинских исследований (the UK Medical Research Council) и компанией Wellcome Trust. Академические лаборатории США спонсировались Национальными институтами здоровья США и Министерством энергетики США.

Итак, 26 июня 2000 г. произошли кардинальные перемены в биологии и медицине.

Геном человека — это один из многих полностью секвенированных геномов. Собранные вместе, геномные последовательности организмов, разбросанных по ветвям гигантского древа жизни, дают нам ощущение, что в деталях жизнь на планете Земля имеет единую основу, о чем мы ранее могли только догадываться. Все это изменило наше представление о мире почти так же, как первые фотографии Земли из космоса помогли получить цельное изображение нашей планеты.

Расшифровка последовательностей генома человека стоит в одном ряду с Манхэттенским проектом, который разрабатывал атомное оружие во время Второй мировой войны, и с освоением космоса; полет людей на Луну — один из великих прорывов в технологических достижениях прошлого столетия. Эти проекты занимали и занимают лидирующие позиции как в фундаментальной науке, так и в широкомасштабном и дорогостоящем развитии прикладных исследований. Биология никогда не будет в таком выгодном положении, не будет она располагать и таким бюджетом. Скоро биологический проект по секвенированию и сопоставлению геномов двух млекопитающих вообще можно будет сравнить по бюджету только с неторопливой студенческой работой в университетском практикуме.

	Исторические даты в проекте генома человека
1953	Уотсон и Крик определили структуру ДНК.
1975	Ф. Сенгер и независимо А. Максам и У. Гилберт разработали методы секвенирования ДНК.
1977	Секвенирован бактериофаг ϕX-174: первый полный геном.
1980	Верховный суд США принял решение о том, что генетически модифицированные бактерии могут патентоваться. Это решение послужило основой для патентования генов.
1981	Секвенирована митохондриальная ДНК человека: 16 569 пар оснований.
1984	Секвенирован геном вируса Эпштейна-Бара: 172 281 пар оснований
1990	Запущен международный проект по геному человека с намеченным сроком 15 лет.
1991	Дж. К. Вентер и его коллеги идентифицировали активные гены через ярлыки экспрессированных последовательностей (EST) — первоначальную долю ДНК, комплементарную матричной РНК[1)]
1992	Завершена карта сцепления человеческого генома с низким разрешением.
1992	Начало проекта по секвенированию *Caenorhabditis elegans*.
1992	Компания Wellcome Trust и Британский Комитет Медицинских Исследований (United Kingdom Medical Research Council) основали The Sanger Centre под руководством Дж. Салстона для широкомасштабного секвенирования геномов.
1992	Дж. К. Вентер (J. C. Venter) организовал Институт геномных исследований (The Institute for Genome Research) с целью коммерческого использования секвенирования путем идентификации генов и разработки лекарств.
1995	TIGR получили первую последовательность генома бактерии, *Haemophilus influenzae*.
1996	Получена карта человеческого генома с высоким разрешением — маркеры, разделенные фрагментами длиной в ~600 000 пар оснований (по).
1996	Полностью определена последовательность генома дрожжей, первого генома эукариот.
Май 1998	Корпорация Celera провозгласили, что они в состоянии закончить расшифровку генома человека к 2001 г. Wellcome реагирует на это увеличением финансирования Sanger Centre.
1998	Совместное заявление о полной расшифровке генома человека.
1 сентября, 1999	Корпорация Celera объявила, что геном *Drosophila melanogaster* секвенирован; опубликовано весной 2000 г.

1999	Проект по человеческому геному объявил своею целью: расшифровать геном человека к 2001 г. (90% последовательностей генов с более чем 95% точностью).
1 декабря, 1999	Опубликована первая полная последовательность одной из хромосом.
26 июня, 2000	Совместное заявление о полной расшифровке генома человека.
2003	Пятидесятая годовщина открытия структуры ДНК. Намеченная дата завершения общественным консорциумом секвенирования генома человека с высоким разрешением[2)]

[1)] На самом деле в рамках проекта «Геном человека» был запущен производный проект определения последовательностей EST. — *Прим. ред.*

[2)] Позже было сделано еще несколько заявлений о завершении секвенирования генома человека. Дело в том, что первое заявление гласило, что определено 98% генома. Более поздние заявления говорили о том, что произведено уточнение последовательности генома и закрыты некоторые бреши в последовательности, хотя эти детали для широкой публики не столь важны. — *Прим. ред.*

Геном человека содержит фундаментальную информацию, и компьютеры необходимы как на этапе определения последовательности, так и этапе ее анализа и применения в биологии и медицине. Компьютеры необходимы не только на самой первой стадии обработки и хранения информации, но также при применении сложных математических методов, позволяющих получить биологически важные результаты. Объединение биологии и информатики породило новую научную область — биоинформатику.

Биоинформатика сегодня — прикладная наука. Мы используем компьютерные программы, чтобы делать различные выводы на основе анализа архива данных современной молекулярной биологии, определять взаимосвязь между ними, а также выдвигать полезные и интересные гипотезы.

Эта книга предназначена для студентов и ученых. Ведь им необходимо знать, как получить доступ к архивам данных геномов и белков, к инструментам, которые были придуманы для работы с этими архивами, и набору вопросов, на которые эти данные и инструменты могут ответить. В действительности, существует много источников подобной информации. Интернет-ресурсы, связанные с биоинформатикой, разбросаны по всей сети. Проблема состоит в выборе существенного материала, а также в изложении биоинформатики на доступном уровне.

Предполагается, что читатель уже имеет некоторые знания по современной молекулярной биологии, а также легко пользуется компьютером. Цель этой книги — расширить и усовершенствовать эти знания. Это удобный учебник для студентов старших курсов и аспирантов. Здесь приведено много упражнений,

а также даны ссылки на полезные Web-сайты и список рекомендованной литературы. (Надо иметь в виду, что область настолько быстро развивается, что значительная часть информации может устареть. — *Прим. ред.*)

Задачи проверяют и закрепляют понимание, предоставляют возможности попрактиковаться и объясняют дополнительные вопросы. В конце глав помещены три типа задач. Решения коротких упражнений непосредственно основаны на приведенном материале. Задачи также не требуют изучения дополнительных литературных источников, надо просто осмыслить и порассуждать, а в некоторых случаях следует сделать расчеты. Для Интернет-заданий (Web-lems) нужен доступ к всемирной компьютерной сети (WWW). Эти задания созданы для того, чтобы читатель получил практические навыки, необходимые при дальнейшем обучении и в научных исследованиях.

Написание этой книги стало возможным благодаря тому, что всемирная компьютерная сеть сильно облегчила доступ к архивам данных и к программам для работы с ними. Раньше было необходимо устанавливать программы в своей компьютерной сети, и там же проводить вычисления. Конечно же, это означало, что все зависело от уровня компьютерного оснащения. Сейчас стало возможным пользоваться Web-сайтами Интернета. Web-сайт данной книги облегчит навигацию по глобальной сети. Для того чтобы быть уверенным, что читатели смогут свободно перенести занятия из книги в компьютерную сеть, мы старались избегать описания коммерческих программных пакетов и ссылок на них, хотя многие из них имеют очень высокое качество.

Серьезной проблемой компьютерной сети является ее быстрое изменение. Web-сайты появляются и исчезают, оставляя цепочки неработающих ссылок. Существует так много сайтов, что трудно даже просто найти несколько ключевых и стабильных, — проблема, а еще надо, чтобы они содержали современные данные и ссылки. Я предложил несколько таких сайтов, но есть много других, которые не хуже. Проблема заключается не в том, чтобы создать длинный список полезных сайтов (это относительно легко и было сделано много раз), а в том, чтобы создать *короткий* список, — вот, что значительно сложнее!

В этой книге изложены некоторые основы программирования, опирающиеся на широко используемый язык PERL. Примеры представленных простых программ на PERL'е основаны на биологических задачах. Много заданий для PERL'а поставлено в качестве задач и упражнений в конце глав.

Чем может заняться читатель для своего дальнейшего образования? Я писал свою книгу как дополнение (говоря на выбранном здесь языке, как **приквел**) к Introduction to Protein Architecture: The Structural Biology of Proteins (Oxford University Press, 2001), — книгу, которую я, безусловно, рекомендую читателю. Ориентация других книг по анализу последовательностей варьируется от биологии до программирования. Читатель сможет удовлетворить здесь собственные интересы и при достаточных знаниях работать в области биоинформатики.

Я благодарен многим коллегам за обсуждение и советы в ходе подготовки книги, а также университетам Упсала, Умеа, Рима 'Tor Vergata' и Кембриджа за предоставление некоторых материалов.

Я благодарю S. Aparicio, T. Baglin, D Baker, A. Bench, M. Brand, G. Bricogne, R. W. Carrell, C. Chotia, D. Crowther, T. Dafforn, R. Foley, A. Friday, M. B. Gerstein, T. Gibson, T. J. Hubbard, J. Irving, J. Karn, K. Karplus, B. Kieffer, E. V. Koonin, M. Krichevsky, P. Lawrence, D. Liberles, A. Lister, E. L. Lesk, M. E. Lesk, V. E. Lesk, V. I. Lesk, L. Lo Conte, D. A. Lomas, J. Magré, C. Mitchell, J. Moult, E. Nacheva, H. Parfey, A Pastore, D. Penny, F. W. Roberts, G. D. Rose, B. Rost, J. Sulton, M. Segal, E. L. Sonnhammer, R. Srinivasan, R Staden, G. H. Thomas, A. Tramontano, A. A. Travers, A. Venkitaraman, G. Vriend, J. C. Whisstock, S. H. White, C. Wu и M. Zuker за советы и критическое прочтение моих материалов.

Я благодарю персонал Oxford University Press за их профессионализм и терпение.

Кембридж A. M. L.
Январь 2002

Я хотел, чтобы читатели моей книги узнали следующее:

- Понимание причин, по которым стал доступен очень большой объем детальной информации о человеке (т. е. о нас самих) и других видах живых существ.
- Области применения биоинформатики в молекулярной биологии, клинической медицине, фармакологии, биотехнологии, сельском хозяйстве, судебной медицине, антропологии и других дисциплинах.
- Полезные знания об информационных технологиях, с помощью которых мы через всемирную компьютерную сеть получаем доступ к данным и методам их анализа.
- Понимание роли компьютера и программирования в исследованиях и применении этих данных.
- Уверенные базовые знания в поиске информации, вычислении, исходя из найденных данных, а также возможность улучшать эти знания, используя собственную «полевую» работу в сети.
- Оптимистичные прогнозы, что биоинформатические данные и методы далеко продвинут нас в понимании жизни, приведут к улучшению здоровья людей и других живых существ.

Структура книги

- Глава 1 является вводной; она представляет нам главных «действующих лиц»: последовательности и структуры ДНК и белков, геномы и протеомы, базы данных и поиск информации, всемирная компьютерная сеть и программное обеспечение.

- Глава 2 знакомит с основными свойствами отдельных геномов, в том числе генома человека, и взаимосвязями между ними.

- Глава 3 дает основные навыки работы с всемирной паутиной применительно к биоинформатике. Описаны архивные банки данных, демонстрируются примеры работы, включая поиск информации в некоторых важных базах данных по молекулярной биологии.

- Глава 4 рассматривает взаимоотношения между последовательностями — выравнивания и филогенетические деревья. Эти методы лежат в основе некоторых важных компьютерных задач биоинформатики: определение дальних родственных связей, понимание отношений между геномами разных организмов и прослеживание эволюции на уровне видов и молекул.

- Глава 5 переводит нас в трехмерное пространство при рассмотрении белковых структур и их укладки. Надо понимать, что последовательность и структура — полноценные партнеры, и в биоинформатике разрабатываются методы быстрого перемещения между ними. Детализированное понимание белковых структур важно как для определения механизма их действия, так и для клинического и фармакологического применения.

1 Введение

Биология традиционно описательная, а не аналитическая наука. Несмотря на то что последние успехи науки не изменили это основное направление, радикально изменилась сущность данных. Можно сказать, что до последнего времени все биологические наблюдения носили в основном случайный характер, правда, с различным уровнем точности, некоторые действительно с очень хорошим качеством. Однако данные последнего поколения исследований стали не только количественными и более точными, но, как в случае нуклеотидных и аминокислотных последовательностей, они стали *дискретными*. Расшифровать геномную последовательность индивидуального организма или клона стало возможным не только полностью, но и, что принципиально, *точно*. Ошибки эксперимента не могут никогда быть полностью исключены, но для современного секвенирования генома они чрезвычайно низки.

Это не означает, что биология стала аналитической наукой. Жизнь действительно подчиняется законам физики и химии, но она слишком сложна

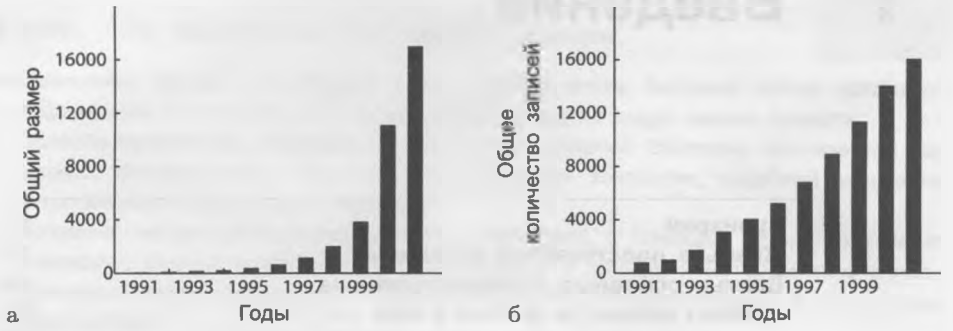


Рис. 1.1. (а) Рост GenBank, банка генетических последовательностей Национального Центра Биотехнологической Информации (NCBI) США. (б) Рост Банка Белковых Структур, архива трехмерных структур биологических макромолекул. (На 11 сентября 2007 г. в базе PDB было депонировано более 42 000 структур. — *Прим. ред.*)

и зависима от цепи исторических случайностей, чтобы сейчас мы могли детально объяснить ее свойства, исходя из основных принципов.

Вторая очевидная особенность биоинформатических данных — это их огромное количество. Сейчас банки данных нуклеотидных последовательностей содержат 16 млрд нуклеотидных пар (пн) оснований[1)]. Если мы возьмем в качестве единицы измерения размер генома человека (HUman Genome Equivalents, HUGE), то этот объем информации эквивалентен 5 HUGE. Для сравнения, один HUGE соответствует числу букв во всех номерах *New York Times* за шесть лет. База данных макромолекулярных структур содержит 16 000 записей, каждая из которых является полным описанием координат $\sim$ 400 аминокислотных остатков белка в трехмерном пространстве. Огромны не только размеры отдельных банков данных, но и темпы их увеличения. На рис. 1.1 показан рост GenBank (архива нуклеотидных последовательностей) и банка белковых структур PDB (архива макромолекулярных структур). Экстраполяция при этом рискованна.

Такое количество и качество данных поощряет стремление ученых к следующим целям:

- Увидеть картину мира живых существ четко и целиком, т. е. понять *интегрирующие* аспекты биологии организмов, рассматриваемых как согласованные комплексные системы.
- Связать между собой последовательность, трехмерную структуру, взаимодействия и функции отдельных белков, нуклеиновых кислот и их комплексов.
- Использовать данные о современных организмах как основу для изучения организмов во времени — назад в прошлое, чтобы вычислить последова-

[1)] В русскоязычной литературе используется несколько эквивалентных обозначений — пар нуклеотидов (пн), пар оснований (по), нуклеотидных пар (нп), а в англоязычной — bp.

тельность событий в эволюционной истории, и вперед к великой научно обоснованной модификации биологических систем.
- Способствовать применению этих знаний в медицине, сельском хозяйстве и других областях.

Сценарий

Чтобы проще понять роль вычислений в молекулярной биологии, вообразим себе кризис (в будущем?!), спровоцированный появлением нового биологического вируса. Этот вирус вызывает эпидемии смертельного заболевания как среди людей, так и среди животных. Ученые в лаборатории выделят его генетический материал (молекулу нуклеиновой кислоты, представляющую собой длинный полимер, состоящий из четырех видов оснований) и определят ее последовательность. Далее к работе приступят компьютерные программы.

Скрининг этого нового генома по базам данных всего известного генетического материала позволит охарактеризовать вирус и выявить его родство с ранее изученными вирусами [10][1)]. Анализ будет продолжен с целью выработки антивирусной терапии. Вирусы содержат молекулы белков, а это подходящие мишени для лекарств, которые будут действовать на структуру и функции вируса. Белки, как и нуклеиновые кислоты, являются линейными полимерами, их последовательность включает в себя 20 различных аминокислотных остатков. Из последовательности ДНК вируса компьютерные программы вычислят аминокислотные последовательности одного или нескольких вирусных белков, критически важных для репликации или сборки вируса [01].

Из аминокислотных последовательностей другие программы вычислят структуры этих белков, следуя тому базовому принципу, что аминокислотная последовательность белка определяет его трехмерную структуру, а тем самым и его функцию. В первую очередь будет проведен скрининг баз данных для поиска родственных белков известной структуры. Если такие белки будут найдены, то проблема предсказания структуры будет сведена до «дифференциальной формы»: предсказания действия изменений в последовательности на структуру. Структуры мишеней будут предсказаны с помощью метода, известного как гомологичное моделирование [25]. Если ни одного родственного белка с известной структурой не будет найдено, а вирусный белок окажется совершенно новым, то предсказание структуры будет сделано *ab initio(с самого начала)* [55]. Последняя ситуация будет возникать все реже, по мере того как растет и пополняется банк данных известных структур и увеличиваются наши возможности устанавливать отдаленное родство организмов.

Знание структуры вирусных белков сделает возможным разработку лекарственных препаратов. На поверхности белков есть участки(сайты), ответственные за их функции, которые чувствительны к блокированию. Будет найдена или сделана маленькая молекула, комплементарная такому участку

[1)] Цифры в квадратных скобках обозначают степень сложности задач. Ниже в тексте дано объяснение. — *Прим. ред.*

(сайту) по структуре и свойствам, которая будет работать как антивирусный препарат [50]. Альтернативный вариант — создать и синтезировать одно или несколько антител для нейтрализации вируса [50].

Сценарий основан на четко установленных принципах, и я не сомневаюсь, что однажды он будет воплощен в жизнь, как описано выше. Многие проблемы еще не решены, и это одна из причин, по которой этот сценарий не может быть использован сейчас против СПИДа. Другая причина в том, что вирусы знают, как себя защитить. Специалисты, читающие эту книгу, могли заметить, что номера в квадратных скобках относятся не к цитатам из литературы, а соответствуют приему, использованному Д. Е. Кнутом в его классических книгах «Искусство программирования» в индексировании сложности проблемы! Номера ниже 30 относятся к проблемам с уже существующим решением, большие номера обозначают темы, исследуемые в настоящее время.

Наконец, следует признать, что чисто экспериментальные подходы к проблеме создания антивирусных препаратов могут еще много лет оставаться успешнее теоретических.

Жизнь в пространстве и времени

Трудно определить, что такое жизнь. Можно пользоваться старым определением, но в век компьютерных технологий оно, возможно, должно быть изменено. Сейчас можно попробовать следующее: биологический организм — это естественно возникающее, самовоспроизводящееся устройство, которое действует за счет управляемого превращения вещества, энергии и информации.

С наиболее общей точки зрения, жизнь на Земле — это комплексная отдельно существующая система, распространенная в пространстве и во времени. Большое значение имеет то обстоятельство, что во многих случаях система состоит из отдельных организмов, каждый с определенной продолжительностью жизни и, в большинстве случаев, с уникальными чертами.

Пространственно, начиная издалека и постепенно проникая все глубже, можно различить внутри биосферы локальные *экосистемы*, являющиеся стабильными пока не меняются окружающие условия или пока они не нарушаются извне. Каждая экосистема содержит определенный набор *видов*, эволюционирующих по законам Дарвина или благодаря дрейфу генов. Генерация вариантов может возникать или в результате естественных мутаций, или в результате рекомбинации генов при половом размножении, или при прямом переносе генов. Каждый вид состоит из *организмов*, осуществляющих индивидуальные или даже независимые действия. Организмы состоят из *клеток*. Каждая клетка является маленькой локализованной экосистемой, не изолированной от своего окружения, но взаимодействующей с ним специфическими контролируемыми путями. Сами эукариотические клетки также имеют сложное строение, включающее ядро, другие внутриклеточные органеллы и цитоскелет. И наконец, мы доходим до уровня молекул.

Жизнь является протяженной не только в пространстве, но и во времени. Сегодня мы наблюдаем лишь краткий миг одного периода истории жизни,

который простирается назад во времени как минимум на 3,5 млрд лет. Теория естественного отбора была очень успешна в объяснении процесса развития жизни. Исторические случайности, однако, играют слишком большую роль в определении направленности события, чтобы сделать возможным детальное предсказание. ДНК ископаемых организмов также не дает достаточного доступа к историческим записям на молекулярном уровне. Вместо этого мы можем попытаться установить прошлое по современным геномам. Верховный судья США Феликс Франкфуртер написал «... американская конституция — не только документ, это исторический поток». Это справедливо и для геномов, которые содержат информацию о собственном развитии.

Догмы: основные и второстепенные[1)]

В информационном архиве каждого организма содержится детальный план будущего развития и функционирования этого индивидуума, представленный генетическим материалом (ДНК) или у некоторых вирусов— РНК. Молекулы ДНК — длинные, линейные, цепочечные молекулы, несущие сообщения в четырехбуквенном алфавите (см. врезку на с. 20). Даже у микроорганизмов сообщение длинное, обычно состоит из 10^6 букв. В структуре ДНК полностью оговорены механизмы репликации и переноса информации с гена на белок. Двойная спираль и ее внутренний принцип комплементарности, необходимый для точной репликации, хорошо известны (см. цветную иллюстрацию I). Почти безупречная репликация необходима для стабильности наследственности. Небольшая неточность в репликации, как и механизм импорта инородного генетического материала, также необходима, иначе организмы, не имеющие полового размножения, не могли бы эволюционировать.

Цепи двойной спирали антипараллельны. Концы носят названия $3'$ и $5'$ по позициям в дезоксирибозном кольце. ДНК считывается всегда в направлении от $5'$ к $3'$.

Генетическая информация воплощается через синтез РНК и белков. Белки — это молекулы, отвечающие за жизнедеятельность большинства структур организма. Наши волосы, мышцы, пищеварительная система, рецепторы и антитела — все это белки. Как и нуклеиновые кислоты, белки — длинные линейные цепочечные молекулы. Генетический код — это шифр: триплеты букв из последовательности ДНК обозначают аминокислоты. В участках ДНК зашифрованы аминокислотные последовательности белков. Обычно белки состоят из 200–400 аминокислот, что требует 600–1200 нуклеотидов ДНК для их кодирования. Синтез молекул РНК, например РНК — компонентов рибосом, также определяется последовательностью в ДНК. Однако в большинстве организмов не вся ДНК кодирует РНК или белки. Некоторые участки последовательности ДНК существуют для механизмов управления, а большая часть

[1)] Этот раздел носит вводный, и, стало быть, весьма поверхностный характер. На самом деле за каждым утверждением, сделанным здесь, стоит множество нюансов, изложение которых является предметом не одного учебника.

генома, похоже, является «ненужной» (это может означать, что нам просто пока ничего не известно о ее функции).

Четыре природных нуклеотида в структуре ДНК (РНК)

a аденин g гуанин c цитозин t тимин (u урацил)

Двадцать природных аминокислот в структуре белков

Неполярные аминокислоты

G глицин	A аланин	P пролин	V валин
I изолейцин	L лейцин	F фенилаланин	M метионин

Полярные аминокислоты

S серин	C цистеин	T треонин	N аспарагин
Q глутамин	H гистидин	Y тирозин	W триптофан

Заряженные аминокислоты

D аспарагиновая кислота	E глутаминовая кислота	K лизин	R аргинин

Другие классификации аминокислот также могут быть полезными. Например, гистидин, фенилаланин, тирозин и триптофан являются ароматическими аминокислотами, и установлено, что они играют специфическую роль в мембранных белках.

Названия аминокислот часто сокращаются до первых трех букв (например Gly для глицина), кроме изолейцина, аспарагина, глутамина и триптофана, которые сокращаются до Ile, Asn, Gln и Trp соответственно. Редко встречающаяся аминокислота селеноцистеин имеет трехбуквенное сокращение Sec и однобуквенный код U.

Принято обозначать нуклеотиды строчной буквой (не всегда), а аминокислоты — прописной буквой (всегда). Так, atg = аденин-тимин-гуанин, а ATG = аланин-треонин-глицин.

Молекулы ДНК, содержащие стандартные четыре буквы, сходны по химическому строению, а сама структура ДНК в первом приближении однородна. Белкам, наоборот, свойственно большое разнообразие трехмерных конформаций. Эти конформации необходимы белкам для выполнения их разнообразной структурной и функциональной роли.

Последовательность аминокислот в белке определяет его трехмерную структуру. Для каждой природной аминокислотной последовательности существует уникальное стабильное нативное состояние, в которое эта последовательность спонтанно переходит в нормальных условиях. Если очищенный белок нагреть или каким-нибудь другим образом перевести в условия, которые сильно отличаются от естественных физиологических условий организма, то он «разворачивается», образуя беспорядочную биологически неактивную

Стандартный генетический код

ttt	Phe	tct	Ser	tat	Tyr	tgt	Cys
ttc	Phe	tcc	Ser	tac	Tyr	tgc	Cys
tta	Leu	tca	Ser	taa	STOP	tga	STOP
ttg	Leu	tcg	Ser	tag	STOP	tgg	Trp
ctt	Leu	cct	Pro	cat	His	cgt	Arg
ctc	Leu	ccc	Pro	cac	His	cgc	Arg
cta	Leu	cca	Pro	caa	Gln	cga	Arg
ctg	Leu	ccg	Pro	caa	Gln	cga	Arg
att	Ile	act	Thr	aat	Asn	agt	Ser
atc	Ile	acc	Thr	aac	Asn	agc	Ser
ata	Ile	aca	Thr	aaa	Lys	aga	Arg
atg	Met	acg	Thr	aag	Lys	agg	Arg
gtt	Val	gct	Ala	gat	Asp	ggt	Gly
gtc	Val	gcc	Ala	gac	Asp	ggc	Gly
gta	Val	gca	Ala	gaa	Glu	gga	Gly
gtg	Val	gcg	Ala	gag	Glu	ggg	Gly

Альтернативные генетические коды встречаются, например, в органеллах — хлоропластах и митохондриях.

структуру (вот почему в нашем организме существуют механизмы для поддержания относительно постоянных внутренних условий). При восстановлении нормальных условий пептидные молекулы в целом приобретают снова свою нативную структуру, которая не отличима от нативной структуры природного происхождения.

Спонтанное сворачивание белков (фолдинг) с целью формирования их нативной структуры является точкой, в которой Природа совершает гигантский прыжок от одномерных генетических и пептидных последовательностей к трехмерному миру, в котором мы все живем. Однако парадокс: трансляцию последовательностей ДНК в последовательности аминокислот очень легко описать логически — она определяется генетическим кодом. Сворачивание полипептидной цепи в точно определенную трехмерную структуру очень трудно описать логически. Для осуществления же трансляции необходимы исключительно сложный механизм работы рибосомы, транспортные рибонуклеиновые кислоты (тРНК) и связанные с ними молекулы, а сворачивание белков происходит спонтанно.

Функции белков зависят от приобретения ими нативной трехмерной структуры. Например, нативная структура фермента может иметь на своей поверхности полость, которая связывает одну маленькую молекулу и помещает ее

рядом с аминокислотными остатками каталитического центра. Таким образом, мы имеем следующую парадигму:

- Последовательность нуклеотидов ДНК определяет последовательность аминокислот белка.
- Последовательность аминокислот определяет структуру белка.
- Структура белка определяет его функцию.

В большинстве своем биоинформатика как раз и занимается анализом данных, связанных с этими процессами.

На данный момент эта парадигма не охватывает уровни выше, чем молекулярный уровень структуры и организации, в том числе, например, такие вопросы, как специализация тканей во время развития или, в более обобщенном смысле, влияние условий окружающей среды на генетические события. В некоторых случаях простых обратных связей легко понять молекулярные механизмы того, как увеличение количества субстрата приводит к повышению продуктивности фермента, который катализирует трансформацию этого субстрата. Более сложными являются программы развития организма в течение его жизни. Эти интригующие вопросы о потоке информации и регуляции внутри организма сейчас попали в рамки основного направления биоинформатики.

Архивы данных и доступ к ним

Каждый банк данных содержит архив данных (логически организованную структуру данных), и инструментальные средства, необходимые для получения доступа к этим данным. Банки данных в молекулярной биологии содержат информацию о нуклеиновых кислотах и белковых последовательностях, макромолекулярных структурах и их функциях. Они включают в себя:

- Архивные банки данных, содержащие первичную биологическую информацию:
  - нуклеотидные (ДНК) и аминокислотные (белковые) последовательности (с аннотациями).
  - пространственные структуры белков и нуклеиновых кислот с аннотациями.
  - банки данных профилей экспрессии генов.
- Производные банки данных: они содержат информацию, собранную из архивных банков данных и из анализа их содержимого. Например:
  - мотивы последовательностей (характерные «подписи» белковых семейств)
  - мутации и варианты белковых и ДНК последовательностей. (На самом деле база данных по однонуклеотидным полиморфизмам является архивной. — *Прим. ред.*)
  - классификации и взаимосвязи (связи и характерные черты отдельных записей в архивах; например, банк данных ряда семейств белковых по-

следовательностей или иерархическая классификация способов белковой укладки)

- Библиографические банки данных
- Банки данных Web-сайтов
  - банки данных банков данных, содержащих биологическую информацию
  - связи между банками данных

Запросы к базам данных производятся с целью поиска набора записей (т. е. структур, последовательностей или иной информации) обладающих специфическими отличительными чертами или характеристиками либо с целью поиска сходства с последовательностью или структурой. Наиболее распространенным является запрос: «Я установил новую последовательность или структуру — содержат ли банки данных что-нибудь подобное?» Когда ряд последовательностей или структур, сходных с пробным объектом, уже получены из подходящего банка данных, исследователь может определить и исследовать их общие черты.

Механизм доступа к банку данных представляет собой набор инструментов для ответа на следующие вопросы:

- Содержит ли банк данных необходимую мне информацию? (Пример: В каком банке данных я могу найти аминокислотные последовательности алкогольдегидрогеназ?)
- Как я могу получить из банка данных избранную информацию в удобной форме? (Пример: Как можно составить выборку последовательностей глобина, или даже больше — таблицу, содержащую выровненные глобиновые последовательности?)
- Каталоги баз данных полезны в вопросах типа «Где я могу найти некоторую специфическую информацию?» (Пример: Какие банки данных содержат аминокислотную последовательность трипсина дикобраза?) Конечно, это простая задача, когда точно известно что именно требуется.

Банк данных без эффективных способов доступа, скорее всего, будет «свалкой» данных. Способ организации эффективного доступа — задача разработки банка данных, и в идеале он должен оставаться скрытым от конечного пользователя. Стало ясно, что эффективный доступ не может быть реализован путем встраивания системы запросов в неструктурированный архив. Вместо этого логическая организация хранения информации должна быть сконструирована разумно с точки зрения ее получения — какие типы вопросов пользователь может задать. И структура архива должна согласовываться с программным обеспечением для извлечения информации.

В биоинформатике может возникнуть множество различных типов запросов к базам данных, а именно:

(1) Дана последовательность или фрагмент последовательности, найти в базе данных последовательности, похожие на нее. Это центральная проблема биоинформатики. Задача сравнения строк присутствует и в других областях информатики. Например, программы для обработки и редактирования текста поддерживают функции поиска строк.

(2) Дана пространственная структура белка или ее фрагмент, найти в базе белки со сходными структурами. Это обобщение задачи поиска строки в трех измерениях.
(3) Дана последовательность белка с неизвестной структурой, найти *структуры белков*, которые могут быть сходны со структурой данного. Тут можно «сжульничать» — искать ответ среди структур белков, последовательности которых сходны с данной, предполагая, что если два белка имеют достаточно похожие последовательности, то они будут иметь и одинаковые структуры. Однако обратное неверно, и можно надеяться создать более мощные алгоритмы поиска, которые будут находить белки с гомологичными пространственными структурами, даже если их последовательности сильно различаются.
(4) Дана пространственная структура белка, найти последовательности белков, пространственные структуры которых схожи с данной. Опять же, можно использовать исходную структуру для поиска по базе данных пространственных структур, но это даст лишь частичный ответ, так как в настоящий момент известных последовательностей гораздо больше, чем структур. Поэтому желательно иметь метод, который позволит выделять из последовательности информацию о структуре.

Задачи (1) и (2) решаемы; такие поиски осуществляются тысячи раз в день. Задачи (3) и (4) — области активного исследования.

Задачи даже большей сложности возникают при желании изучать взаимодействия данных, хранящихся в разных банках. Это требует наличия связей, которые облегчают одновременный доступ к разным банкам данных. Пример: Для каких белков человека, связанных с нарушениями пуринового биосинтеза, с известной пространственной структурой существуют родственные белки у дрожжей? Задаем следующие условия для поиска: известная структура, установленная функция, обнаружение связей, корреляция с болезнью, определенные виды организмов. Рост важности одновременного доступа к банкам данных привел к рассмотрению их взаимодействия — как могут банки данных сообщаться друг с другом, не жертвуя возможностью структурировать свои собственные данные подходящим способом, который учитывает индивидуальные особенности содержащегося в них материала.

Проблема, которая еще не появилась в молекулярной биологии, — контроль обновлений архивов. База данных бронирования авиабилетов должна предотвратить продажу разными посредниками одного и того же места различным туристам. В биоинформатике пользователи могут читать и извлекать информацию из архивных банков данных или вводить материал для обработки персоналом архива, но не добавлять или изменять записи непосредственно. Эта ситуация может измениться. На практике количество генерируемых данных возрастает настолько быстро, что этот рост может лишить архивные проекты возможности усваивать их. Многие, не занимавшиеся этим ранее ученые сейчас начинают принимать участие в подготовке данных для архивов.

Несмотря на наличие разумных доводов в пользу единого контроля всех архивов, нет необходимости лимитировать количество способов доступа

к ним — проще говоря, дизайна внешнего интерфейса базы. Специализированные объединения пользователей могут извлекать подмножества данных, комбинировать данные из различных источников и предоставлять альтернативные способы доступа. Такие узконаправленные базы данных зависят от первоначальных архивов как источника содержащейся в них информации, но пересматривают организацию и представление данных по своему усмотрению. В самом деле, базы данных полученные различными способами могут распределять одну и ту же информацию по-разному. Разумную экстраполяцию можно реализовать, создав специализированные «виртуальные базы данных», которые основаны на исходных архивах, но предоставляют функции, приспособленные для нужд индивидуальных исследовательских групп или даже отдельных ученых.

Курирование, аннотация и контроль качества

Результаты исследований научных и медицинских сообществ зависят от качества банков данных. Индексы качества могут помочь нам избежать неверных заключений, даже если они и не позволяют производить коррекцию ошибок.

Записи банков данных включают в себя результаты экспериментов и дополнительную информацию — аннотации. Все эти данные могут содержать ошибки, вызванные различными причинами.

Самый важный критерий качества данных — технический уровень проведения экспериментов. Качество данных, полученных ранее, было лимитировано несовершенством методов; например, аминокислотные последовательности белков определялись секвенированием пептидов, но сейчас почти все они транслируются из ДНК. Сейчас количество данных быстро растет, и большинство из них — новые, полученные с помощью современных хорошо (или не очень. — *Прим. ред.*) работающих технологий.

Аннотации включают информацию об источнике данных и использованных методах их получения. В них есть ссылки на исследователей, получивших эти данные, и на наиболее важные публикации. Также они предоставляют ссылки на связанную информацию из других банков данных. В банках данных последовательностей аннотации включают *таблицы свойств* (*feature tables*): списки сегментов последовательности, имеющих биологическое значение — например, области последовательности ДНК, которые кодируют белки. Они имеют единый формат, а их содержание может быть ограничено списком типов свойств.

До недавнего времени типичная запись банка данных, содержащая последовательность ДНК, создавалась отдельной группой исследователей, изучавших и ген и его продукты как единое целое. Аннотации были основаны на экспериментальных данных и писались специалистами. Проекты расшифровки полных геномов, напротив, не предоставляют ни экспериментального подтверждения экспрессии большинства предполагаемых генов, ни характеристики их продуктов. Банки данных базируют свои аннотации на компьютерном анализе.

Аннотации — наиболее слабое место геномных проектов. Невозможно добиться полностью автоматического аннотирования; создание правильных ан-

нотаций — трудоемкий ручной процесс. Но значение правильной аннотации не может быть переоценено. Пир Борк (P. Bork) замечает, что ошибки в аннотировании сводят на нет высокое качество экспериментальных данных.

Рост количества геномных данных позволит повысить качество аннотаций, так как возрастет точность статистических методов. Это позволит произвести *реаннотацию* записей баз данных. Процесс уточнения аннотаций станет совершеннее. Но беспокоит то, что неизбежным последствием этих преобразований будет поточная аннотация. Будут ли законченные исследовательские проекты периодически перепроверяться и выводы пересматриваться? Проблема усугубляется ростом количества Web-сайтов со все более плотными сетями ссылок. Конечно, они предоставляют удобный доступ к приложениям. Но Интернет также является и «переносчиком инфекции», распространяя ошибки предварительных данных, которые впоследствии исправляются, но эти исправления уже не доходят до конечного пользователя, и возникают разночтения в аннотациях (Есть замечательный пример — в архее Archeoglobus был аннотирован ген Главного комплекса гистосовместимости — важного белка иммунной системы. Это было очевидной ошибкой автоматического аннотирования. Основываясь на этой аннотации в другой архее, Metanococcus Janshi, весьма уверенно аннотирован также ген Главного комплекса гистосовместимости, поскольку он почти идентичен соответствующему гену в Archeoglobus. — *Прим. ред.*)

Единственное возможное решение — *распределенный*, *динамический* процесс аннотации и коррекции ошибок. Распределенный, так как персонал банка данных не имеет ни времени, ни опыта для такой работы; курировать этот процесс должны будут специалисты. Динамический, так как прогресс в автоматизации аннотирования и идентификации/коррекции ошибок позволит проводить реаннотацию. Мы будем вынуждены отказаться от безопасной концепции создания банка данных, состоящего из изначально верных записей, остающихся неизменными и впоследствии. Банки данных будут содержать непрерывно изменяющиеся и растущие в размере данные, качество которых, будем надеяться, повысится.

Всемирная Паутина (The World Wide Web)

Похоже, все читатели используют Интернет: для поиска справочных материалов, новостей, для доступа к банкам данных в молекулярной биологии, для получения личной информации об отдельных людях — о друзьях, коллегах или о знаменитостях — или же просто для просмотра различной информации. По существу, Интернет — это способ связи между людьми (и между компьютерами) через сеть. Он объединяет всех в некий совершенный информационный город, в котором есть аналог библиотеки, почты, магазинов и школ.

Вы, пользователь, запускаете браузер (программу-обозреватель) у себя на персональном компьютере. Самые распространенные браузеры — это Netscape и Internet Explorer (в настоящее время на смену Netscape пришли программы

Mozila и FireFox. — *Прим. ред.*) При помощи этих программ-навигаторов вы можете читать и отображать материалы из любой точки мира. С помощью меню браузера вы можете передвигаться на предыдущую или следующую страничку или прервать загрузку данных. Обозреватель также позволяет вам загружать и сохранять информацию на свой локальный компьютер.

Отображаемые материалы содержат встроенные ссылки, которые позволяют вам перемещаться по Web-страницам и сайтам, открывая новые возможности в вашем путешествии. Взаимосвязи оживляют Интернет. Уникальность человеческого мозга вызвана не общим количеством нейронов, а плотностью образуемых ими взаимосвязей. Подобным образом и могущество Интернета основано не столько на количестве пользователей, а скорее на их многочисленных связях друг с другом.

В рассматриваемых вами документах почти всегда присутствуют ссылки. Запуская программу-обозреватель, вы попадаете на какую-то страницу. Она будет содержать различные активные элементы: выделенные слова, кнопки, картинки. Обычно такие элементы — ссылки — выделены яркими цветами. Нажав на ссылку, вы переместитесь на новую страницу. В то же время вы автоматически оставляете за собой тропинку из «электронных хлебных крошек» (Здесь по-видимому имеется в виду ссылка на известную сказку. — *Прим. ред.*), так что вы можете вернуться туда, откуда пришли, чтобы продолжить дальнейшее внимательное рассмотрение той страницы, с которой вы стартовали.

Интернет можно представить как огромную всемирную информационную доску. Там содержатся тексты, изображения, фильмы и звукозаписи. Практически все, что может быть сохранено на компьютере, можно сделать доступным через Интернет. Интересный пример — сайт, посвященный поэзии Уильяма Батлера Ейтса. Самая верхняя (первая) страница содержит данные, которые можно назвать оглавлением. Следуя по ссылкам, изображенным на этой первой странице, вы можете увидеть напечатанный текст различных поэм. Вы можете сравнивать различные издания. Вы можете получить доступ к критической литературе на эти поэмы. Вы можете увидеть некоторые из этих поэм в рукописях Ейтса. Для отдельных произведений есть даже ссылки на аудиофайлы, где вы сможете послушать, как Ейтс сам читает свои поэмы.

Ссылки могут быть внешними или внутренними. Внутренние ссылки могут отправить вас к следующему участку текста или к картинкам, видеозаписям или звуковым файлам. Внешние ссылки либо отправляют вас *глубже*, к более специализированным документам, либо *выше*, к более общим (например, обеспечивая базовые знания по техническим вопросам), *в сторону*, к похожим документам (другим материалам, посвященным тому же самому вопросу), или *наружу*, к директориям, показывающим, какие другие важные документы также доступны.

Главное, что необходимо для эффективного использования Интернета, — это правильно выбрать стартовую страницу. Начав работу, далее вы можете передвигаться по ссылкам куда хотите. Среди наиболее важных ресурсов Интернета — *поисковые сайты*, которые индексируют весь Интернет и позволяют вам осуществлять поиск сайтов по ключевым словам. Вы можете ввести один или несколько терминов, например, «фосфорилаза», «аллостерические изме-

нения», «кристаллическая структура», и поисковая система даст вам список ссылок на Интернет-ресурсы, в которых содержатся эти термины. Затем вы сами определяете значимость найденных сайтов в соответствии с вашими интересами.

Закончив работу в Интернете на какой-либо странице, в следующий раз вы можете снова попасть на нее благодаря средствам обслуживания, встроенным в браузер. Межсессионная память браузера позволяет вам начать работу в Интернете прямо там, где вы закончили. Если во время работы вы нашли важный документ, к которому впоследствии захотите вернуться снова, вы можете сохранить ссылку на него в файле с *закладками* или добавить ее в список «*избранное*». В последующих сеансах вы сможете вернуться на любой сайт из этого списка сразу, непосредственно, без необходимости проделывать весь тот путь ссылок, который привел вас на этот сайт в первый раз.

Личная *домашняя страничка* — это (обычно. — *Прим. ред.*) короткий автобиографический очерк (разумеется, тоже со ссылками). У ваших коллег по работе наверняка есть их собственные домашние странички, где обычно написано имя, подразделение института, ссылки на доклады и адреса электронной почты, номера телефонов и факсов, список публикаций и перечень текущих интересов. Также нет ничего необычного в том, чтобы поместить на домашней страничке какую-нибудь личную информацию о своих увлечениях, картинки, семейные фотографии — с женой, детьми, даже с любимой собакой!

Но Интернет — это не только «улица с односторонним движением». На многих сайтах есть специальные формы, где вы можете вводить информацию и затем запускать выполнение программ на удаленном компьютере (Web-сервере), которые тут же выдадут вам запрашиваемый результат. Типичный пример такого сайта — программа-поисковик. Многие вычисления в биоинформатике стали возможны именно благодаря таким Web-серверам. Если вычисления слишком сложны, результаты могут быть готовы не сразу же, а высланы впоследствии по электронной почте.

Что такое URL?

Даже небольшого опыта работы в Интернете достаточно, чтобы заметить странно выглядящие характерные записи, которые индивидуальны для каждой web-страницы. Эти записи и есть URL — унифицированный локатор (определитель местонахождения) ресурса (Uniform Resource Locator). Они отражают информацию о формате записи и о ее местонахождении. Кроме всего прочего, каждый документ в Интернете — это файл, который размещен на каком-нибудь компьютере. Рассмотрим пример URL:

```
http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html
```

Эта страница содержит полезные инструкции относительно поиска информации в Интернете. Префикс `http://` указывает, что для обмена информацией используется Протокол Передачи Гипертекста (HyperText Transfer Protocol). Это говорит вашему браузеру, что следует ожидать информации в формате `http`, который используется наиболее часто. Следующая часть

URL, `www.lib.berkeley.edu`, — это название компьютера, в данном случае — компьютера центральной библиотеки Университета Калифорнии в Беркли (University of California at Berkeley). Окончание ссылки описывает размещение на компьютере и название файла, содержимое которого ваш браузер собирается отображать.

Электронные публикации

Все больше и больше публикаций появляется на просторах Интернета. Научные журналы могут, как объявлять только содержание номеров, так и выкладывать названия статей и их краткое содержание, или же даже полные тексты статей (в настоящее время появился ряд исключительно электронных журналов. — *Прим. ред.*). В Интернете появляются разнообразные полезные данные — информационные бюллетени, технические описания. Многие журналы и газеты также постепенно переходят в электронный вид. Возможно, вы захотите посмотреть `http://www.nytimes.com` — сайт газеты «The New York Times». Многие печатные издания сейчас содержат ссылки на Интернет-ресурсы, содержащие дополнительные материалы, которые никогда не появляются в печати.

Мы сейчас переживаем эру перехода к «небумажным публикациям». Уже считается хорошим тоном в напечатанной статье давать свой электронный адрес или URL вашей домашней странички.

В связи с электронными публикациями возникает ряд вопросов. Один из них — вопрос о равнозначности электронных и напечатанных обзоров. Как мы можем гарантировать такое же качество электронных публикаций, какое мы привыкли ожидать от печатных изданий? Следует ли считать электронные публикации (наряду с печатными) при оценке продуктивности (если даже не качества) соискателей? Один известный наблюдатель высказал разумное (хотя возможно несколько преувеличенное) предположение: «Как только Гарвард или Стэнфорд начнут признавать электронные публикации, 90% научных журналов исчезнут в одночасье»[1)].

Компьютеры и компьютерные науки

Биоинформатика была бы невозможна без достижений в компьютерном оборудовании и программном обеспечении. Быстрые и высокоэффективные запоминающие и сохраняющие информацию средства необходимы даже для обслуживания архивов. Информационный поиск и анализ требует программ; как блестяще простых в одних случаях, так и чрезвычайно усложненных в других. Для распределения информации (и вычислений. — *Прим. ред.*) требуются возможности компьютерных сетей и Интернета.

[1)] В настоящее время появился ряд исключительно электронных журналов, в которых проводится полный цикл реферирования. Статьи в этих журналах являются общепризнанными публикациями. — *Прим. ред.*

Компьютерная наука — молодая цветущая область, задача которой — сделать наиболее эффективным применение информационных технологий. Определенные области теоретической компьютерной науки тесно связаны напрямую с биоинформатикой. Давайте рассмотрим их относительно некоторой конкретной биологической задачи: «Отыскать в банке данных все последовательности, похожие на данную последовательность-образец». Хорошее решение этой проблемы попросило бы от компьютерной науки следующего:

- *Проанализировать алгоритмы. Алгоритм — это полное и точное определение последовательности действий для решения данной задачи.* Для поиска похожих последовательностей нам нужно оценить степень сходства заданной последовательности с каждой последовательностью из банка данных. Такое сравнение можно сделать гораздо более эффективными способами, чем простейший алгоритм рассмотрения каждой пары в каждой позиции в каждом возможном сопоставлении — метод, который даже без учета возможных пропусков требует для своей работы времени, пропорционального количеству символов в заданной последовательности, умноженному на количество символов в последовательностях из банка данных. Особое направление в компьютерной науке — разговорно выражаясь, «стрингология», или наука о строках — сфокусировано на разработке эффективных методов для проблем такого типа и анализе их эффективного исполнения[1)].
- *Структура данных и поиск информации.* Как можно организовать данные для наиболее эффективного поиска ответов на запрос? Например, есть ли способ проиндексировать или каким-то иным способом предварительно обработать имеющиеся данные, чтобы сделать поиск по сходству последовательностей более эффективным? Как мы можем обеспечить такой интерфейс, который поможет пользователю в оформлении и реализации запроса?
- *Разработка программного обеспечения.* Вряд ли кто-либо когда-нибудь писал программы на «родном» языке компьютера. Программисты работают на языках более высокого уровня, таких как C, C++, PERL (Practical Extraction and Report Language) или даже FORTRAN (в последнее время все большую популярность приобретает программирование на языке Java. — *Прим. ред.*). Выбор языка программирования зависит от алгоритма и связан со структурой данных и требованиями к программе. Конечно, в большинстве случаев сложное программное обеспечение, использующееся сейчас в биоинформатике, пишется специалистами. Итак, вопрос: как много программирования необходимо в биоинформатике?

[1)] В большинстве случаев перед тем, как строить и анализировать алгоритм требуется *формализовать* задачу — перевести задачу с языка биологии на язык математики, и только потом размышлять о путях ее решения. Значительная часть успеха или неуспеха в решении биологической задачи с помощью компьютера зависит именно от этого этапа.

Программирование

Программирование в информатике — это все равно, что кирпичная кладка в архитектуре. В обоих случаях речь идет о создании; только одно — искусство, а другое — ремесло.

Многих студентов, начинающих заниматься биоинформатикой, интересует вопрос, надо ли изучать написание сложных компьютерных программ. Мой совет (с которым далеко не каждый может согласиться) такой: «Нет. Но только до тех пор, пока вы не пожелаете специализироваться в этом». Для работы в биоинформатике вам понадобится развить умение грамотно использовать средства, находящиеся в наличии на web. Необходимо знать, как создавать и поддерживать Web-сайт. Конечно, вам также понадобится уверенное владение той операционной системой, которая установлена на вашем компьютере. Некоторые навыки в написании простых скриптов на языках программирования вроде Perl, существенно увеличат продуктивность операционной системы.

С другой стороны, размеры банков данных и возрастающая сложность вопросов, стоящих перед нами, нуждаются в здоровом уважении к ним. В самом деле, творческий подход к программированию на бескрайнем поле действия есть лучшее для специалиста, хорошо владеющего информатикой. А использование программ через «отполированный» пользовательский web-интерфейс (в широком смысле) не говорит о той деятельности, которая связана с созданием и распространением программного обеспечения. Бисмарк однажды сказал: «Тот, кто любит удовольствие или закон, не следит за тем, чтобы все это было соблюдено одновременно». Наверное, то же можно сказать и о компьютерных программах.

Я рекомендую развить основные навыки программирования на примере Perl. Perl — это очень мощное средство. Он сделан для облегчения выполнения часто встречающихся несложных задач. Perl также имеет преимущество в том, что он работает под управлением многих операционных систем.

Насколько глубоко вам следует изучать Perl, чтобы применять его в биоинформатике? Многие институты проводят у себя курсы. Другой прекрасный способ получения знаний — от своих коллег, и он зависит от соотношения между вашей искусностью и их терпением. В наличие есть книги. Очень полезный подход — поискать уроки на Web; попробуйте дать мощной поисковой машине фразу «Perl tutorial», и вы получите множество полезных сайтов, которые постепенно дадут вам основы. Естественно, используйте данные советы как можно чаще при работе. Данная книга не научит вас языку Perl, но она создаст ощущение возможности практиковаться в том, что вы будете изучать где бы то ни было еще.

В книге представлена *простая* программа на языке Perl. Сильная сторона Perl заключается в командной строке, делающей этот язык подходящим для анализа последовательностей в биологических задачах. Эта программа (см.ниже) транслирует нуклеотидную последовательность в аминокислотную в соответствие со стандартным генетическим кодом. Первая строчка, `#!/usr/bin/perl`, означает сигнал для операционной системы UNIX (или LINUX), о том, что нужно запустить компьютерную программу на языке Perl.

```
#!/usr/bin/perl
#translate.pl - translate nucleic acid sequence to protein sequence
#               according to standard genetic code

#   set up table of standard genetic code

%standardgeneticcode = ( "ttt"=> "Phe",   "tct"=> "Ser", "tat"=> "Tyr",   "tgt"=> "Cys",
 "ttc"=> "Phe",   "tcc"=> "Ser", "tac"=> "Tyr",   "tgc"=> "Cys",
 "tta"=> "Leu",   "tca"=> "Ser", "taa"=> "TER",   "tga"=> "TER",
 "ttg"=> "Leu",   "tcg"=> "Ser", "tag"=> "TER",   "tgg"=> "Trp",
 "ctt"=> "Leu",   "cct"=> "Pro", "cat"=> "His",   "cgt"=> "Arg",
 "ctc"=> "Leu",   "ccc"=> "Pro", "cac"=> "His",   "cgc"=> "Arg",
 "cta"=> "Leu",   "cca"=> "Pro", "caa"=> "Gln",   "cga"=> "Arg",
 "ctg"=> "Leu",   "ccg"=> "Pro", "cag"=> "Gln",   "cgg"=> "Arg",
 "att"=> "Ile",   "act"=> "Thr", "aat"=> "Asn",   "agt"=> "Ser",
 "atc"=> "Ile",   "acc"=> "Thr", "aac"=> "Asn",   "agc"=> "Ser",
 "ata"=> "Ile",   "aca"=> "Thr", "aaa"=> "Lys",   "aga"=> "Arg",
 "atg"=> "Met",   "acg"=> "Thr", "aag"=> "Lys",   "agg"=> "Arg",
 "gtt"=> "Val",   "gct"=> "Ala", "gat"=> "Asp",   "ggt"=> "Gly",
 "gtc"=> "Val",   "gcc"=> "Ala", "gac"=> "Asp",   "ggc"=> "Gly",
 "gta"=> "Val",   "gca"=> "Ala", "gaa"=> "Glu",   "gga"=> "Gly",
 "gtg"=> "Val",   "gcg"=> "Ala", "gag"=> "Glu",   "ggg"=> "Gly"
);


#   process input data

while ($line = <DATA>) {                                    # read in line of input
    print "$line";                                          # transcribe to output
    chop();                                                 # remove end-of-line character
    @triplets = unpack("a3" x (length($line)/3), $line);    # pull out successive triplets
    foreach $codon (@triplets) {                            # loop over triplets
        print "$standardgeneticcode{$codon}";               # print out translation of each
    }                                                       # end loop on triplets
    print "\n\n";                                           # skip line on output
}                                                           # end loop on input lines

#   what follows is input data

__END__
atgcatccctttaat
tctgtctga
```

Running this program on the given input data produces the output:

```
atgcatccctttaat
MetHisProPheAsn

tctgtctga
SerValTER
```

Текст внутри программы, начинающийся со знака # и идущий до конца строки, представляет собой лишь комментарий. Слово `__END__` подает сигнал о том, что текст программы закончен, а далее следует входной файл. (Все материалы, которые читатель может найти здесь, полезно иметь в цифровом виде, включая все программы, информация на web-сайте книги: `http://www.oup.com/uk/lesk/bioinf`)

Даже этот простой пример компьютерной программы показывает несколько особенностей языка программирования Perl. Файл содержит вложенные данные (трансляционную таблицу генетического кода), указания

компьютеру, что делать с входящими данными (последовательность, которую требуется транслировать) и входящие данные (пишется после `__END__`). Примечания после # показывают секции программы и описывают результат каждой команды.

Программа разбита на блоки, вложенные в фигурные скобки: {...}, оказывающиеся полезными при исполнении потока. Индивидуальные команды внутри блока (каждая заканчивается на точку с запятой;) осуществляются в порядке их записи в программе. Наружный блок называется *циклом*.

```
while ($line = <DATA>) {
    ...
}
```

`<DATA>` относится к строке со входными данными (появляющимися после `__END__`). Блок осуществляется один раз для каждой линии входных данных, т. е. `while` появляется в любой строке, где упоминается о входных данных.

В программе есть три типа данных структуры. Строка входных файлов, обозначаемая `$line`, — простейшая *строка букв*. Она делится на два *массива*, или вектора триплетов. Массив хранит несколько элементов в линейном порядке, доступ к индивидуальным элементам данных может быть получен по позиции (индексу) в массиве. Для облегчения просмотра кодирования той или иной аминокислоты соответствующим триплетом, генетический код хранится в *ассоциативном массиве*. Ассоциативный массив (или хеш-таблица) — это обобщение простого или последовательного массива. Если элементы простого массива пронумерованы последовательными целыми числами, то элементы ассоциативного массива пронумерованы *любыми* символами, в данном случае 64 триплетами. Мы обработали входящие триплеты *в порядке их появления* в нуклеотидную последовательность, но мы нуждаемся в доступе к элементам таблицы генетического кода *в произвольном порядке*, — так, как они продиктованы последовательностью триплетов. Простой массив, или вектор символов, предназначен для перевода последовательности триплетов, а ассоциативный массив предназначен для просмотра аминокислот, которые относятся к ним.

Здесь же другая программа Perl, которая иллюстрирует дополнительные аспекты данного языка.[1)] Данная программа разбирает предложение:

> Весь мир — сцена,
> И все мужчины и женщины только актеры;
> У них есть их выходы и их входы,
> И один человек в свое время играет много ролей.

После чего оно разбивается на случайные перекрывающееся фрагменты (знаки \n во фрагментах являются окончаниями строк в оригинале):

> мужчины и женщины только актеры;\n
> один человек в свое время
> Весь мир

[1)] Эта часть может быть пропущена при первом чтении.

```
их входы,\nИ один человек
сцена\nИ все мужчины и женщины
У них есть их выходы и их входы,\n
мир — сцена,\nИ все
их входы,\nИ один человек
в свое время играет много ролей.
только актеры;\nУ них есть
```

Этот тип вычислений важен для восстановления полной последовательности ДНК по перекрывающимся фрагментам. (Для знакомства с существенными проблемами, связанными с повторами — см. задачу 1.4.)

Чтобы убедиться, позволят ли ваши программистские амбиции пройти через простые задачи, познакомьтесь с проектом BioPerl — источником свободно распространяемых кодов Perl программ и компонентов (см. `http://bio.perl.org/`).

Биологическая классификация и номенклатура

Вернемся в девятнадцатый век, когда жизнь ученых была, по крайней мере, в некотором отношении проще.

Биологическая номенклатура основана на идее, что живые организмы подразделяются на виды — группы схожих организмов с одинаковым геномом. (Вопрос, почему живые организмы должны быть «квантованы» на *дискретные* виды — очень сложен). Линней, шведский натуралист, классифицировал организмы согласно иерархии: царство, тип, класс, порядок, семейство, род и вид (см. таблицу). Современные классификаторы (taxonomists) вводят также некоторые дополнительные уровни (таконы). В настоящее время общепринята *двойная* (*биноминальная* номенклатура), суть которой в том, что название вида состоит из двух латинских слов: первое — название рода, второе — название вида. Например, человек относится к виду *Homo sapiens*, плодовая мушка — к виду *Drosophila melanogaster*. Каждый вид однозначно определяется двойным названием, кроме того, для некоторых видов существуют тривиальное название (например, *Bos Taurus* — бык (корова)). Конечно, большинство видов такого названия не имеют.

Первоначально линнеевская систематика была единственной, и она была основанной на наблюдении сходств и различий организмов. С развитием теории эволюции было выяснено, что эта система довольно точно отражает родословную вида. Но возникает вопрос, при наличии каких сходств можно считать, что у организмов общий предок? (Строго говоря, у всех организмов был общий предок, который возник в момент возникновения жизни на земле. Здесь надо понимать, что имеется в виду общий предок на разумном эволюционном расстоянии. — *Прим. ред.*). Органы, имеющие одинаковое происхождение, называются *гомологичными* (например, рука человека и крыло орла). Другие, очень похожие органы могли произойти независимо друг от друга в результате *конвергентной эволюции*. Например, крыло орла и крыло

```
#!/usr/bin/perl
#assemble.pl - assemble overlapping fragments of strings

#  input of fragments
while ($line = <DATA>) {                    #   read in fragments, 1 per line
   chop($line);                             #   remove trailing carriage return
   push(@fragments,$line);                  #   copy each fragment into array
}
#  now array  @fragments  contains fragments

#  we need two relationships between fragments:
#  (1) which fragment shares no prefix with suffix of another fragment
#      * This tells us which fragment comes first
#  (2) which fragment shares longest suffix with a prefix of another
#      * This tells us which fragment  follows  any fragment

#  First set array of prefixes to the default value   "noprefixfound".
#      Later, change this default value when a prefix is found.
#      The one fragment that retains the default value must be come first.

#  Then loop over pairs of fragments to determine maximal overlap.
#      This determines successor of each fragment
#      Note in passing that if a fragment has a successor then the
#          successor must have a prefix

foreach $i (@fragments) {                   #   initially set  prefix of each fragment
    $prefix{$i} = "noprefixfound";          #       to  "noprefixfound"
    }                                       #   this will be overwritten when a prefix is found

#  for each pair, find longest overlap of suffix of one with prefix of the other
#       This tells us which fragment FOLLOWS any fragment

foreach $i (@fragments) {                   #   loop over fragments
    $longestsuffix = "";                    #   initialize longest suffix to null

    foreach $j (@fragments) {               #   loop over fragment pairs
        unless ($i eq $j) {                 #   don't check fragment against itself

            $combine = $i . "XXX" . $j;  #   concatenate fragments, with fence XXX
            $combine =~ /([\S ]{2,})XXX\1/;   #   check for repeated sequence
            if (length($1) > length($longestsuffix)) {    # keep longest overlap
                $longestsuffix = $1;        #   retain longest suffix
                $successor{$i} = $j;        #   record that $j follows $i
            }

        }
    }
    $prefix{$successor{$i}} = "found";   #  if $j follows $i then $j must have a prefix
}


foreach (@fragments) {                      #  find fragment that has no prefix; that's the start
    if ($prefix{$_} eq "noprefixfound") {$outstring = $_;}
}

$test = $outstring;                         #   start with fragment without prefix
while ($successor{$test}) {                 #   append fragments in order
     $test = $successor{$test};             #   choose next fragment
     $outstring = $outstring . "XXX" . $test;  # append to string
     $outstring =~ s/([\S ]+)XXX\1/\1/;  #   remove overlapping segment
}

$outstring =~ s/\\n/\n/g;                   #   change signal \n to real carriage return
print "$outstring\n";                       #   print final result

__END__
the men and women merely players;\n
one man in his time
All the world's
their entrances,\nand one man
stage,\nAnd all the men and women
They have their exits and their entrances,\n
world's a stage,\nAnd all
their entrances,\nand one man
in his time plays many parts.
merely players;\nThey have
```

Классификация человека и плодовой мухи

	Человек	Плодовая муха
Царство	животное	животное
тип	позвоночное	беспозвоночное
класс	млекопитающее	насекомое
порядок	примат	двукрылое
семейство	гоминид	дрозофолида
род	человек	дрозофила
вид	*разумный*	*melanogaster*

пчелы являются результатом конвергентной эволюции и выполняют схожие функции (хотя их общий предок вообще не имел крыльев). Наоборот, в результате дивергенции гомологичные органы могут существенно различаться по строению и функциям. Например, слуховые косточки среднего уха человека гомологичны костям челюсти примитивных рыб, а евстахиева труба — жаберным щелям. В большинстве случаев ученые могут различить истинно гомологичные органы и органы, ставшие похожими в результате конвергенции.

Наиболее точные сведения относительно родства организмов дает анализ их последовательностей. Хорошо изучена систематика высших организмов, для которых анализ последовательностей и классические методы сравнительной анатомии, палеонтологии и эмбриологии обычно дают полную картину. Классификация микроорганизмов более трудна, отчасти потому, что не очень понятно по каким признакам их классифицировать, а отчасти потому, что происходят интенсивные миграции генов, из-за которых картина может полностью перевернуться.

Рибосомные РНК являются необходимым компонентом всех организмов с правильной организацией (слишком большое или слишком малое расхождение (divergence) и родство — плохо определяются).

Основываясь на 15S рибосомных РНК, С. Воз разделил все живые организмы на три империи: бактерии, археи и эукариоты (см. рис. 1.2). Бактерии и археи — прокариоты, их клетки не содержат оформленного ядра. Типичные представители бактерий — микроорганизмы, являющиеся причиной многих заболеваний, а также *Escherichia coli* — главное 'подопытное животное' молекулярной биологии. В империю архей входят термофилы, галлофилы, серовосстанавливающие и метанобразующие архебактерии. Мы относимся к эукариотам — организмам, клетки которых содержат ядро. Кроме нас к ним относятся дрожжи и все многоклеточные организмы (а также амебы, инфузории и многие другие организмы. — *Прим. ред.*).

Наиболее изучены геномы бактерий, так как это клинически важно. При этом выяснилось, что их геном относительно прост. Тем не менее, мы можем

Рис. 1.2. Основная классификация живых организмов, полученная Возом (C. Woese) на основе анализа последовательностей 15S РНК

узнать о нас больше от архей, чем от бактерий. Вопреки очевидным различиям в жизненных формах и отсутствию в клетках архей ядра, они в некотором смысле на молекулярном уровне ближе к эукариотам, чем к бактериям. Похоже, что археи из всех живых организмов, наиболее близки к корню дерева жизни.

Рисунок 1.2 демонстрирует самый глубокий уровень дерева жизни. Ветвь Eukarya включает в себя животных, растения, грибы и одноклеточные организмы. В вершине ветви Eukarya находятся metazoa (многоклеточные организмы) (рис. 1.3.). Мы и наши ближайшие родичи являемся вторичноротыми (рис. 1.4).

Использование последовательностей для определения филогенетических взаимосвязей

В предыдущих разделах рассматривались банки данных последовательностей и биологические взаимосвязи. Здесь можно найти примеры применения извлечения последовательностей из банка данных и сопоставления последовательностей для анализа биологических взаимосвязей.

ПРИМЕР 1.1.

Восстановление аминокислотной последовательности панкреатической рибонуклеазы лошади.

Используйте сервер ExPASy швейцарского института биоинформатики: URL `http://www.expasy.\ch/cgi-bin/sprot-search-ful`. Наберите ключевое слово `horse pancreatic ribonuclease` нажмите `ENTER`. Выберите `RNP_HORSE`

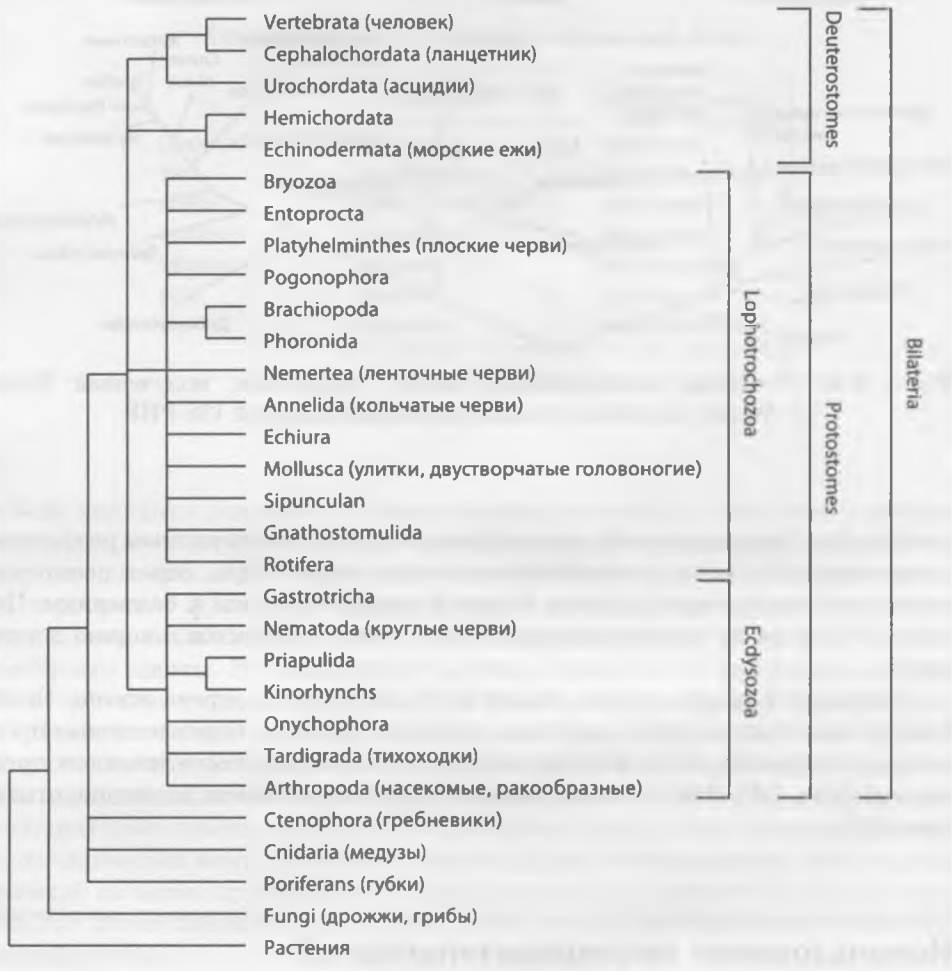


Рис. 1.3. Филогенетическое дерево metazoa (многоклеточных животных). Группа Bilateria включают в себя всех животных, которые имеют двустороннюю симметрию строения тела. Первичноротые (protostomes) и вторичноротые (deuterostomes) представляют собой два основных рода, разделившихся на ранней стадии эволюции, приблизительно 670 млн лет назад. Они демонстрируют различные модели эмбрионального развития, включая различные процессы раннего деления клетки, противоположную ориентацию зрелого кишечника в отношении ранней инвагинации бластулы, и происхождение скелета из мезодермы (вторичноротые) или эктодермы (первичноротые). Первичноротые включают в себя две подгруппы, различающиеся на базе 18S РНК (из малой рибосомной субъединицы) и генных последовательностей HOX. Морфологически Ecdysozoa имеет линяющие кутикулы — жесткий внешний слой органической материи. Lophotrochozoa имеет мягкое тело. (Основано на Adouette.A, Balavoine.G, Lartillot.N, Lespinet.O, Prud'homme.B и de Rosa.R (2000) «Новый филогенез животных: достоверность и применение», *Деятельность Национальной Академии Наук США 97, 4453–6*)

Рис. 1.4. Филогенетическое дерево позвоночных и наших ближайших родичей. Хордовые, включая позвоночных, и иглокожие все являются вторичноротыми

и затем `FASTA format` (см. Box: FASTA format). Получим следующий результат (первая строка усечена):

```
>sp|P00674|RNP_HORSE RIBONUCLEASE PANCREATIC (EC 3.1.27.5) (RNASE 1) ...
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEP
LADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTS
QKERHIIVACEGNPYVPVHFDASVEVST
```

который можно вырезать и скопировать в другие программы.

Например, мы можем восстановить несколько последовательностей и выровнять их (см. врезку: Выравнивание последовательностей). Анализ структуры подобия среди выровненных последовательностей является полезным в оценке близости взаимосвязи.

ПРИМЕР 1.2.

Даны последовательности панкреатической рибонуклеазы лошади *(Equus caballus)*, малого полосатика *(Balaenoptera acutorostrata)* и большого рыжего кенгуру *(Macropus rufus)*. Какие из перечисленных животных наиболее близки?

Зная, что лошадь и кит являются плацентарными млекопитающими, а кенгуру — сумчатое, мы предполагаем, что кит и лошадь будут более близкой парой. Найдем три последовательности как в предыдущем примере:

```
>RNP_HORSE
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEP
LADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTS
QKERHIIVACEGNPYVPVHFDASVEVST
>RNP_BALAC
RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHES
```

Формат FASTA

Очень распространенный формат данных последовательностей для программы FAST (быстрого) Aligment (выравнивания) В. Р. Пирсона (W. R. Pearson). Многие программы используют формат FASTA для чтения последовательностей, или для оформления результатов.

Последовательность в формате FASTA:

- Начинается с простой строки описаний. В первой колонке должно стоять >. Остальное содержимое заголовочной строки является произвольным, но должно быть информативным.
- Следующие строки содержат последовательность, по одному символу на каждый остаток.
- Используйте однобуквенные коды для нуклеотидов и аминокислот, заданные Международным Объединением Биохимии и Международным Объединением Чистой и Прикладной Химии (IUB/IUPAC).
  См. `http://www.chem.qmw.ac.uk/iupac/misc/naabb.html`
  и `http://www.chem.qmw.ac.uk/iupac/AminoAcid/`
  Используйте Sec и U как трехбуквенный и однобуквенный коды для селеноцистенина:
  `http://www.chem.qmw.ac.uk/iubmb/newsletter/1999/item3.html`
- Строки могут иметь разную длину; это граница с «рваным» правым краем.
- Многие программы воспринимают маленькие буквы в качестве кодов аминокислот.

Пример формата FASTA: глутатион пероксидаза быка

```
>gi|121664|sp|P00435|GSHC\_BOVIN GLUTATHIONE PEROXIDASE
MCAAQRSAAALAAAAPRTVYAFSARPLAGGEPFNLSSLRGKVLLIENVASLUGTTVRDYTQMNDLQRRLG
PRGLVVLGFPCNQFGHQENAKNEEILNCLKYVRPGGGFEPNFMLFEKCEVNGEKAHPLFAFLREVLPTPS
DDATALMTDPKFITWSPVCRNDVSWNFEKFLVGPDGVPVRRYSRRFLTIDIEPDIETLLSQGASA
```

Строка заголовка имеет следующие поля:
> обязательный символ в столбце 1

gi|121664 это номер *geninfo*, идентификатор, назначенный Национальным Центром США по Биотехнологической информации (NCBI). Каждая последовательности в банке данных ENTREZ имеет уникальный идентификатор gi. NCBI собирает последовательности из разных источников, включая первичные архивы данных и заявления на получение патента. Его номера gi обеспечивают общий и непротиворечивый идентификатор-«зонтик», накладывающийся на различные соглашения для баз данных-источников. Если база данных — источник обновляет информацию, NCBI создает новую запись с новым номером gi, если эти изменения затронули последовательность, но обновляет и сохраняет запись, если изменения коснулись только информации, не входящей в последовательность, например, цитирование литературы.

Запись **sp|P00435** что источником информации является SWISS-PROT, и что номером доступа к записи SWISS-PROT является P00435.

GSHC_BOVIN GLUTATHIONE PEROXIDASE это идентификатор SWISS-PROT для последовательности и видов, (GSHC_BOVIN), за которым следует имя молекулы.

Выравнивание последовательностей

Выравнивание последовательностей — это установление соответствия остаток-остаток. Мы можем искать:

- *Глобальное совпадение*: выровнять всю последовательность против другой последовательности.

```
And.--so,.from.hour.to.hour,.we.ripe.and.ripe
||||    ||||||||||||||||||||||||    ||||||
And.then,.from.hour.to.hour,.we.rot-.and.rot-
```

  Это иллюстрирует несоответствия, вставки и удаления.

- *Локальное совпадение*: поиск части последовательности, которая совпадает с частью другой последовательности.

```
   My.care.is.loss.of.care,.by.old.care.done,
     |||||||||    ||||||||||||    |||||| ||
 Your.care.is.gain.of.care,.by.new.care.won
```

  Для локального совпадения выступающие концы не рассматриваются как пропуски(делеции). В дополнение к несовпадениям, видным в данном примере, возможны также вставки и удаления внутри совпадающей части.

- *Поиск мотивов совпадения*: поиск совпадения короткой последовательности в одном или более отрезках длинной последовательности. В этом случае допускается несовпадение одного символа. По выбору можно потребовать полного совпадения, или допустить большее число несовпадений или даже пропусков.

```
        match
         ||||
for the watch to babble and to talk is most tolerable
```

  или:

```
                                match
                                 ||||
Any thing that's mended is but patched: virtue that transgresses is
    match                                                 match
     ||||                                                  ||||
but patched with sin; and sin that amends is but patched with virtue
```

- *Множественное выравнивание*: взаимное выравнивание многих последовательностей.

```
no.sooner.---met.---------but.they.-look'd
no.sooner.look'd.---------but.they.-lo-v'd
no.sooner.lo-v'd.---------but.they.-sigh'd
no.sooner.sigh'd.---------but.they.--asked.one.another.the.reason
no.sooner.knew.the.reason.but.they.-------------sought.the.remedy
no.sooner.                .but.they.
```

Последняя строка показывает символы, сохраненные во всех последовательностях выравнивания.

См. гл. 4 где продолжается обсуждение выравнивания.

```
LEDVKAVCSQKNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTS
QKEKHIIVACEGNPYVPVHFDNSV
>RNP_MACRU

ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPK
SVVDAVCHQENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSN
LNKQIIVACEGQYVPVHFDAYV
```

Построим для этих последовательностей множественное выравнивание с помощью программы CLUSTAL–W `http://www.ebi.ac.uk/clustalw/` (или в программу T-coffee: `http://www./ch.embnet.org/software/TCoffee.html`)

Программы выдадут следующее множественное выравнивание:

```
CLUSTAL W (1.8) multiple sequence alignment

RNP_HORSE      KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEPLADVQAICLQ 60
RNP_BALAC      RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHESLEDVKAVCSQ 60
RNP_MACRU      -ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPKSVVDAVCHQ 59
                  *:** **:*****:  :......*** **  *.**.* ***:***:**.   *.*:* *

RNP_HORSE      KNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTSQKERHIIVACEGNPYVPVHF 120
RNP_BALAC      KNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTSQKEKHIIVACEGNPYVPVHF 120
RNP_MACRU      ENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSNLNKQIIVACEG-QYVPVHF 118
                  :*: ****::***:*.* : **:** *..****** *:**: :::********  ******

RNP_HORSE        DASVEVST 128
RNP_BALAC        DNSV-- 124
RNP_MACRU        DAYV-- 122
                 *  *
```

В этой таблице «*» под последовательностями обозначает неизменную (одинаковую во всех последовательностях) аминокислоту, и : и . указывают позиции, в которых все последовательности содержат остатки очень сходного физико-химического характера (:) или довольно сходного физико-химического характера (.).

Большие куски последовательностей идентичны. Есть большое число замещений, но только одно внутреннее удаление. При попарном сравнивании последовательностей, число идентичных остатков между парами в этом выравнивании (это не то же самое, что подсчет звездочек) следующее:

Число идентичных остатков в выровненных последовательностях (из общего количества 122–128 остатков)

Лошадь и малый полосатик	95
Малый полосатик и большой коричневый кенгуру	82
Лошадь и большой коричневый кенгуру	75

Лошадь и кит имеют больше идентичных остатков. Результат представляется значимым, и, кроме того, подтверждает наши ожидания. *Предупреждение: может быть мы путаем причину и следствие?*

Попробуем более сложный пример:

ПРИМЕР 1.3.

Два ныне живущих рода слонов представлены африканским слоном (*Loxodonta Africana*) и индийским слоном (*Elephas maximus*). Оказалось возможным получить последовательность митохондриального цитохрома *b* от экземпляра сибирского шерстистого мамонта (*Mammuthus primigenius*), сохранившегося в арктической вечной мерзлоте.

Отыскиваем последовательности и запускаем CLUSTAL-W:

```
African elephant  MTHIRKSHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60
Siberian mammoth  MTHIRKSHPLLKILNKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60
Indian elephant   MTHTRKSHPLFKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60
                  *** ******:**:**************************************************

African elephant  TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
Siberian mammoth  TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
Indian elephant   TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
                  ************************************************************

African elephant  LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPCIGTNLVEWIWGGFSVDKATLNRFFA 180
Siberian mammoth  LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTDLVEWIWGGFSVDKATLNRFFA 180
Indian elephant   LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA 180
                  ********************************** ***:**********************

African elephant  LHFILPFTMIALAGVHLTFLHETGSNNPLGLISDSDKIPFHPYYTIKDFLGLLILILLLL 240
Siberian mammoth  LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILFLL 240
Indian elephant   FHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL 240
                  :********:********************** ************************:**

African elephant  LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI 300

Siberian mammoth  LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI 300
Indian elephant   LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSILI 300
                  ************************************************************:******

African Elephant  LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEYPYIIIGQMASILYFS 360
Siberian mammoth  LGIMPLLHTSKHRSMMLRPLSQVLFWTLATDLLMLTWIGSQPVEYPYIIIGQMASILYFS 360
Indian elephant   LGLMPFLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFS 360
                  **:**:**********************: *** ************** *************

African elephant  IILAFLPIAGVIENYLIK 378
Siberian mammoth  IILAFLPIAGMIENYLIK 378
Indian elephant   IILAFLPIAGMIENYLIK 378
                  **********:*******
```

Последовательности мамонта и африканского слона имеют 10 несовпадений, а последовательности мамонта и индийского слона имеют 14 несовпадений. Оказывается, что мамонт ближе к африканскому слону. Однако, этот результат менее удовлетворителен, чем предыдущий. Есть несколько различий. Являются ли они существенными? (Трудно решить, являются ли отличия существенными, потому что у нас нет никаких предварительных идей о том, каким должен быть ответ.)

Этот пример вызывает ряд вопросов.

1. Мы «знаем», что африканский и индийский слоны и мамонты должны быть близкими родственниками — для этого достаточно простого взгляда. Но можем ли мы сказать *только из этих последовательностей*, что они принадлежат близким видам?
2. Принимая утверждение о том, что различия малы, задаемся вопросом, представляют ли они собой эволюционные отклонения, возникшие из отбора, или просто случайный шум или случайное отклонение? Нам нужен чувствительный статистический критерий для определения значимости совпадений и различий.

Для пояснения данных вопросов, давайте отметим разницу между *сходством* (*similarity*) и *гомологией* (*homology*). *Сходство* — это наличие или измерение сходства и различия, независимо от источника сходства. *Гомология* означает, конкретно, что последовательности и организмы, в которых они обнаружены, являются потомками общего предка, при этом предполагается, что подобные характеристики имели и предки. Подобие последовательностей (или макроскопических биологических характеристик) можно наблюдать в информации, которую собирают *сейчас*, и при этом не подразумеваются никакие исторические гипотезы. Наоборот, утверждение о гомологии — это утверждение исторических событий, которые почти всегда необозримы. Гомология должна быть *предположением*, возникающим из наблюдения подобия. Только в некоторых немногочисленных случаях гомология может быть непосредственно наблюдаема: например, в фамильной родословной, демонстрирующей необычный фенотип, как например, губа Габсбургов, или в лабораторной популяции, или в клинических испытаниях, в курсе наблюдения за вирусными инфекциями на уровне последовательностей у индивидуальных пациентов.

Утверждение, что цитохромы b африканского и индийского слонов и мамонтов гомологичны, *означает*, что существовал общий предок, который, вероятно, содержал уникальный цитохром b, который путем альтернативных мутаций дал начало белкам мамонтов и современных слонов. Доказывает ли высокая степень сходства последовательностей утверждение о том, что они гомологичны, или есть другие объяснения?

- Может быть такое, что функциональный цитохром b содержит так много консервативных участков, что цитохромы b других животных так же похожи друг на друга, как и цитохромы слона и мамонта. Мы можем проверить это, изучив последовательности этого белка других видов. В ре-

зультате цитохромы *b* других животных достаточно сильно отличаются от цитохромов слонов и мамонтов.

- Второй вариант состоит в том, что есть специальные условия для хорошего функционирования цитохрома *b* у слоноподобных животных и что три последовательности цитохрома *b* идут от трех самостоятельными предков, а общее избирательное воздействие вынудило их стать похожими. (Помните, мы спрашиваем, что может быть выведено, только из анализа последовательностей цитохромов *b*).
- Мамонт может быть более близким родственником африканского слона, но со времени последнего общего предка последовательность цитохрома *b* индийского слона эволюционировала быстрее, чем последовательности африканского слона и мамонта, накапливая больше мутаций.
- До сих пор есть четвертая гипотеза о том, что все общие предки слонов и мамонтов имели сильно различающиеся цитохромы *b*, но жившие слоны и мамонты размножили общий ген путем переноса из неродственных организмов с помощью вирусов.

Допустим, однако, мы доказали, что сходство последовательностей цитохрома *b* у слона и мамонта может быть достаточным доказательством гомологии, но как тогда насчет последовательностей рибонуклеаз в предыдущем примере? Являются ли *большие* различия панкреатических рибонуклеаз лошади, кита и кенгуру доказательством того, что они *не* гомологичны?

Как мы сейчас можем ответить на эти вопросы? Специалисты взялись аккуратно откалибровать сходства и различия последовательностей по многим белкам из многих видов, для которых таксономическое положение было установлено классическими методами. В примере с панкреатическими рибонуклеазами рассуждения от сходства к гомологии оправданы. Вопрос о том, ближе мамонты к африканским или индийским слонам, еще не разрешен, даже используя все имеющиеся анатомические доказательства и сходство последовательностей. Сейчас анализы сходства последовательностей полностью признаны и считается, что это наиболее надежные методы установления филогенетического родства, несмотря даже на то, что иногда — как на примере со слонами — результаты могут не быть достоверными, а в других случаях даже давать неправильные ответы. Есть множество доступных данных, эффективные инструменты для извлечения того, что необходимо для решения специфического вопроса, а также многочисленные инструменты для анализа. Но ничто это не заменяет необходимость содержательного научного обсуждения.

Использование SINE и LINE для установления филогенетического родства

Основные проблемы, связанные с установлением филогении на основе выравниваний (сравнения) нуклеотидных и белковых последовательностей, это: (1) широкая область вариантов сходства, которые могут не быть статистически значимыми и (2) эффекты разной скорости эволюции в разных ветвях эволюционного дерева. Во многих случаях, даже если сходство после-

довательностей точно устанавливает родство, может быть невозможно установить *порядок*, в котором разделился набор таксонов. Мечта филогенетиков — признак, который имеет характер «все или ничего», и появление которого фиксировано, так что может быть установлен порядок расхождения. В некоторых случаях эта мечта представлена точными некодирующими участками последовательностей в геномах.

SINE и LINE (Short and Long Interspersed Nuclear Elements, короткие и длинные распределенные генетические элементы) — это повторяющиеся некодирующие последовательности, которые образуют большую часть генома эукариотов, как минимум, 30% хромосомной ДНК человека и более 50% генома некоторых высших растений. Обычно SINE состоят ~ из 70–500 пар оснований, может появиться до 10^6повторений. LINE могут быть до 7000 пар оснований, и может появиться до 10^5 повторений. SINE появляются в геноме путем обратной транскрипции РНК. Большинство SINE содержат $5'$ участок, гомологичный тРНК (или другой некодирующей функциональной РНК), середину, негомологичную тРНК, и АТ-богатый $3'$-участок.

Характерные особенности SINE, позволяющие использовать их при изучении филогении:

- SINE присутствует или отсутствует. Присутствие SINE в каких-либо специфических позициях — это свойство, которое не влечет за собой сложного и непостоянного измерения сходства.
- SINES случайным образом вставлены в некодирующие части генома. Следовательно, появление похожего SINES в одном и том же локусе у двух видов означает, что у этих видов есть общий предок, у которого произошла вставка. Аналоги конвергентной эволюции не замутнят эту картину, так как нет отбора на участок вставки.
- Появление вставки SINE необратимо: механизмы *утраты* SINE неизвестны, кроме редких крупных делеций участков, которые включают SINE. Следовательно, если два вида имеют SINE в одинаковом локусе, то *отсутствие* этого SINE у третьего вида означает, что первые два вида ближе друг к другу, чем какой — либо из них к третьему.
- SINE не только показывают родство, они показывают, какие виды появились первыми. Последний общий предок видов, содержащих некоторый конкретный SINE должен появиться *позже* последнего общего предка, связывающего эти виды и у которого отсутствует этот SINE.

Н. Окада и коллеги использовали последовательности SINE, для исследования филогении.

Киты, подобные австралийским, — это млекопитающие, которые приспособились к водному образу жизни. Но кто — в случае с китами — их ближайшие сухопутные родственники? Классическая палеонтология связывает отряд китообразные — содержащий китов, дельфинов и морских свиней — с отрядом Arteriodactyla — парнокопытные (включающим, например, крупный рогатый скот). Считается, что китообразные дивергировали ранее общего предка трех сохранившихся подотрядов Arteriodactyla: Suiformes (свиньи), мозоленогие (включает верблюдов и лам) и жвачные (включающие оленей, крупный ро-

Рис. 1.5. Филогенетическое родство между китовыми и другими подгруппами arteriodactyl, установленное анализом последовательностей SINE. Небольшие стрелки показывают вставку. Каждая стрелка показывает наличие особых SINE и LINE в специфических локусах во всех видах справа от стрелки. Строчные буквами названы локусы, прописными названы паттерны последовательностей. Например, паттерн ARE2 появляется только у свиней в локусе ino. Паттерн ARE появляется дважды в геноме свиньи, в локусах gpi и pro, и в геноме пекари в этих же локусах. Вставка ARE происходит у видов, родственных свиньям и пекари, но не у других видов на диаграмме. Это означает, что свиньи и пекари более близки друг к другу, чем к каким-либо другим исследуемым животным. (Nikaido, M., Rooney, A. P. и Okada, N. (1999) 'Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales', Proceedings of the National Academy of Sciences USA 96, 10261–6. (© 1999, National Academy of Sciences, USA))

гатый скот, овец, жирафов и т. д.). Чтобы правильно поместить китообразных среди этих групп, осуществили несколько исследований с последовательностями ДНК. Сравнения митохондриальной ДНК, генов панкреатической рибонуклеазы, γ-фибриногена и других белков говорят о том, что ближайшими родственниками китов являются гиппопотамы и что китообразные с гиппопотамами образуют отдельную группу внутри Arteriodactyla, наиболее близкая к жвачным. (см. Интернет задание 1.7).

Анализы SINE подтверждают это родство. Некоторые SINE являются общими у жвачных, гиппопотамов и китовых. Четыре типа SINE встречаются только у гиппопотамов и китовых. Эти наблюдения предполагают дерево, показанное на рис. 1.5, на котором помечены вставки SINE. [Добавленные в подтверждение замечания: новые ископаемые сухопутных предков китов подтверждают связь между китами и *arteriodactyls*.Это хороший пример взаимного дополнения молекулярных и палеонтологических методов: анализы последовательностей ДНК могут устанавливать родственные связи между живущими видами довольно точно, но только сравнение с ископаемыми может подтвердить родство с вымершими видами.]

Поиск схожих последовательностей в базах данных: PSI-BLAST

Примеры, к которым мы обратились, имеют общую тематику — поиск в базе данных объектов, похожих на имеющиеся у нас. Например, если вы определили последовательность нового гена или нашли в геноме человека ген, ответственный за какое-то заболевание, то вы, возможно, захотите узнать, нет ли таких генов у других видов. Идеальный метод — тот, который как чувствителен (который определяет даже дальнее родство), так и селективен (благодаря которому все полученные родственные связи — истинные).

Методы поиска в базах данных подразумевают компромисс между чувствительностью и селективностью. Находит ли метод все или большинство из «хитов», которые на самом деле существуют, или же он упускает большую часть? А также, сколько из выданных этим методом «хитов» неправильные? Предположим, база данных содержит 1000 последовательностей глобина. Предположим, поиск в этой базе данных по глобинам выдал 900 находок, 700 из них действительно последовательности глобина, а 200 таковыми не являются. Про такой поиск можно сказать, что у него 300 ложных отрицательных результатов (упущенных последовательностей) и 200 ложных положительных результатов. Уменьшая порог допустимости, мы получим меньше ложных отрицательных результатов, но больше ложных положительных результатов. Часто лучше работать с низкими порогами, чтобы быть уверенным, что ничего, что могло бы быть важным не утеряно; но тогда потребуется детальная проверка результатов, для того чтобы устранить ложные находки.

Мощным инструментом для поиска последовательностей в базах данных, по имеющейся у нас последовательности, является PSI-BLAST, из US National Center for Biotechnological Information (NCBI). Аббревиатура PSI -BLAST расшифровывается как «Position Sensitive Iterated — Basic Linear Alignment Sequence Tool». Более ранняя программа, BLAST, работала, определяя локальные участки сходства без делеций, а затем объединяя их. Аббревиатура PSI в PSI-BLAST указывает на улучшения, которые устанавливают паттерны[1)] в последовательности на предварительных стадиях поиска в базе данных, а затем последовательно улучшают их. Опознавание консервативных паттернов может улучшить как селективность, так и чувствительность поиска. PSI-BLAST представляет собой повторяющийся (итерационный) процесс, в котором появившийся паттерн уточняется в последующих стадиях поиска.

ПРИМЕР 1.4.

Гомологи PAX-6 гена человека. PAX-6 гены контролируют развитие глаза в широком наборе видов (см. Дополнение на с. 50). PAX-6 ген человека кодирует белок, занесенный в базу данных SWISS-PROT запись P26367.

[1)] В русском языке нет адекватного перевода этого слова в контексте анализа последовательностей. Под паттерном в биоинформатике понимается общее свойство, например подпись. — *Прим. ред.*

Для того чтобы запустить PSI-BLAST, используйте следующую ссылку: `http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi`.

Введите последовательность и используйте опции по умолчанию для поиска в базах данных и используемой матрицы попарного сходства.

Программа выдает список записей схожих с последовательностью, заданной для поиска, сортированный в порядке убывания статистической значимости. (Извлечения из выдачи программы показаны в дополнении: результаты поиска PSI-BLAST для PAX-6 белка человека.) Обычно строка выглядит так:

```
pir||I45557  eyeless, long form - fruit fly (Drosophila melano...   255  7e-67
```

Первое значение в строке база данных и номер записи в ней (разделены ||) в данном случае запись I45557 в базе данных PIR (Protein Identification Resource). Это гомолог *eyeless Drosophila*. Число 255 количество очков, присвоенное обнаруженному совпадению, и значимость данного совпадения измерена как $E = 7 \times 10^{-67}$. E определяется вероятностью того, что данная степень сходства может быть случайной: E это ожидаемое количество последовательностей, которые совпадут также или лучше чем данная, если поиск будет производиться базе данных такого же размера, но со случайными последовательностями. $E = 7 \times 10^{-67}$ означает, что крайне невероятно, что даже одна случайная последовательность совпадет так как гомолог Дрозофилы. Значения E ниже 0.05 могут рассматриваться как значимые: по крайней мере, их стоит рассмотреть. При граничных случаях, следует задать вопросы: «являются ли несовпадения консервативными? Существует ли какой-нибудь паттерн, или распределены ли совпадения и несовпадения беспорядочно по последовательностям?» Существует размытое понятие, «текстура выравнивания», которое вы поймете на практике.

Заметьте, что если в банке данных много последовательностей, очень похожих на ту, по которой мы производим поиск, то они будут возглавлять список. В данном случае у других млекопитающих много очень похожих PAX генов. Вам, возможно, придется далеко пролистать список, чтобы найти, показавшегося вам интересным, дальнего родственника.

Фактически программа сопоставила только часть последовательностей. Полное выравнивание показано в Дополнении: Полное парное выравнивание последовательности PAX-6 белка человека и eyeless белка Drosophila melanogaster. (см. упр. 1.5)

ПРИМЕР 1.5.

Какие особи содержат гомологов PAX-6 человека, найденных PSI-BLAST'ом?

PSI-BLAST сообщает названия особей, в которых присутствуют найденные последовательности (см. Дополнение: «Результаты поиска PSI-BLAST

Дополнительные сведения о гене PAX

Глаза человека, мухи и осьминога сильно различаются по строению. Общепринятая мудрость, принимая во внимание безграничную селективную пользу, дарованную способностью видеть, полагала, что глаза возникли независимо в каждой эволюционной ветви. Поэтому большим сюрпризом стал тот факт, что ген, контролирующий развитие человеческого глаза, имеет гомолога, управляющего развитием глаза дрозофилы.

Ген PAX-6 был клонирован вначале у мыши и человека. Он является главным регуляторным геном, контролируя комплекс каскада событий в развитии глаза. Мутации в гене человека вызывают клиническое состояние — аниридию: дефект в развитии глаза, при котором радужная оболочка отсутствует или деформирована. Гомолог гена PAX-6 в дрозофиле называется eyeless геном (имеет сходную функцию контроля развития глаза). Мухи, мутантные по этому гену, развиваются без глаз; и обратно, экспрессия этого гена на лапке мухи или на антенне мухи — вызывает появление эктопических (= находящихся не на месте) глаз. Дрозофила, мутантная по гену eyeless, была впервые описана в 1915 г. Никто и не подозревал о его родстве с геном млекопитающих.

Гены насекомого и млекопитающего схожи не только по последовательности, они так близкородственны, что их активность выходит за рамки видов. Экспрессия мышиного PAX-6 в мухе вызывает эктопическое развитие глаза, также как и собственный eyeless ген мухи.

Гомологи PAX-6 представлены и в других классах, включая плоских червей, асцидий, морских ежей и нематод. Наблюдение, что родопсины (семейство белков, содержащих ретин в качестве хромофора) функционируют, как светочувствительные пигменты в различных классах организмов, является дополнительным доказательством общего происхождения различных систем фоторецепторов. Настоящие структурные различия в макроскопическом строении различных глаз отражают дивергенцию и независимость развития высокоорганизованных структур.

по белку PAX-6 человека»). Они вставлены в текст вывода в квадратных скобках; например:

```
emb|CAA56038.1| (X79493) transcription factor [Drosophila melanogaster]
```

(В секции, содержащей *E*-values, названия особей могут быть оборваны.) Следующая программа на языке PERL, извлекает названия видов из вывода PSI-BLAST.

```
#!/usr/bin/perl
#извлекает названия особей из вывода psiblast
```

```
# Метод:
#   Для каждой строки ввода, проверить на наличие паттерна формы [Drosophila melanogaster]
#   Использовать найденный паттерн как индекс в связанном множестве
#   Значение, соответствующее данному индексу, не важно
#   Используя связанное множество, последующие названия тех же особей запишутся на
#   место первого названия, сохраняя только уникальный набор
#   После завершения обработки отсортировать результаты и вывести на экран.

while (<>) {                              # прочитать строку ввода
    if (/\[([A-Z][a-z]+ [a-z]+)\]/) {     # выделить строки, содержащие параметр вида
                                          #      [Drosophila melanogaster]
     $species{$1} = 1;                    # создать или перезаписать запись
     }                                    #          в связанном множестве
}

foreach (sort(keys(%species))){           # в~алфавитном порядке,
    print "$_\n";                         #    вывести названия особей на экран
}
```

Найдено 52 биологических вида (см. с. 52: *Виды, распознанные PSI-BLAST'ом*).

Программа использует богатые ресурсы узнавания паттернов (образцов) PERL'а для поиска последовательности букв *вида* `[Drosophila melanogaster]`. Рассмотрим следующий паттерн:

- квадратная скобка,
- затем — слово, начинающееся с заглавной буквы, после которой может стоять произвольное число строчных букв,
- затем — пробел между словами,
- затем — слово, целиком состоящее из строчных букв,
- затем — закрывающая квадратная скобка.

Этот вид паттерна называется *регулярным выражением* и появляется в программе PERL в следующей форме: `[([A-Z][a-z]+ [a-z]+)]`

При построении образца используются следующие обозначения:

`[A-Z]` = любая буква в ряду A, B, C, ... Z

`[a-z]` = любая буква в ряду a, b, c, ... z

Мы можем указывать повторения:

`[A-Z]` = *ровно одна* буква в верхнем регистре

`[a-z]+` = *одна или более* букв в нижнем регистре

и скомбинировать результаты:

`[A-Z][a-z]+ [a-z]+` = заглавная буква, затем одна или более строчных букв (название рода), затем пробел, затем одна или более строчных букв (название вида).

Заключение этого выражения в круглые скобки: `([A-Z][a-z]+ [a-z]+)` дает команду PERL'у сохранить материал, удовлетворяющий паттерну, для дальнейшей ссылки на него. В PERL этот подобранный материал обозначается переменной $1. Таким образом, если входная строка содержала `[Drosophila melanogaster]`, выражение

```
$species{$1} = 1;
```

в действительности будет:

```
$species{"Drosophila melanogaster"} = 1;
```

Наконец, мы хотим включить скобки, окружающие названия рода и вида, но скобки указывают на ряд из букв. Поэтому мы должны поставить обратный слеш (косую черту) перед скобками: `\[...\]`, чтобы получить конечный паттерн: `\[([A-Z][a-z]+ [a-z]+)\]`

Использование ассоциативного массива (хэш таблицы) для сохранения только уникального набора видов — еще одно средство, встроенное в язык программирования PERL. Ассоциативный массив есть обобщение обычного массива или вектора, в котором элементы индексированы не целыми числами, а произвольными последовательностями. Вторая ссылка на ассоциированный элемент с прежде встречавшейся индексной строкой может изменить значение в массиве, но не список индексных строк. В этом случае мы не заботимся о значении, а просто используем индексные строки, чтобы составить уникальный список видов, которые были обнаружены. Множественные ссылки на одни и те же виды будут просто переписывать первую ссылку, а не делать повторяющийся список.

Виды, найденные PSI-BLAST при исследовании последовательности human PAX-6

```
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro
A. Schaeffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David
J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= sp|P26367|PAX6_HUMAN PAIRED BOX PROTEIN PAX-6
(OCULORHOMBIN) (ANIRIDIA, TYPE II PROTEIN) - Homo sapiens (Human).
        (422 letters)

              Sequences with E-value BETTER than threshold

                                                                  Score  E
Sequences producing significant alignments:                       (bits) Value

 ref|NP_037133.1|  paired box homeotic gene 6 >gi|2495314|sp|P7...  730    0.0
 ref|NP_000271.1|  paired box gene 6, isoform a >gi|417450|sp|P...  730    0.0
 pir||A41644  homeotic protein aniridia - human                     728    0.0
 gb|AAA59962.1|  (M77844) oculorhombin [Homo sapiens] >gi|18935...  728    0.0
 prf||1902328A  PAX6 gene [Homo sapiens]                            724    0.0
 emb|CAB05885.1|  (Z83307) PAX6 [Homo sapiens]                      723    0.0
 ref|NP_001595.2|  paired box gene 6, isoform b                     721    0.0
 ref|NP_038655.1|  paired box gene 6 >gi|543296|pir||S42234 pai...  721    0.0
 dbj|BAA23004.1|  (D87837) PAX6 protein [Gallus gallus]             717    0.0
 gb|AAF73271.1|AF154555_1  (AF154555) paired domain transcripti...  714    0.0
 sp|P55864|PAX6_XENLA  PAIRED BOX PROTEIN PAX-6 >gi|1685056|gb|...  713    0.0
 gb|AAB36681.1|  (U76386) paired-type homeodomain Pax-6 protein...  712    0.0
 gb|AAB05932.1|  (U64513) Xpax6 [Xenopus laevis]                    712    0.0
 sp|P47238|PAX6_COTJA  PAIRED BOX PROTEIN PAX-6 (PAX-QNR) >gi|4...  710    0.0
 dbj|BAA24025.1|  (D88741) PAX6 SL [Cynops pyrrhogaster]            707    0.0
 gb|AAD50903.1|AF169414_1  (AF169414) paired-box transcription ...  706    0.0
```

```
dbj|BAA13680.1|  (D88737) Xenopus Pax-6 long [Xenopus laevis]        703    0.0
sp|P26630|PAX6_BRARE  PAIRED BOX PROTEIN PAX[ZF-A] (PAX-6) >gi...   699    0.0
dbj|BAA24024.1|  (D88741) PAX6 LL [Cynops pyrrhogaster]              697    0.0
gb|AAD50901.1|AF169412_1  (AF169412) paired-box transcription ...    696    0.0
emb|CAA68835.1|  (Y07546) PAX-6 protein [Astyanax mexicanus] >...    693    0.0
pir||I50108  paired box transcription factor Pax-6 - zebra fis...    689    0.0
sp|O73917|PAX6_ORYLA  PAIRED BOX PROTEIN PAX-6 >gi|3115324|emb...    686    0.0
gb|AAC96095.1|  (AF061252) Pax-family transcription factor 6.2...    684    0.0
emb|CAA68837.1|  (Y07547) PAX-6 protein [Astyanax mexicanus]         683    0.0
emb|CAA68836.1|  (Y07547) PAX-6 protein [Astyanax mexicanus]         675    0.0
emb|CAA68838.1|  (Y07547) PAX-6 protein [Astyanax mexicanus]         675    0.0
emb|CAA16493.1|  (AL021531) PAX6 [Fugu rubripes]                     646    0.0
gb|AAF73273.1|AF154557_1  (AF154557) paired domain transcripti...    609    e-173
dbj|BAA24023.1|  (D88741) PAX6 SS [Cynops pyrrhogaster]              609    e-173
prf||1717390A  pax gene [Danio rerio]                                609    e-173
gb|AAF73268.1|AF154552_1  (AF154552) paired domain transcripti...    608    e-173
gb|AAD50904.1|AF169415_1  (AF169415) paired-box transcription ...    605    e-172
gb|AAF73269.1|AF154553_1  (AF154553) paired domain transcripti...    604    e-172
dbj|BAA13681.1|  (D88738) Xenopus Pax-6 short [Xenopus laevis]       600    e-171
dbj|BAA24022.1|  (D88741) PAX6 LS [Cynops pyrrhogaster]              599    e-170
gb|AAD50902.1|AF169413_1  (AF169413) paired-box transcription ...    595    e-169
gb|AAF73270.1|  (AF154554) paired domain transcription factor ...    594    e-169
gb|AAB07733.1|  (U67887) XLPAX6 [Xenopus laevis]                     592    e-168
gb|AAA40109.1|  (M77842) oculorhombin [Mus musculus]                 455    e-127
emb|CAA11364.1|  (AJ223440) Pax6 [Branchiostoma floridae]            440    e-122
emb|CAA11366.1|  (AJ223442) Pax6 [Branchiostoma floridae]            437    e-122
gb|AAB40616.1|  (U59830) Pax-6 [Loligo opalescens]                   437    e-122
pir||A57374  paired box transcription factor Pax-6 - sea urchi...    437    e-121
emb|CAA11368.1|  (AJ223444) Pax6 [Branchiostoma floridae]            435    e-121
emb|CAA11367.1|  (AJ223443) Pax6 [Branchiostoma floridae]            433    e-120
emb|CAA11365.1|  (AJ223441) Pax6 [Branchiostoma floridae]            412    e-114
pir||JC6130  paired box transcription factor Pax-6 - Ribbonwor...    396    e-109
gb|AAD31712.1|AF134350_1  (AF134350) transcription factor Toy ...    380    e-104
gb|AAB36534.1|  (U77178) paired box homeodomain protein TPAX6 ...    377    e-104
emb|CAA71094.1|  (Y09975) Pax-6 [Phallusia mammilata]                342    4e-93
dbj|BAA20936.1|  (AB002408) mdkPax-6 [Oryzias sp.]                   338    6e-92
pir||S60252  paired box transcription factor vab-3 - Caenorhab...    336    2e-91
pir||T20900  hypothetical protein F14F3.1 - Caenorhabditis ele...    336    2e-91
pir||S36166  paired box transcription factor Pax-6 - rat (frag...    335    5e-91
sp|P47237|PAX6_CHICK  PAIRED BOX PROTEIN PAX-6 >gi|2147404|pir...    333    2e-90
dbj|BAA75672.1|  (AB017632) DjPax-6 [Dugesia japonica]               329    4e-89
gb|AAF64460.1|AF241310_1  (AF241310) transcription factor PaxB...    290    2e-77
gb|AAF73274.1|  (AF154558) paired domain transcription factor ...    287    1e-76
pdb|6PAX|A  Chain A, Crystal Structure Of The Human Pax-6 Pair...    264    1e-69
pir||C41061  paired box homolog Pax6 - mouse (fragment)              261    9e-69
gb|AAC18658.1|  (U73855) Pax6 [Bos taurus]                           259    4e-68
pir||I45557  eyeless, long form - fruit fly (Drosophila melano...    255    7e-67
gb|AAF59318.1|  (AE003843) ey gene product [Drosophila melanog...    255    7e-67

                          ...many additional «hits» deleted ...

                          ...two selected alignments follow ...

----------------------------------------
                              Alignments

>ref|NP_037133.1| paired box homeotic gene 6
 sp|P70601|PAX6_RAT PAIRED BOX PROTEIN PAX-6
```

```
 gb|AAB09042.1| (U69644) paired-box/homeobox protein [Rattus norvegicus]
          Length = 422

 Score =  730 bits (1865), Expect = 0.0
 Identities = 362/422 (85%), Positives = 362/422 (85%)

Query: 1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY 60
           MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY
Sbjct: 1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY 60

Query: 61  YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV 120
           YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV
Sbjct: 61  YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV 120

Query: 121 SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTXXXXXXX 180
           SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPT
Sbjct: 121 SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ 180

Query: 181 XXXXXNTNSISSNGEDSDEAQMXXXXXXXXXXNRTSFTQEQIEALEKEFERTHYPDVFAR 240
                NTNSISSNGEDSDEAQM          NRTSFTQEQIEALEKEFERTHYPDVFAR
Sbjct: 181 EGQGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR 240

Query: 241 ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNXXXXXXXXXXXXXXXVYQPIP 300
           ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASN               VYQPIP
Sbjct: 241 ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP 300

Query: 301 QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT 360
           QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT
Sbjct: 301 QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT 360

Query: 361 SPSVNGRSYDTYTPPHMQTHMNSQPMXXXXXXXXXLIXXXXXXXXXXXXXXXDMSQYWPR 420
           SPSVNGRSYDTYTPPHMQTHMNSQPM         LI               DMSQYWPR
Sbjct: 361 SPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR 420

Query: 421 LQ 422
           LQ
Sbjct: 421 LQ 422

>pir||I45557 eyeless, long form - fruit fly (Drosophila melanogaster)
 emb|CAA56038.1| (X79493) transcription factor [Drosophila melanogaster]
          Length = 838

 Score =  255 bits (644), Expect = 7e-67
 Identities = 124/132 (93%), Positives = 128/132 (96%)

Query: 5   HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 64
           HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct: 38  HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG 97

Query: 65  SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN 124
           SIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E VCTNDNIPSVSSIN
Sbjct: 98  SIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSIN 157

Query: 125 RVLRNLASEKQQ 136
           RVLRNLA++K+Q
Sbjct: 158 RVLRNLAAQKEQ 169
```

Полное попарное выравнивание последовательности человеческого белка PAX-6, и белка, кодированного геном eyeless из *Drosophila melanogaster*

```
PAX6_human      ----------------MQNSHSGVNQLGGVFVNGRPLPDSTRQ 27
eyeless         MFTLQPTPTAIGTVVPPWSAGTLIERLPSLEDMAHKGHSGVNQLGGVFVGGRPLPDSTRQ 60
                                        ::.************.**********

PAX6_human      KIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKI 87
eyeless         KIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKI 120
                ************************************************.******

PAX6_human      AQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQ------ 136
eyeless         SQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQKEQQSTGSGSSSTS 180
                :*******************.*.**********************::*:*

PAX6_human      ------MG----------------------ADG 141
eyeless         AGNSISAKVSVSIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGATNSGEGSEQEA 240
                      :*                                            :.

PAX6_human      MYDKLRMLNGQTGS----------WGTRP----------- 160
eyeless         IYEKLRLLNTQHAAGPGPLEPARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQQHQQQSW 300
                :*:***:** * .:                    ***.

PAX6_human      ----GWYPG----TSVP---------------GQP-- 172
eyeless         PPRHYSGSWYPTSLSEIPISSAPNIASVTAYASGPSLAHSLSPPNDIKSLASIGHQRNCP 360
                      .***        :*.*                                  *:

PAX6_human      -----TQDGCQQQEGG--GENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQ 219
eyeless         VATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDDQARLILKRKLQRNRTSFTN 420
                           ** *.:* *   ***:*. :**  :::: * ** **************:

PAX6_human      EQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQAS 279
eyeless         DQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEKLRNQRRTPN 480
                :**::**********************.**.**************************** ..

PAX6_human      NTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLG------------ 316
eyeless         STGASATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAPT 540
                .* :  . **: :*:     *:. :. **: ***  *

PAX6_human      ------------------------------
eyeless         LGAGIDSSESPTPIPHIRPSCTSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHA 600


PAX6_human      -------------RTDTALTNTYSALPPMPSFTMANNLPMQPPVP 348
eyeless         QGHALVPAISPRLNFNSGSFGAMYSNMHHTALSMSDSYGAVTPIPSFNHSAVGPLAPPSP 660
                                     :*  :::::*.*:.*:***. :   *: ** *

PAX6_human      S----QTSSYSCMLPTSP-----------------SVNGRS 368
eyeless         IPQQGDLTPSSLYPCHMTLRPPPMAPAHHHIVPGDGGRPAGVGLGSGQSANLGASCSGSG 720
                        :* *.* :.  *                                 * .* .

PAX6_human      YDTYTP---------------PHMQTHMNSQP-----MGTS 389
eyeless         YEVLSAYALPPPPMASSSAADSSFSAASSASANVTPHHTIAQESCPSPCSSASHFGVAHS 780
                *:. :.                                **     :* *          :. *

PAX6_human      GTTSTGLISPGVS--------VPVQVPGS--EPDMSQYWPRLQ--- 422
eyeless         SGFSSDPISPAVSSYAHMSYNYASSANTMTPSSASGTSAHVAPGKQQFFASCFYSPWV 838
                .  *:: ***.**               .* ::.*:     *. .*::.
```

Виды, найденные PSI-BLAST при исследовании последовательности human PAX-6.

Acropora millepora	Herdmania curvata
Archegozetes longisetosus	Homo sapiens
Astyanax mexicanus	Hydra littoralis
Bos taurus	Hydra magnipapillata
Branchiostoma floridae	Hydra vulgaris
Branchiostoma lanceolatum	Ilyanassa obsoleta
Caenorhabditis elegans	Lampetra japonica
Canis familiaris	Lineus sanguineus
Carassius auratus	Loligo opalescens
Chrysaora quinquecirrha	Mesocricetus auratus
Ciona intestinalis	Mus musculus
Coturnix coturnix	Notophthalmus viridescens
Cynops pyrrhogaster	Oryzias latipes
Danio rerio	Paracentrotus lividus
Drosophila mauritiana	Petromyzon marinus
Drosophila melanogaster	Phallusia mammilata
Drosophila sechellia	Podocoryne carnea
Drosophila simulans	Ptychodera flava
Drosophila virilis	Rattus norvegicus
Dugesia japonica	Schistosoma mansoni
Ephydatia fluviatilis	Strongylocentrotus purpuratus
Fugu rubripes	Sus scrofa
Gallus gallus	Takifugu rubripes
Girardia tigrina	Tribolium castaneum
Halocynthia roretzi	Triturus alpestris
Helobdella triserialis	Xenopus laevis

Структуры белков. Введение

При переходе к белковым структурам оставляем позади «одномерность» нуклеотидных и аминокислотных последовательностей и открываем мир пространственных молекулярных структур. Один из путей хранения и поиска молекулярной биологической информации, сохранить этот переход адекватным. При этом кое-что в способе хранения информации должно быть в значительной степени изменено, а кое-что остаться без изменений.

Белки играют целый ряд ролей в процессах жизнедеятельности: есть структурные белки (например, белки оболочек вирусов, белки ороговевшего внешнего слоя кожи человека и животных, белки цитоскелета); белки, катализирующие химические реакции (ферменты); транспортные и информационные белки (гемоглобин); регуляторные белки, включая гормоны и рецепторные белки; белки, контролирующие генетическую транскрипцию; белки, участвующие в узнавании, включая клеточную адгезию, антитела, и другие белки иммунной системы.

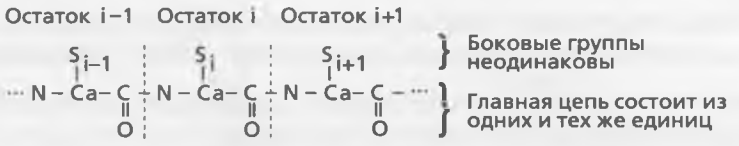


Рис. 1.6. Полипептидная цепь белков состоит из основной цепи (скелета) и боковых групп, последовательность которых может варьироваться. Здесь S_{i-1}, S_i и S_{i+1} — боковые группы. Боковые группы выбираются из набора 20 возможных стандартных аминокислотных остатков. При чередовании различных аминокислотных остатков каждый белок приобретает индивидуальные функциональные характеристики

Белки — достаточно крупные молекулы. В большинстве случаев лишь малая часть их структуры — функциональный центр — несет какую-либо функцию, остальная часть существует только лишь для того, чтобы создавать и фиксировать пространственные связи между остатками активных центров. Белки эволюционируют благодаря изменениям, вызванным мутациями в аминокислотной последовательности. Первый принцип эволюции состоит в том, что изменения в ДНК порождают изменчивость в белковой структуре и функции, что сказывается на репродуктивной способности индивидуума, что будет сказываться при естественном отборе.

На данный момент известно около 15 000 структур белков. Большинство было получено с помощью методов рентгеновской кристаллографии и ядерного магнитного резонанса (ЯМР, NMR). Отсюда пришло понимание обеих функций индивидуальных белков — например, химическое объяснение каталитической активности ферментов — и главных принципов структурного строения белковых молекул и их формы (укладки белковой цепи).

Белковые молекулы представляют собой длинные полимеры, обычно состоящие из нескольких тысяч атомов, образующих равномерно повторяющиеся группы, образующие *остов* молекулы (главная цепь), к которым присоединены специфические ответвления, называемые *боковыми группами* (см. рис. 1.6). Аминокислотная последовательность белка кодирует последовательность боковых групп.

Полипептидная цепь определяет повороты в пространстве; направление цепи, определяющее модель изгиба. Хотя имеется большое разнообразие поворотов в пространственной структуре, существует набор основных структурных особенностей. Они включают повторение основных структурных элементов (например, α-спирали и β-листы) и общие принципы и особенности, такие как плотная упаковка атомов внутри белка. Изгиб в цепи можно трактовать как способ внутримолекулярной конденсации или кристаллизации.

Иерархия в белковой архитектуре

Датский химик Линдерстром-Ланг, исследовавший белки, описал следующие уровни организации структуры белка. Аминокислотная последовательность — набор первичных цепей — называется *первичной структурой*. Распределение спиралей и листов — водородные связи в главной цепи — *вторичная структура*. Собрание и взаимодействие спиралей и листов — *третичная структура*.

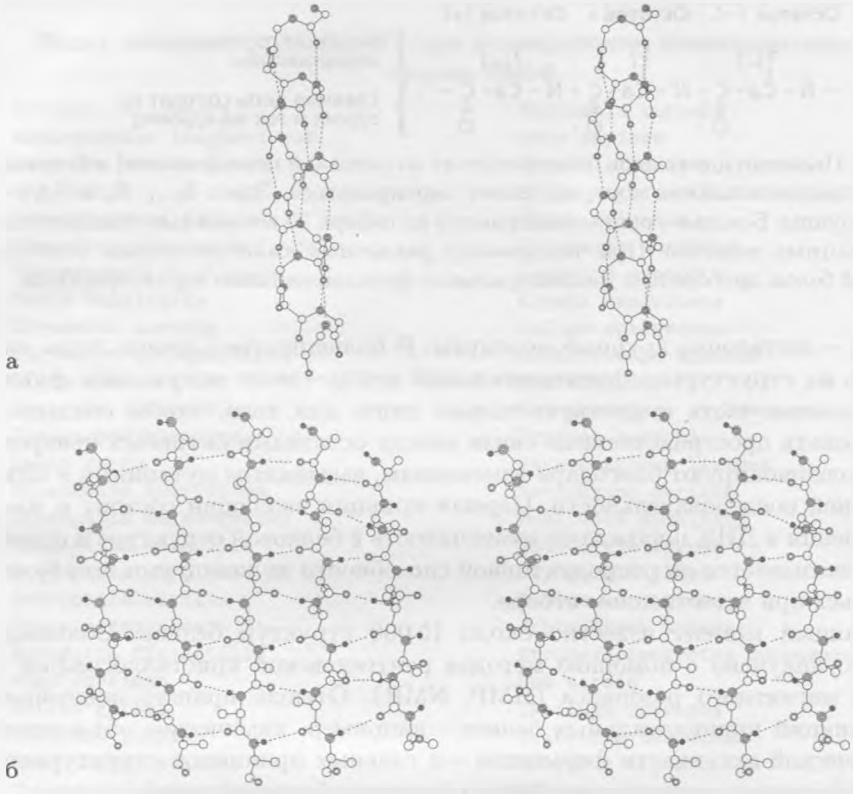


Рис. 1.7. Стандартные вторичные структуры белков. (а) α-спираль (б) β-лист. Пунктирные линии обозначают водородные связи. (б) иллюстрирует параллельный β-лист, в котором все цепочки направлены в одну сторону. Антипараллельные β-листы, в которых все пары смежных цепей направлены в противоположном направлении, также часто встречаются. В действительности β-листы могут формироваться из любой комбинации параллельных и антипараллельных цепей

Для белков, объединенных в более чем одну субъединицу, Берналом было подобрано название для собрания мономеров — *четвертичная структура.* В некоторых случаях, белки могут эволюционно объединяться — четвертичная структура переходит в третичную. Например, пять отдельных ферментов в бактерии *E. coli*, которые катализируют соответствующие шаги в процессе биосинтеза ароматических аминокислот, соответствуют пяти областям одного белка гриба Aspergillus nidulans. Иногда гомологичные мономеры формируют олигомеры различными способами; например, глобины формируют тетрамеры в гемоглобине млекопитающих, в то время как в моллюске *Scapharca inaeguivalvis* используются димеры.

Было доказано, что помимо четырех уровней структурной организации, приведенных выше, удобно использовать следующие дополнительные уровни:

- *Супервторичные структуры.* В белках показана повторяемость взаимодействий между листами и спиралями; супервторичные структуры включают α-спиральные шпильки, β-шпильки и $\beta - \alpha - \beta$-единицу.
- *Домены.* Многие белки включают несколько компактных единиц в одной цепи, которые могут существовать независимо стабильно. Они называются доменами. (Не путать домены, как элемент структуры белков, с доменами, обозначающими основные классы живых организмов: археи, бактерии и эукариоты). РНК-связывающий белок L1, имеет особенности многодоменных белков: связывающее звено появляется в промежутке между двумя доменами, и их геометрия взаимодействия достаточно гибкая, что позволяет осуществлять изменения в конформации лиганда. В иерархии структур, домены располагаются между супервторичными структурами и третичными структурами мономера.
- *Модульные белки.* Модульные белки. Модульные белки являются многодоменными белками, которые часто содержат много копий близко родственных доменов. Эти домены появляются в различных структурных контекстах, так что различные модульные белки представляют из себя мозаику таких доменов. Например, фибронектин, большой внеклеточный белок, участвующий в адгезии и миграции, содержит 29 доменов, включающих в себя множественные тандемные повторы из трех типов доменов,, называемых F1, F2, F3. Их линейная последовательность $(F1)_6(F2)_2(F3)_{15}(F1)_3$. Фибронектиновые домены появляются также в других модульных белках. (См. `http://www.bork.embl-heidelberg.de/Modules/` Там показаны рисунки и номенклатура.)

Классификация белковых структур

Наиболее общая классификация семейств белковых структур основана на вторичной и третичной структуре белка.

Класс	Характеристика
α-Спираль	вторичная структура почти исключительно содержит α-спирали
β-Структура	вторичная структура почти исключительно содержит β-листы
$\alpha + \beta$	α-спирали и β-листы находятся в разных частях молекулы; отсутствие β-α-β-супервторичной структуры
α/β	спирали и листы собраны из β-α-β-структурных единиц
α/β, линейный	линия, проходящая через центры тяжей (strands) листов, — почти прямая
Неструктурированна	мало или нет элементов вторичной структуры

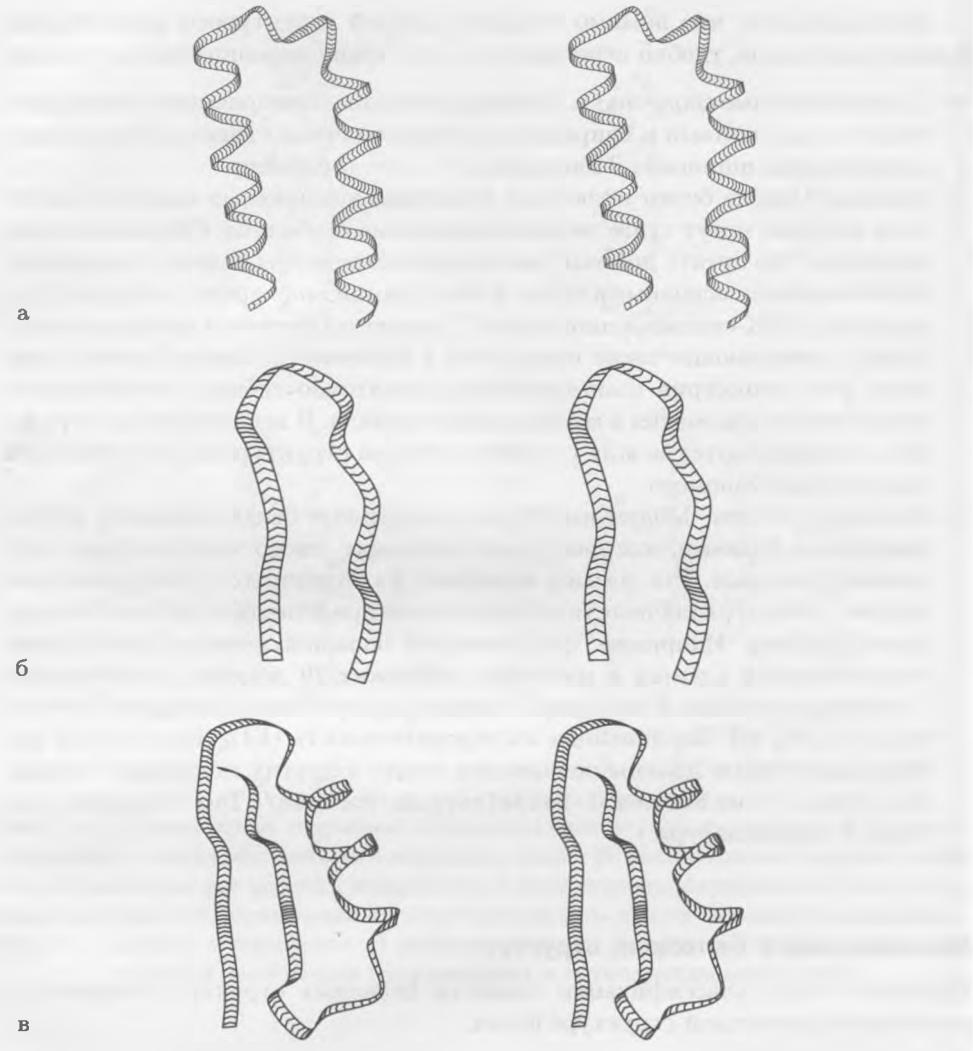


Рис. 1.8. Стандартные вторичные структуры белков. (а) α-спираль (б) β-лист. Пунктирные линии обозначают водородные связи. (в) иллюстрирует параллельный β-лист, в котором все цепочки направлены в одну сторону. Антипараллельные β-листы, в которых все пары смежных цепей направлены в противоположном направлении, также часто встречаются. В действительности, β-листы могут формироваться из любой комбинации параллельных и антипараллельных цепей.

В этих достаточно широких категориях белки имеют большое разнообразие способов укладки. Среди белков со сходной укладкой представлены семейства, имеющие достаточно большое количество деталей структур, последовательностей и функций, обусловленное эволюционными взаимоотношениями. Однако и неродственные белки зачастую имеют похожие способы укладки.

Классификация белковых структур занимает одно из центральных мест в биоинформатике, по крайней мере как мост между последовательностью и функцией. Мы вернемся к этой теме и опишем основные результаты и подходящие Web-ресурсы (Web-сайты). Между прочим, следующий альбом небольших структур дает возможность для того, чтобы попрактиковаться в визуальном анализе и распознавании важных пространственных паттернов (рис. 1.10). Проследите глазами ход цепи, выделите спирали и листы. (Метка указывает направление цепи.) Вы видите вторичную структуру? К каким общим классам вы можете отнести эти структуры? (см. упр. 1.12 и 1.13 и задачу 1.2) Множество других примеров появится в книге «Введение в архитектуру белков: Структурная биология белков» («Introduction to Protein Architecture: The Structural Biology of Proteins (Oxford University Press, 2001)»).

Предсказание структур белков и белковая инженерия

Аминокислотная последовательность белка определяет его пространственную структуру. Если поместить белок в подходящие условия, например, такие, которые есть в клетке, то он восстанавливает свое нативное активное состояние. Некоторые белки нуждаются в помощи специальных белков — шаперонов, для правильного сворачивания. Но это скорее катализирует (ускоряет) процесс, чем направляет его.

Если аминокислотная последовательность содержит достаточно информации для определения ее пространственной структуры, то должна существовать возможность создать алгоритм предсказания пространственной структуры по последовательности. Однако это очень трудно. Поэтому для решения фунда-

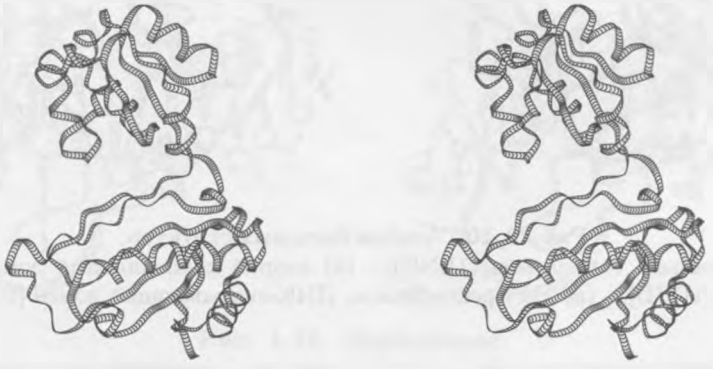


Рис. 1.9. Рибосомный белок L1 из *Methanococcus jannaschii* [1CJS]. ([1CJS] — идентификатор для Protein Data Bank.)

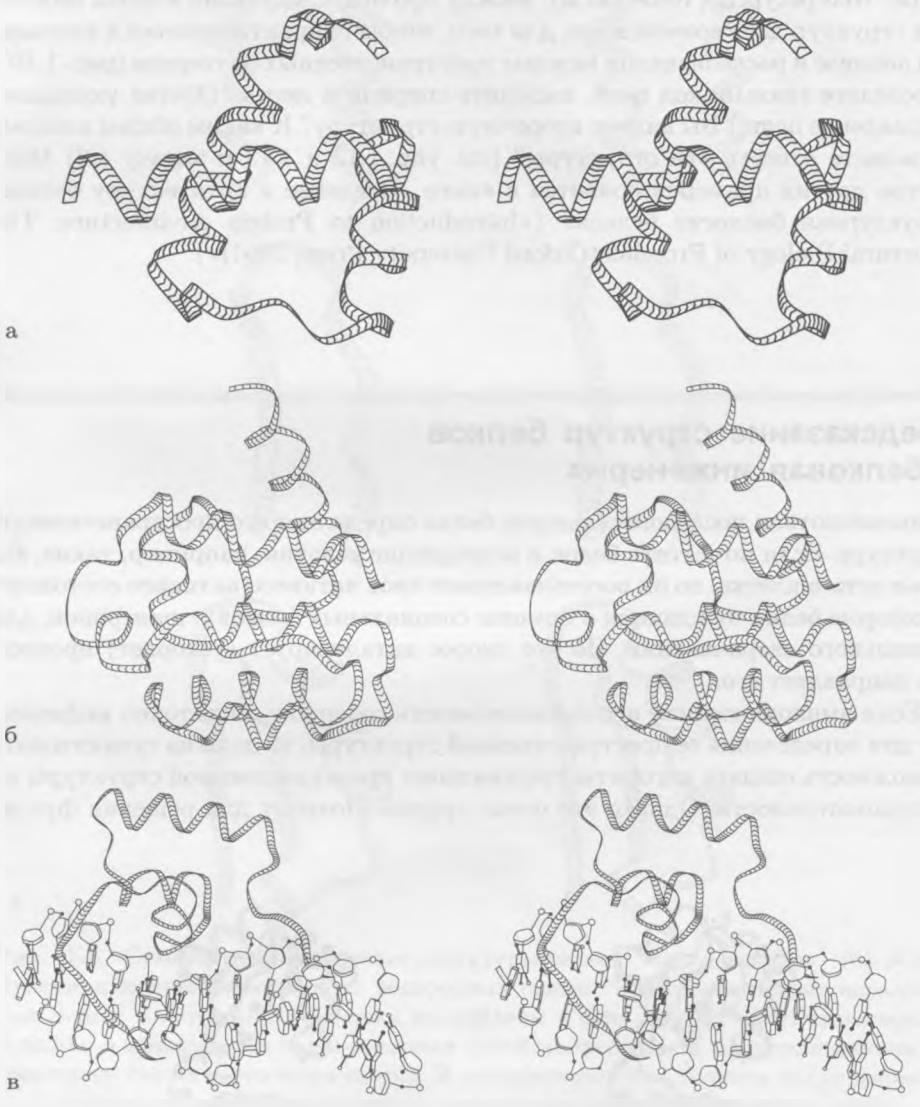


Рис. 1.10. Альбом белковых структур.
(а) «нарезанный» гомеодомен (1ENH); (б) второй калпониновый домен из утропина [1BHD]; (в) HIN рекомбиназа, ДНК-связывающий домен [1HCR];

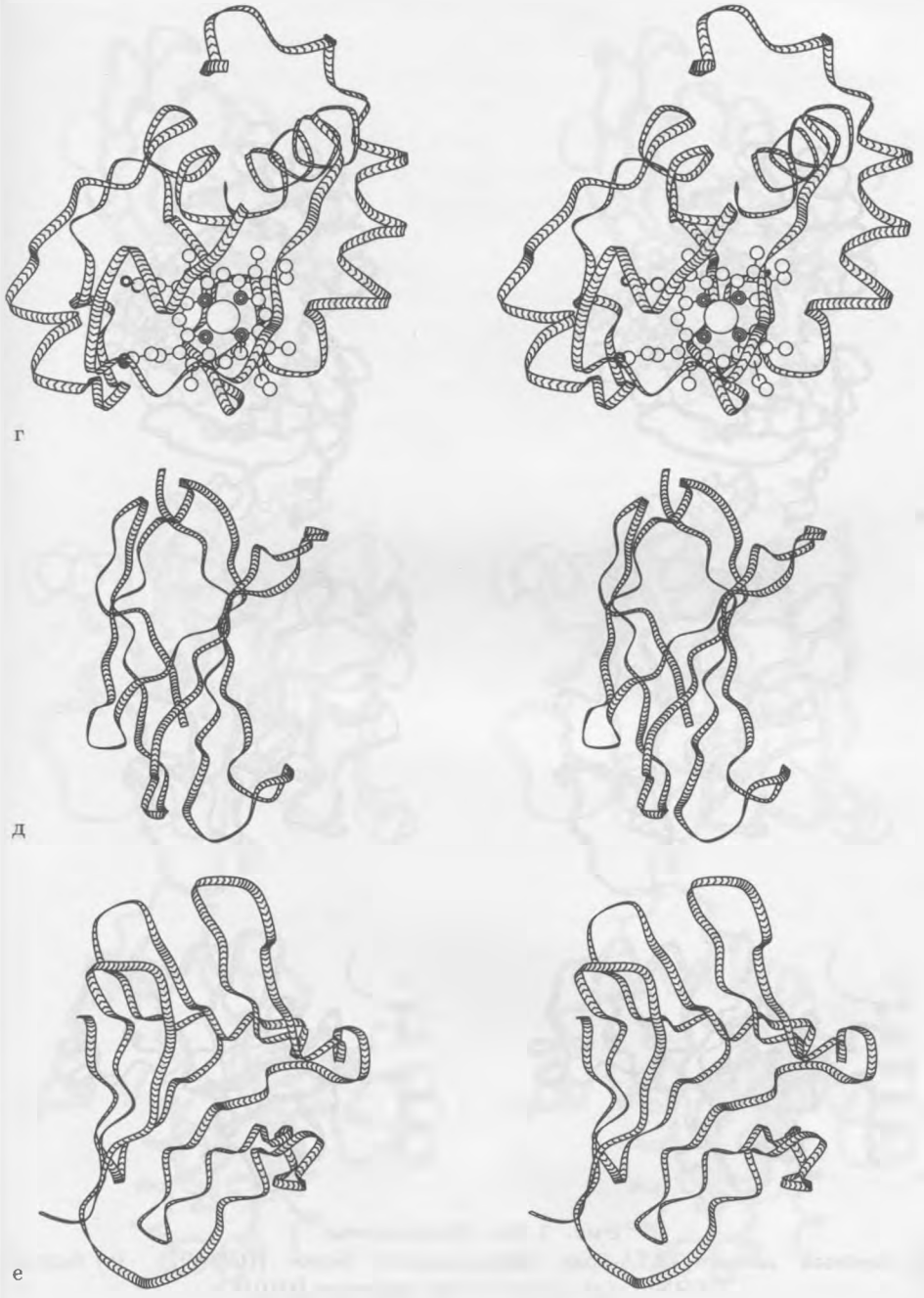


Рис. 1.10. *Продолжение.*
(г) эмбриональный цитохром с из риса [1CCR]; (д) клеточно-адгезивный белок фибронектин типа III-10 [1FNA]; (е) манноза-специфический агглютинин (лектин) [1NPL];

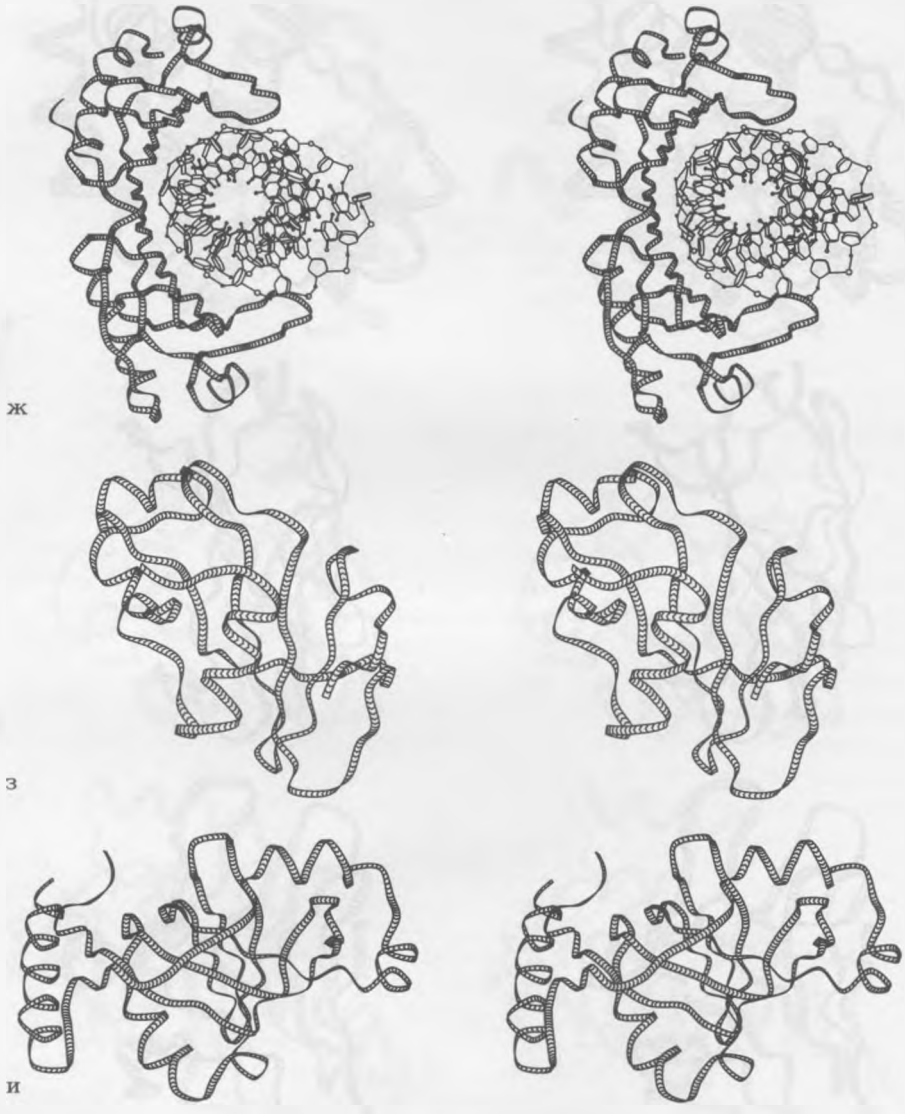


Рис. 1.10. *Продолжение.*
(ж) коровый домен TATA-бокс связывающего белка [1CDW]; (з) барназа [1BRN]; (и) лизил-тРНК синтетаза [1BBW];

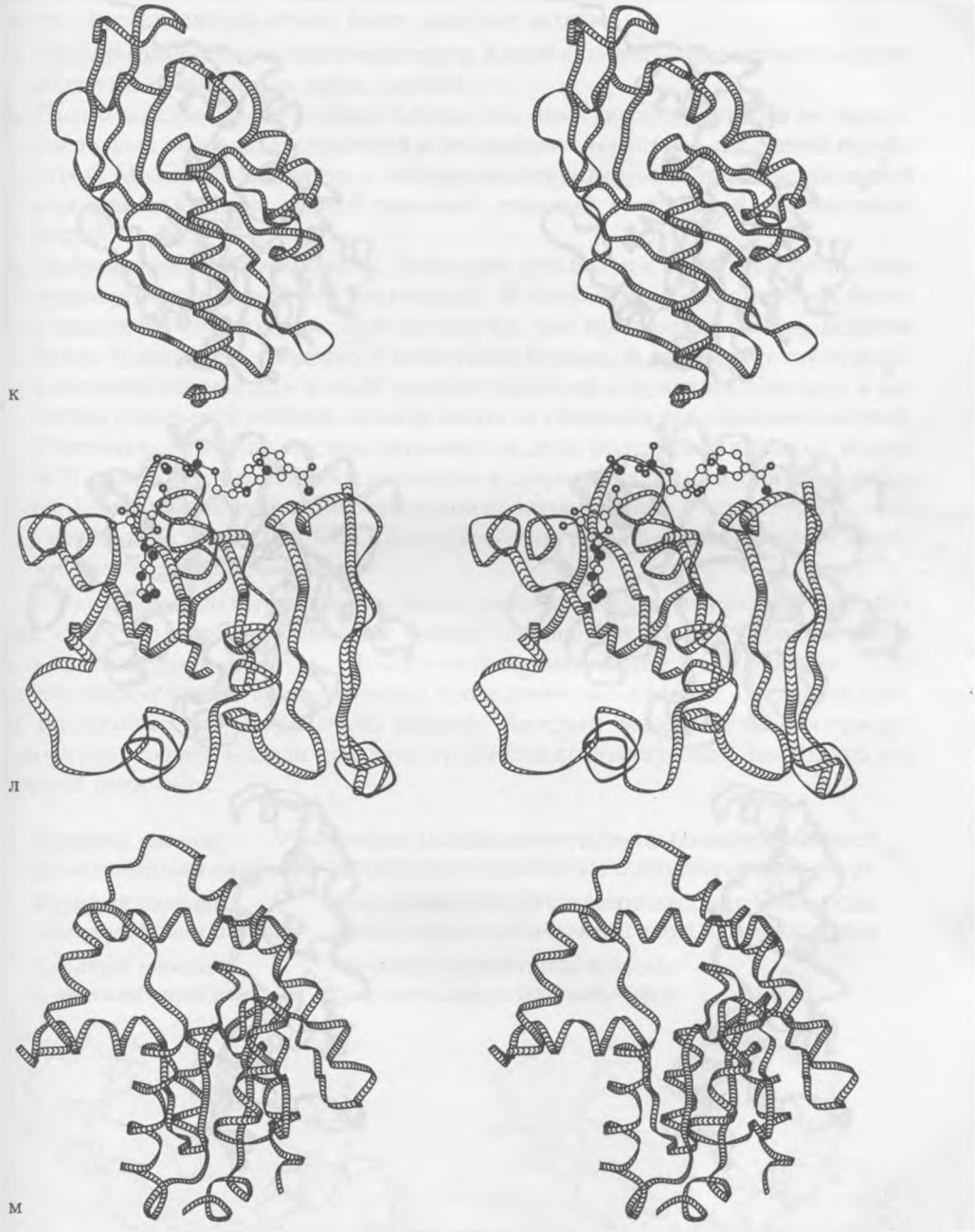


Рис. 1.10. *Продолжение.*
(к) сциталондегидротаза [3STD]; (л) алкогольдегидрогеназа, NAD-связывающий домен [1EE2]; (м) аденилаткиназа [3ADK];

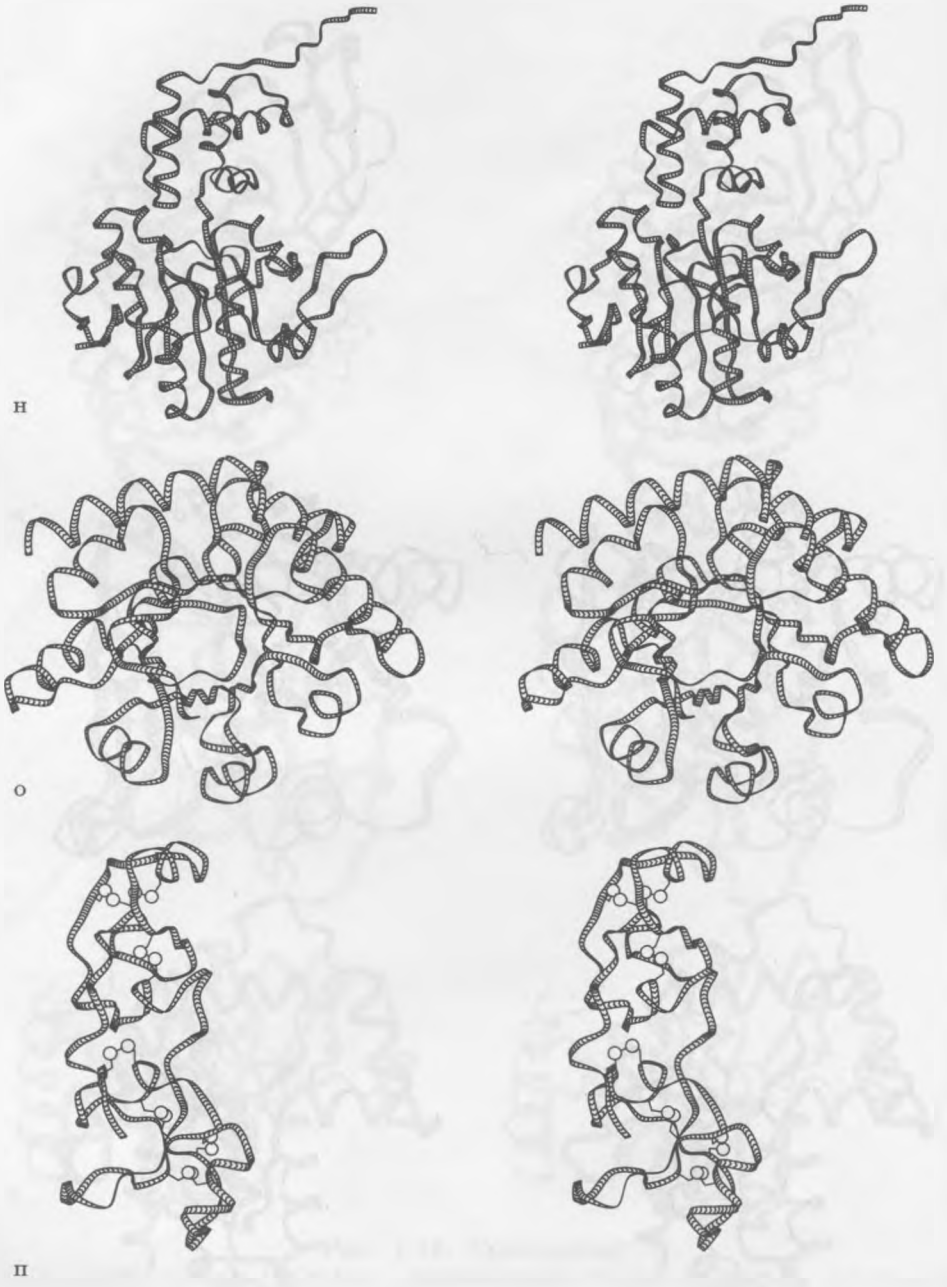


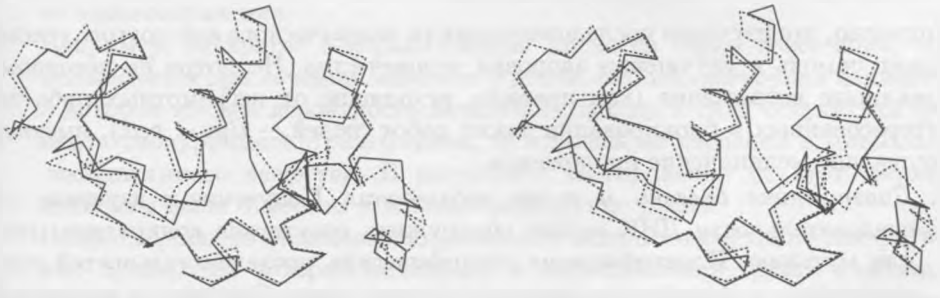
Рис. 1.10. *Окончание.*
(н) метилтрансфераза рецептора хемотаксиса [1AF7]; (о) тиамин-фосфат-синтетаза [2TPS]; (п) порцин панкреатический спазмолитический полипептид [2PSP]

ментальной проблемы — предсказания структуры белков по его последовательности — исследователи ставят более простые задачи:

1. *Предсказание вторичной структуры.* Какие сегменты последовательности образуют спирали или тяжи листов?
2. *Распознавание фолда*[1]. Дана библиотека известных структур и их аминокислотных последовательностей и последовательностей с известной структурой. Можем ли мы найти в библиотеке структуру, которая с наибольшей вероятностью имеет способ укладки, сходный с укладкой неизвестного белка?
3. *Моделирование по гомологии.* Допустим, дан белок с известной последовательностью и неизвестной структурой. И пусть есть гомологи этого белка с известной структурой. В этом случае мы предполагаем, что целевой белок будет иметь сходство с известным белком, и это может послужить в качестве основы для модели соответствующей структуры. Полнота и качество результата зависят, прежде всего, от схожести последовательностей. Считается, что, если последовательности двух родственных белков имеют 50% или более идентичных остатков в выравнивании, то они, вероятно, обладают аналогичной конформацией пространственной структуры с вероятностью не менее, чем 90%. (Как следует из рисунка, приведенного ниже, это консервативная оценка.)

Ниже приведено выравнивание последовательностей и наложение трехмерных структур двух родственных белков: лизоцим из белка куриного яйца и α-лактальбумин павиана. Последовательности достаточно близкие (37% идентичных остатков в выровненных последовательностях), и, следовательно, их трехмерные структуры очень похожи. Каждый белок мог бы послужить в качестве хорошей модели для другого настолько значительно, насколько ход главной цепи схож.

```
Куриный лизоцим          KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGS
α-лактальбумин павиана   KQFTKCELSQNLY--DIDGYGRIALPELICTMFHTSGYDTQAIVEND-ES

Куриный лизоцим          TDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVS
α-лактальбумин павиана   TEYGLFQISNALWCKSSQSPQSRNICDITCDKFLDDDITDDIMCAKKILD

Куриный лизоцим          DGN-GMNAWVAWRNRCKGTDVQA-WIRGCRL-
α-лактальбумин павиана   I--KGIDYWIAHKALC-TEKL-EQWL--CE-K
```



[1] Термин фолд означает способ укладки полипептидной цепи. — *Прим. ред.*

Критическая оценка предсказания структуры (CASP)

Оценка методов для предсказания белковых структуры требует специальных тестов. С этой целью J. Moult предложила двухлетнюю программу CASP (критической оценки структурного прогноза). Кристаллографы и специалисты по ЯМР спектроскопии участвуют в определении пространственной структуры белка, благодаря чему: (1) аминокислотная последовательность может быть опубликована на несколько месяцев раньше ожидаемой даты завершения эксперимента и (2) результаты могут не сообщаться до согласованной даты. Ученые, предсказывающие пространственную структуру, представляют свои модели до последнего срока опубликования экспериментальной структуры. Затем прогнозы и эксперименты сравнивались. (Можно интерпретировать CASP как Олимпийские игры по предсказанию структур. — *Прим. ред.*)

Эффективность предсказания повышается благодаря росту банков данных, а также улучшению методов. Мы обсудим предсказание структуры белков в гл. 5.

Белковая инженерия

Молекулярные биологи были схожи с астрономами — мы могли наблюдать наши объекты, но не модифицировать их. Это теперь не так. В лаборатории мы можем модифицировать нуклеиновые кислоты и белки по желанию. Мы можем изучать их, создавая мутации и наблюдая изменения функций. Мы можем старым белкам придать новые функции, как при разработке каталитических антител. Мы можем даже пытаться создавать новые белки.

Большинство правил о белковой структуре было выведено благодаря наблюдениям за природными белками. Эти правила не обязательно относятся к синтетическим белкам. У природных белков характеристики подчиняются основным принципами физической химии и механизмам белковой эволюции. Синтетические белки должны подчиниться законам физической химии, но не должны ограничиваться правилами эволюции. Белковая инженерия может выделиться в новое научное направление.

Медицинские аспекты

Признано, что изучение последовательности человеческого или другого генома может помочь в улучшении здоровья человечества. Несмотря на некоторые крикливые возражения (как правило, исходящие от неграмотных либо заинтересованных в блокировании таких работ людей. — *Прим. ред.*), имеются следующие медицинские приложения.

1. *Диагностика болезни и риска заболевания.* Получение и изучение последовательности ДНК может обнаружить отсутствие конкретного гена или мутацию. Идентификация специфических последовательностей гена, связанного с болезнями, позволит осуществить быструю и надежную диагностику в следующих случаях: (а) когда пациент ощущает симптомы; (б) заранее предупредить появления симптомов, как в тестах на

наследуемые поздно-приобретаемые заболевания, такие как болезнь Хантингтона (см. врезку); (в) для внутриутробной диагностики потенциальных аномалий таких, как, например, кистозный фиброз, и (г) для генетической консультации пар, собирающихся завести детей.

Во многих случаях наши гены не бесповоротно приговаривают нас к заболеванию, а оставляют возможность, которой мы можем воспользоваться. Примером фактора риска обнаруживаемого на генетическом уровне, является α_1-антитрипсин — белок, который нормально функционирует для ингибирования эластазы в альвеолах легкого. Люди гомозиготные по Z-мутанту α_1-антитрипсина (342Glu→Lys) экспрессируют только нефункциональный белок. Они — в группе риска возникновения эмфиземы, из-за повреждений в легких, вызванных неконтролируемым ингибированием эластазы, а также болезни печени из-за накопления полимерной формы α_1-антитрипсина в гепатоцитах. Курение вызывает развитие эмфиземы почти наверняка. В этих случаях болезнь появляется при сочетании генетических факторов с факторами влияния окружающей среды.

Часто отношение между генотипом и риском заболевания более сложное. Некоторые болезни, такие как астма, зависят от взаимодействия многих генов, а также от факторов влияния окружающей среды. В других случаях, ген может присутствовать и быть исправным, но мутация где-нибудь еще (например, в регуляторной области. — *Прим. ред.*) может изменить уровень экспрессии или распределения по тканям. Такие аномалии могут быть обнаружены измерением белковой активности. Анализ модели белковой экспрессии является также важным путем к поиску лечения.

2. *Генетика реакции на терапию — индивидуально специфическое лечение.* Поскольку люди различаются по особенностям метаболизма лекарства, разные пациенты в одинаковых условиях могут потребовать разные дозировки. Анализ последовательности позволяет индивидуально выбирать лекарства и дозировки, оптимальные для пациентов. Эта быстро развивающаяся область была названа фармогеномика (*pharmacogenomics*). Врачи теперь могут избежать экспериментирования с разными терапиями — процедурами, которые опасны с точки зрения побочных эффектов, часто даже фатальных, и в любом случае дорогих. Лечение пациентов от неблагоприятных реакций на предписанные лекарства требуют миллиардов долларов от здравоохранения.

   Например, лекарство 6-меркаптопурин является очень токсичным, хотя используется при лечении лейкемии у детей. Небольшая группа пациентов, в которой высока вероятность летального исхода в силу отсутствия фермента тиопуринметилтрансферазы, нуждалась во введении в метаболизм лекарственного вещества. В результате тестирования на этот фермент пациенты были отнесены к повышенной группе риска.

   Напротив, теперь появилась возможность использовать такие лекарственные препараты, которые безопасны и эффективны в группе указанных пациентов, но эти вещества были отвергнуты до или во время клинических испытаний из-за медленного действия и тяжелых побочных явлений у некоторых пациентов.

Болезнь Хантингтона

Болезнь Хантингтона является наследственным нейродегенеративным расстройством, которым в США болеют приблизительно 30 000 человек. Симптомы болезни очень серьезные, включают неконтролируемые, наподобие танца, перемещения, умственные расстройства, изменения личности и снижение интеллекта. Смерть обычно наступает в течение 10–15 лет после начала симптомов. Поврежденный ген появился в Новой Англии во время колониального периода в XVII в. Он мог быть ответственным за некоторые обвинения в колдовстве. Ген не утратился у населения, поскольку болезнь проявляется в возрасте 30–50 лет, что значительно позже типичного репродуктивного периода.

Прежде члены семей, затронутых болезнью Хантингтона, в молодом возрасте боялись иметь детей, они не знали, унаследовали ли эту болезнь. Открытие гена[1] болезни Хантингтона в 1993 г. сделало возможным идентифицировать носителей заболевания. Ген содержит многократные повторы тринуклеотида CAG, кодирующие полиглутаминовые блоки в соответствующем белке. (Болезнь Хантингтона — одно из семейных нейродегенеративных расстройств, при которых наблюдаются тринуклеотидные повторы.) Чем больше блок CAG-повторов, тем раньше дебют заболевания и более серьезные симптомы. Нормальный ген содержит 11–28 повторов CAG. Люди с повторами 29–34 почти никогда не заболевают, а у тех, у кого повторов 35–41, могут проявляться только сравнительно мягкие симптомы. Люди, у которых тринуклеотидные повторы встречаются более 41 раза, почти всегда страдают от болезни Хантингтона в полной мере.

Наследственность обладает феноменом, названным *ожиданием*: повторы становятся длиннее в последующих поколениях, прогрессивно увеличивая тяжесть болезни и уменьшая возраст появления симптомов. По некоторым причинам этот эффект преобладает больше в отцовских генах, чем в материнских. Следовательно, люди в предельной группе, обладающие геном с 29–41 повторами, должны думать о риске для своих потомков.

[1] Широко распространенное выражение «ген такого-то заболевания» может привести к непониманию. На самом деле это утверждение означает, что *мутация* в данном гене или его утрата приводит к заболеванию. — *Прим. ред.*

3. *Идентификация мишеней для лекарственных веществ.* Мишень — это белок, функцию которого можно тонко изменить лекарственным веществом, с тем чтобы подавить симптомы или скрытые причины болезни. Точное определение мишени позволит планировать действия при разработке лекарственного препарата. Среди ныне используемых лекарственных препаратов половина действует на рецепторы, около четверти — на ферменты и около четверти — на гормоны. Приблизительно 7% оказывает влияние на неизвестные мишени.

Растущая устойчивость бактерий к антибиотикам привела к кризису в контролировании инфекционных заболеваний. Высока вероятность того, что наши потомки будут смотреть на вторую половину XX в., как на тот небольшой отрезок времени, когда еще можно было контролировать бактериальные инфекции, но этого не удавалось ни до, ни после.

Опираясь на знания, можно модифицировать существующие препараты и снизить остроту необходимости поиска новых лекарств. Анализ генома может пригодиться при поиске мишеней. Дифференциальная геномика и сравнение профилей экспрессии белков у чувствительных и устойчивых к лекарственным препаратам линий патогенных бактерий может указать на те белки, которые отвечают за сопротивляемость. Изучение вариаций генома у опухолевых и нормальных клеток, как ожидается, может помочь в идентификации участков, обладающих разной степенью экспрессии, и тем самым выявить те белки, которые могут быть потенциальными мишенями для противораковых веществ.

4. *Генная терапия.* Если ген пропущен или имеет дефект, то мы хотели бы уметь заменять его нормальным геном или хотя бы увеличивать концентрацию его продукта. Если ген сверхактивен, мы хотели бы уметь выключать его.

   Простое введение белков помогает при многих заболеваниях, из которых, наверное, наиболее известно про введение инсулина больным сахарным диабетом и про фактор VIII общей формы гемофилии.

   Пересадка генов была успешно проделана с помощью животных: человеческие белки продуцировались в молоке коров и овец. У пациентов, страдающих кистозным фиброзом, генная заместительная терапия с использованием аденовируса дала обнадеживающие результаты.

   Способ блокирования генов назван «антисмысловая терапия». Идея заключается во введении ДНК или РНК, которые особым образом связываются с определенным участком гена. Присоединение к эндогенной ДНК может препятствовать транскрипции; присоединение к мРНК может препятствовать трансляции. У антисмысловой терапии есть некоторые успехи в лечении цитомегаловирусного колита и болезни Крона.

   Антисмысловая терапия также весьма привлекательна тем, что может оказывать непосредственное действие на синтез мишени и позволяет быстро обойти стадии разработки лекарственного препарата.

Будущее

Новый век станет свидетелем революционного развития здравоохранения. Рухнет преграда между «голубым небом» исследователя и клинической практикой. Возможно, что уже читатель этой книги сам откроет курс лечения болезни, которая ранее могла бы привести к смерти. В самом деле, похоже, что станет правдой шутка Szent-Györgi: «Рак поддерживает больше людей, чем убивает». Следует надеяться, что это случится, потому что скорее всего исследователи достигнут успехов в развитии антираковой терапии, когда они смогут имитировать бесконтрольный рост клеток.

WEB-РЕСУРСЫ:

Общая информация:

D. Casey из лаборатории Oak Ridge написал два исключительно полезных и компактных введения в молекулярную биологию, обеспечивая важнейшими сведениями для биоинформатики: *Primer on Molecular Genetics (1992). Genomics and Its Impact on Medicine and Society: A 2001 Primer (2001). Washington, D. C*: Human Genome Program, US Department of Energy.

Проект по геному человека:

`http://www.ornl.gov/hgmis/project/info.html`

Статистика генома:

`http://bioinformatics.weizmann.ac.il/mb/statistics.html`

Таксономические сайты:

Species 2000 — полный учет всех известных растений, животных, грибов и микроорганизмов: `http://www.sp2000.org`

Дерево жизни — филогения и биологическая вариативность: `http://phylogeny.arizona.edu/tree`

Базы данных генных заболеваний:

`http://www.ncbi.nlm.nih.gov/omim/`
`http://www.geneclinics.org/profiles/all.html`

Списки банков данных:

`http://www.infobiogen.fr/services/dbcat/`
`http://www.ebi.ac.uk/biocat/`

Список инструментов для анализа:

`http://www.ebi.ac.uk/Tools/index.html`

Форум по сетевому доступу к научной литературе:

`http://www.nature.com/nature/debates/e-access/`

Рекомендуемая литература

Проблеск будущего?

Blumberg, B. S. (1996) 'Medical research for the next millenium', *The Cambridge Review* 117, 3–8. [Очаровательное предсказание вещей, некоторые из которых уже сбылись.]

Переход на электронные публикации

Lesk, M. (1997) *Practical Digital Libraries: Books, Bytes and Bucks* (San Francisco: Morgan Kaufmann). [Повествование о переходе от традиционных библиотек к информационному обеспечению компьютером.]

Berners-Lee, T. and Hendler, J. (2001) 'Publishing on the semantic web', *Nature* 410, 1023–4. [Комментарии от изобретателя web.]

Butler, D. and Campbell, P. (2001) 'Future e-access to the primary literature', *Nature* 410, 613. [Описание развития электронных публикаций научных журналов.]

Определение геномной последовательности

Doolittle, W. F. (2000) 'Uprooting the tree of life', *Scientific American* 282(2), 90–5. [Привлечение анализа последовательностей для понимания взаимоотношений между живыми организмами.]

Green, E. D. (2001) 'Strategies for systematic sequencing of complex organisms', *Nature Reviews (Genetics)* 2, 573–83. [Понятное обсуждение возможных подходов к крупномасштабным проектам секвенирования. Включает список и ссылки на постоянные проекты по секвенированию многоклеточных организмов.]

Sulston, J. and Ferry, G. (2002) *The common thread: a story of science, politics, ethics, and the human genome* (New York: Bantam). [Первоисточник счета.]

Больше о структуре белков

Branden, C.-I. and Tooze, J. (1999). *Introduction to Protein Structure,* 2nd. ed. (New York: Garland). [Прекрасный для понимания текст.]

Lesk, A. M. (2001) *Introduction to Protein Architecture: The Structural Biology of Proteins* (Oxford: Oxford University Press). [Дополнительный том к введению по биоинформатике, с акцентом на белковые структуры и эволюцию.]

Обсуждение баз данных

Frishman, D., Heumann, K., Lesk, A, and Mewes, H.-W. (1998) 'Comprehensive, comprehensible, distributed and intelligent databases: current status', *Bioinformatics* 14, 551–61. [Статус и проблемы организации информации в молекулярной биологии.]

Lesk, A. M. and 25 co-authors. (2001) 'Quality control in databanks for molecular biology', *BioEssays* 22, 1024–34. [Рассмотрение проблем и возможностей развития в гарантировании качества архивных данных, от которых мы все зависим.]

Stein, L. (2001) 'Genome annotation: from sequence to biology', *Nature Reviews (Genetics)* 2, 493–503. [Также подчеркивает важность аннотации.]

Правовая сторона патентования

Human Genome Project Information Website: Genetics and Patenting
`http://www.ornl.gov/hgmis/elsi/patents.html`

Maschio, T. and Kowalski, T. (2001) 'Bioinformatics — a patenting view', *Trends in Biotechnology* 19, 334–9.

Caulfield, T., Gold, E. R., and Cho, M. K. (2000) 'Patenting human genetic material: refocusing the debate', *Nature Reviews (Genetics)* 1, 227–31.

[Обсуждение правовых аспектов геномики и биоинформатики. (1) гены, (2) компьютерные методы — алгоритмы и (3) патентование и авторская регистрация компьютерных программ.]

Упражнения, задачи и компьютерные задания

Упражнение 1.1. (а) Sloan Digital Sky Survey представляет собой карту неба в Северном полушарии за 5 лет. Объем всех необработанных данных в ней превосходит 40 терабайт (1 байт = 1 символ; 1 Тб = 10^{12} байт). Скольким эквивалентам генома человека (HUGE) это соответствует? (б) The Earth Observing

System/Data Information System (EOS/DIS) — серия долгосрочных глобальных наблюдений Земли — для хранения предположительно нуждается в 15 петабайтах (1 петабайт = 10^{15} байт.) Скольким эквивалентам генома человека (HUGE) это будет соответствовать? (в) Сравните объем памяти, необходимый для хранения EOS/DIS, с объемом для хранения полных последовательностей ДНК всех жителей США. (Не рассматривать возможности компрессии данных при их сохранении. Считать, что в последовательности ДНК каждого индивидуума 1 нуклеотидная пара соответствует 1 байту.)

Упражнение 1.2. (а) Сколько дискет понадобилось бы для хранения полного генома человека? (б) Сколько CD-дисков понадобилось бы для хранения полного генома человека? (в) Сколько DVD-дисков понадобилось бы для хранения полного генома человека? (Во всех случаях считать, что каждый знак занимает 1 байт, компрессию не рассматривать.)

Упражнение 1.3. Представьте себе, что вы собираетесь подготовить заметку о болезни Хантингтона (см. с. 70) для интернет-сайта. Какие слова и фразы вы бы сопроводили ссылками?

Упражнение 1.4. На конце гена β-гемоглобина человека имеется следующая последовательность:

```
... ctg gcc cac aag tat cac taa
```

(а) Какова аминокислотная последовательность, соответствующая представленной? (б) Напишите нуклеотидную последовательность, в которой единичная замена нуклеотида приводит к «молчащей» мутации в участке. («Молчащая» мутация оставляет аминокислотную последовательность неизменной.) (в) Напишите нуклеотидную последовательность, в которой единичная замена основания может привести к бессмысленной (миссенс)-мутации в участке, и соответствующую ей аминокислотную. (г) Напишите нуклеотидную последовательность, в которой единичная замена основания приводит к преждевременной остановке синтеза белка, а также соответствующую ей аминокислотную. (д) Напишите нуклеотидную последовательность, в которой единичная замена нуклеотида приводит к ошибке терминации и продолжению синтеза цепи.

Упражнение 1.5. На копии страницы *Полного парного выравнивания белка PAX-6 человека и* Drosophila melanogaster *eyeless*, отметьте с помощью маркера участки, выровненные PSI-BLAST'ом.

Упражнение 1.6. (а) Какое ограничение E-value вы бы использовали в поиске с помощью PSI-BLAST'а, если бы вам нужно было узнать, есть ли уже ваша последовательность в банке? (б) Какое ограничение E-value вы бы использовали при поиске с помощью PSI-BLAST'а, если бы вам необходимо было найти отдаленных гомологов вашей последовательности?

Упражнение 1.7. Оцените минимальную длину, необходимую для составления бессмысленной последовательности, чтобы избежать полного соответствия многим случайным последовательностям в геноме человека.

Упражнение 1.8. Говорят, человеческий род ведет свое начало от всеобщего предка по имени Ева, которая жила примерно 140 000–200 000 лет назад. (а) Сколько поколений было между Евой и нашим современником, если считать, что за столетие сменяется 6 поколений? (б) Если бактериальная клетка делится каждые 20 минут, сколько времени понадобится, чтобы бактерии прошли такое же количество поколений?

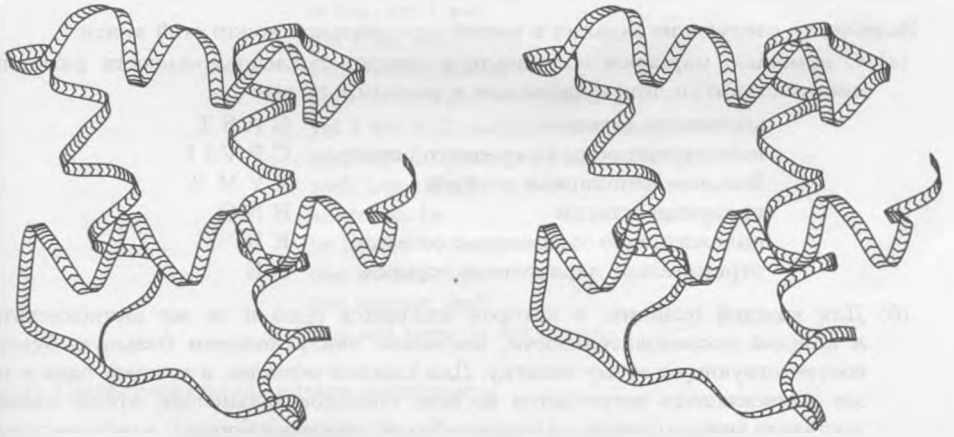
Упражнение 1.9. Назовите аминокислоту, чьи физико-химические свойства сходны с (а) лейцином, (б) аспартатом, (в) треонином. Предполагается, что соответствующие замены в большинстве случаев приведут к достаточно небольшому изменению структуры и функций белка. Назовите аминокислоту, чьи физико-химические свойства сильно отличаются от (а) лейцина, (б) аспартата, (в) треонина. Такие замены могут привести к значительному влиянию на структуру и функции белка, особенно если они находятся внутри белковой молекулы.

Упражнение 1.10. На рис. 1.7, *а* идет ли цепь от N-конца к C-концу вверх или вниз? На рис. 1.7, *б* идет ли цепь от N-конца к C-концу вверх или вниз?

Упражнение 1.11. Из рассмотрения рис. 1.9, сколько раз цепь проходит между доменами Рибосомального протеина L1 *M. jannaschii*?

Упражнение 1.12. На копии рис. 1.10, *л* и 1.10, *м* обозначьте маркером спирали (красным) и тяжи β-листа (синим). На копии рис. 1.10, *ж* и 1.10, *н* разбейте белок на домены.

Упражнение 1.13. Какая из структур на рис. 1.10 содержит следующий домен?



Упражнение 1.14. На копии совмещенных структур лизоцима цыпленка и α-лактальбумина бабуина отметьте с помощью маркера два участка, в которых различается конформация главной цепи.

Упражнение 1.15. В программе на PERL'е со с. 35 оцените долю текста программы, содержащего комментарии. (Сосчитайте полные строчки и их половины.)

Упражнение 1.16. Измените программу на PERL'е, которая вытаскивает имена видов из выдачи PSI-BLAST, так, чтобы она понимала также и имена в виде `[D. melanogaster]`.

Упражнение 1.17. Какова нуклеотидная последовательность молекулы на цветной вклейке 1?

Задача 1.1. Ниже представлено множественное выравнивание участков последовательностей из семейства белков, называемого ETS-доменами. Каждая строчка соответствует аминокислотной последовательности из одного белка, обозначенной как последовательность букв, соответствующих аминокислотам. Просматривая

одну колонку, можно определить аминокислоту, которая появляется на этой позиции в каждом из белков семейства.

```
TYLWEFLLKLLQDR.EYCPRFIKWTNREKGVFKLV..DSKAVSRLWGMHKN.KPD
VQLWQFLLEILTD..CEHTDVIEWVG.TEGEFKLT..DPDRVARLWGEKKN.KPA
IQLWQFLLELLTD..KDARDCISWVG.DEGEFKLN..QPELVAQKWGQRKN.KPT
IQLWQFLLELLSD..SSNSSCITWEG.TNGEFKMT..DPDEVARRWGERKS.KPN
IQLWQFLLELLTD..KSCQSFISWTG.DGWEFKLS..DPDEVARRWGKRKN.KPK
IQLWQFLLELLQD..GARSSCIRWTG.NSREFQLC..DPKEVARLWGERKR.KPG
IQLWHFILELLQK..EEFRHVIAWQQGEYGEFVIK..DPDEVARLWGRRKC.KPQ
VTLWQFLLQLLRE..QGNGHIISWTSRDGGEFKLV..DAEEVARLWGLRKN.KTN
ITLWQFLLHLLLD..QKHEHLICWTS.NDGEFKLL..KAEEVAKLWGLRKN.KTN
LQLWQFLVALLDD..PTNAHFIAWTG.RGMEFKLI..EPEEVARLWGIQKN.RPA
IHLWQFLKELLASP.QVNGTAIRWIDRSKGIFKIE..DSVRVAKLWGRRKN.RPA
RLLWDFLQQLLNDRNQKYSDLIAWKCRDTGVFKIV..DPAGLAKLWGIQKN.HLS
RLLWDYVYQLLSD..SRYENFIRWEDKESKIFRIV..DPNGLARLWGNHKN.RTN
IRLYQFLLDLLRS..GDMKDSIWWVDKDKGTFQFSSKHKEALAHRWGIQKGNRKK
LRLYQFLLGLLTR..GDMRECVWWVEPGAGVFQFSSKHKELLARRWGQQKGNRKR
```

Выполните следующие задания в вашей персональной копии этой книги:

(а) С помощью маркеров обозначьте в каждой последовательности разными цветами остатки, принадлежащие к разным классам:

маленькие остатки	G A S T
неполярные остатки среднего размера	C P V I L
большие неполярные остатки	F Y M W
полярные остатки	H N Q
положительно заряженные остатки	K R
отрицательно заряженные остатки	D E

(б) Для каждой позиции, в которой находится одна и та же аминокислота в каждой последовательности, поставьте внизу колонки большую букву, соответствующую этому остатку. Для каждой позиции, в которой одна и та же аминокислота встречается во всех последовательностях, кроме одной, поставьте внизу колонки маленькую букву, соответствующую наиболее предпочтительному остатку.

(в) Что представляют собой паттерны периодичности консервативных аминокислот?

(г) Какое распределение консервативности заряженных остатков вы наблюдаете? Предложите разумную гипотезу, с какой молекулой взаимодействуют эти домены.

Задача 1.2. Объедините структуры, представленные на рис. 1.10 в следующие категории: α-спиральные, β-листовые, $\alpha+\beta$, α/β линейные, α/β-бочонки, с небольшой вторичной структурой.

Задача 1.3. Измените программу на PERL'е со с. 32 так, чтобы она печатала результат трансляции последовательности ДНК во всех шести возможных рамках считывания (прочтенная последовательность в трех рамках и комплементарная ей также в трех рамках).

Задача 1.4. Задача 1.4. Для каких из следующих наборов фрагментов строк PERL-программа со с. 35 работает правильно? (а) Сможет ли она правильно распознать:

Kate, when France is mine and I am
yours, then yours is France and you are mine.

из:

```
Kate, when France
France is mine
is mine and
and I am\nyours
yours then
then yours is France
France and you are mine\n
```

(б) Сможет ли она правильно распознать:

One woman is fair, yet I am well; another is wise, yet I am well; another virtuous, yet I am well; but till all graces be in one woman, one woman shall not come in my grace.

из:

```
One woman is
woman is fair,
is fair, yet I am
yet I am well;
I am well; another
another is wise, yet I am well;
yet I am well; another virtuous,
another virtuous, yet I am well;
well; but till all
all graces be
be in one woman,
one woman, one
one woman shall
shal not come in my grace.
```

(в) Сможет ли она правильно распознать:

```
That he is mad, 'tis true: 'tis true 'tis pity;
And pity 'tis 'tis true.
```

из:

```
That he is
is mad, 'tis
'tis true
true: 'tis true 'tis
true 'tis
'tis pity;\n
 pity;\nAnd pity
 pity 'tis
 'tis 'tis
 'tis true.\n
```

В (в): будет ли она работать, если убрать все знаки препинания из строк?

Задача 1.5. Измените программу, написанную на PERL'е со с. 35 так, чтобы она корректно распознавала все фрагменты из предыдущей задачи. (Внимание: это нелегко!)

Задача 1.6. Напишите программу на PERL'е для поиска совпадения мотивов, как это показано во врезке на с. 41. (а) Необходимы точные совпадения. (б) Разрешено одно несовпадение, не обязательно в первой позиции, как в примере, но не вставки (инсерции) и не удаления (делеции).

Задача 1.7. Программа PERL очень лаконична. Вот альтернативная версия программы для сбора перекрывающихся фрагментов (см. с. 35):

```
#!/usr/bin/perl

$/ = "";
@fragments = split("\n",<DATA>);

foreach (@fragments) { $firstfragment{$_} = $_; }

foreach $i (@fragments) {
    foreach $j (@fragments) { unless ($i eq $j) {
        ($combine = $i . "XXX" . $j) =~ /([\S ]{2,})XXX\1/;
        (length($1) <= length($successor{$i})) || { $successor{$i} = $j };
    }                          }
    undef $firstfragment{$successor{$i}};
}

$test = $outstring = join "", values(%firstfragment);
while ($test = $successor{$test}) { ($outstring .= "XXX" . $test) =~ s/([\S ]+)XXX\1/\1/; }

$outstring =~ s/\\n/\n/g; print "$outstring\n";

__END__
the men and women merely players;\n
one man in his time
All the world's
their entrances,\nand one man
stage,\nAnd all the men and women
They have their exits and their entrances,\n
world's a stage,\nAnd all
their entrances,\nand one man
in his time plays many parts.
merely players;\nThey have
```

(Это хороший пример того, чего не следует делать. Любой, написавший код в таком стиле, должен быть немедленно уничтожен. Отсутствие комментариев, сложное кодирование и излишняя краткость, делают невозможным понимание назначения программы. Программа, написанная в таком стиле, сложна в отладке и не может быть поддержана. Если когда-нибудь вы смените кого-нибудь на некоторой работе, и вам достанется работать с подобной программой, то я буду вам сочувствовать.)

(а) Скопируйте краткую программу, предложенную в данной задаче, и первоначальную со с. 35, так. Чтобы они располагались на рядом странице. Где возможно, соотнесите каждую строку из краткой программы с соответствующей группой строк из длинной программы.

(б) Добавьте в краткую программу комментарии, необходимые для объяснения того, что она делает (для этого достаточно адаптировать комментарии из первоначальной программы), и как она это делает.

Интернет-задание 1.1. Определите источник всех цитат из трагедий Шекспира в выравнивании во врезке с. 41.

Интернет-задание 1.2. Определите Web-сайты, дающие простые объяснения и/или он-лайн демонстрации (а) Полимеразной Цепной Реакции (ПЦР), (б) блота по Саузерну, (в) рестрикционной карты, (г) кэш-памяти, (д) Суффиксного дерева. Напишите абзац, объясняющий эти термины на основе данных с сайтов.

Интернет-задание 1.3. К какому типу принадлежат следующие виды? (а) Морская звезда. (б) Минога. (в) Солитер. (г) Гингко. (д) Скорпион. (е) Медуза. (ж) Морской анемон.

Интернет-задание 1.4. Каковы тривиальные названия данных видов? (а) *Acer rubrum.* (б) *Orycteropus afer.* (в) *Beta vulgaris.* (г) *Pyractomena borealis.* (д) *Macrocystic poryfera.*

Интернет-задание 1.5. Обычный завтрак англичанина состоит из: куриных яиц, жареных в сале, бекона, копченой сельди, жареных грибов, жареного картофеля, жареных томатов, печеных бобов, тостов и чая с молоком. Напишите полную таксономическую классификацию организмов, из которых получены указанные продукты.

Интернет-задание 1.6. Получите и выровняйте последовательность митохондриального цитохрома-*b* лошади, кита и кенгуру. (а) Сравните степень попарного сходства каждой пары последовательностей с результатами сравнения последовательностей панкреатических рибонуклеаз этих видов в задании 1.2. Совместимы ли выводы, полученные в результате анализа последовательностей митохондриального цитохрома-*b*, с выводами, полученными в результате анализа последовательностей панкреатических рибонуклеаз этих видов в задании 1.2? (б) Сравните *взаимное* сходство этих последовательностей с результатами сравнения последовательностей панкреатических рибонуклеаз этих видов в задании 1.2. Совместимы ли выводы, полученные в результате анализа последовательностей митохондриального цитохрома-*b*, с выводами, полученными в результате анализа последовательностей панкреатических рибонуклеаз этих видов в задании 1.2?

Интернет-задание 1.7. Получите последовательности панкреатических рибонуклеаз из спермы кита, лошади, гиппопотама. Совместимы ли результаты с взаимосвязью, показанной с помощью SINE-повторов?

Интернет-задание 1.8. Было показано, что аминокислотная последовательность цитохрома-*b* слона и мамонта очень похожи. Одна из гипотез, это объясняющих, состоит в том, что для функционирования цитохрома-*b* требуется такое большое количества консервативных остатков, что все цитохромы-*b* из всех животных настолько же сходны, насколько сходны протеины слона и мамонта. Проверьте эту гипотезу, получив последовательности цитохрома-*b* других млекопитающих и проверив, совпадают ли аминокислотные последовательности более удаленных видов настолько же, насколько они совпадают у слона и мамонта.

Интернет-задание 1.9. Получите последовательности цитохрома-*c* человека, гремучей змеи и варана. Какая пара обладает наибольшей степенью сходства? Кажется ли это вам удивительным? Почему?

Интернет-задание 1.10. Пошлите последовательности панкреатических рибонуклеаз лошади, малого кита-полосатика и красного кенгуру (Пример 1.2) на сервер множественного выравнивания T-коффе: `http://www.ch.embnet.org/software/TCoffee.html` Совпадает ли получившееся выравнивание с выравниванием, полученным в примере 1.2 с помощью CLUSTAL-W.

Интернет-задание 1.11. Линней разделил Царство животных на шесть классов: млекопитающие, птицы, амфибии (включая рептилий), рыбы, насекомые и черви. При этом он предполагал, например, что крокодилы и саламандры — более близкородственные животные, чем крокодилы и птицы. В XIX в. Томас Хаксли объединил птиц и рептилий. Определите степень сходства для трех подходящих белков из саламандры, крокодила и птицы между гомологичными последовательностями. Какая пара животных более близка? Кто был прав, Линней или Хаксли?

Интернет-задание 1.12. Когда был открыт последний вид приматов?

Интернет-задание 1.13. В каких еще видах после составления таблицы на старницах 38–39 были обнаружены гомологи PAX-6?

Интернет-задание 1.14. Определите еще три модульных белка (кроме собственно фибронектина), содержащих домены фибронектина III.

Интернет-задание 1.15. Приведите шесть примеров заболеваний (кроме диабета и гемофилии), которые излечиваются путем направленного введения недостающего белка. Укажите вводимый белок в каждом случае.

Интернет-задание 1.16. Для каких поздно проявляющихся заболеваний варианты гена аполипопротеина E создают особо большой (наибольший) риск? Что известно о механизме, по которому этот вариант влияет на развитие заболевания?

Интернет-задание 1.17. Примерно для 10% европейцев транквилизатор кодеин бесполезен, так как в их организме отсутствует фермент, переводящий кодеин в активную форму — морфин. Какая наиболее обычная мутация, вызывающая эти условия?

2 Организация генома и эволюция

Геномика и протеомика

Геном типичной бактерии представлен единственной молекулой ДНК, которая в растянутом виде может иметь длину порядка 2 мм (сама клетка при этом имеет диаметр порядка 0,001 мм). ДНК высших организмов организована в хромосомы — нормальные клетки человека содержат 23 пары хромосом. Общее количество генетической информации на клетку — последовательность нуклеотидов ДНК — практически всегда постоянна для всех особей одного

вида, но сильно отличается у разных видов (см. врезку для ознакомления с более полным списком):

Организм	Размер генома
Вирус Эпштейна—Барра	0.172×10^6 нуклеотидных пар
Бактерия(*E. coli*)	4.6×10^6
Дрожжи (*S. cerevisiae*)	12.1×10^6
Нематода (*C. elegans*)	95.5×10^6
Резуховидка Таля (*A. thaliana*)	117.0×10^6
Плодовая мушка (*D. melanogaster*)	180.0×10^6
Человек (*H. sapiens*)	3200×10^6

Не вся ДНК кодирует белки. С другой стороны, некоторые гены представлены многочисленными копиями. Таким образом, количество информации о белковых последовательностях в клетке не может быть оценено, исходя лишь из размера генома.

Гены

Один ген, кодирующий отдельный белок, соответствует последовательности нуклеотидов одного или большего числа участков молекулы ДНК. Последовательность ДНК коллинеарна («параллельна») последовательности белка. У видов, у которых генетический материал представлен двуцепочечной ДНК, гены могут находиться на любой из цепей. Бактериальные гены представляют собой непрерывные участки ДНК. Таким образом, функциональная единица генетической информации у бактерий представляет собой последовательность $3N$ нуклеотидов, которая кодирует последовательность N аминокислот, или же последовательность N нуклеотидов, кодирующую молекулу структурной (например, рибосомной) РНК из N остатков. Такая последовательность, снабженная аннотациями, может быть сохранена в виде типичной записи одного из архивов генетических последовательностей.

У эукариот последовательности нуклеотидов, кодирующие аминокислотные последовательности отдельных белков, организованы более сложным образом. Здесь совершенно иная зависимость между размерами гена и закодированного в нем белка, чем у бактерий. Часто один ген представлен в виде отдельных сегментов геномной ДНК. *Экзон* — это участок ДНК, сохраняемый в зрелой информационной РНК, которую рибосом транслирует в белок. *Интрон* — это промежуточный участок между двумя экзонами. Клеточные механизмы сплайсируют определенные сегменты в транскриптах РНК, основываясь на сигнальных последовательностях, которые фланкируют экзоны. Многие интроны являются очень длинными — преимущественно длиннее, чем экзоны.

Регуляторные механизмы организуют экспрессию генов. Гены могут быть включены или выключены (или отрегулированы более тонко) в ответ на различные концентрации питательных веществ, на стресс, или на сложные программы развития в течение жизни организма. Множество управляющих

Размеры геномов

Организм	Число пар оснований	Число генов	Комментарий
ϕX–174	5 386	10	вирус, инфицирующий *E. coli*
Человеческая митохондрия	16 569	37	субклеточная органелла
Вирус Эпштейна—Барра (EBV)	172 282	80	вызывает мононуклеоз
Mycoplasma pneumoniae	816 394	680	возбудитель эпидемии циклической пневмонии
Rickettsia prowazekii	1 111 523	878	бактерия, возбудитель эпидемического тифа
Treponema pallidum	1 138 011	1 039	бактерия, вызывает сифилис
Borrelia burgdorferi	1 471 725	1 738	бактерия, вызывает болезнь Лайма
Aquifex aeolicus	1 551 335	1 749	бактерия из горячих источников
Thermoplasma acidophilum	1 564 905	1 509	архея, не имеет клеточной стенки
Campylobacter jejuni	1 641 481	1 708	частая причина пищевых отравлений
Helicobacter pylori	1 667 867	1 589	основная причина язвы желудка
Methanococcus jannaschii	1 664 970	1 783	архея, термофил
Hemophilus influenzae	1 830 138	1 738	бактерия, причина инфекций среднего уха
Thermotoga maritima	1 860 725	1 879	морская бактерия
Archaeoglobus fulgidus	2 178 400	2 437	другая архея
Deinococcus radiodurans	3 284 156	3 187	радиационно-устойчивая бактерия
Synechocystis	3 573 470	4 003	цианобактерия, сине-зеленая водоросль
Vibrio cholerae	4 033 460	3 890	возбудитель холеры
Mycobacterium tuberculosis	4 411 529	4 275	возбудитель туберкулеза
Bacillus subtilis	4 214 814	4 779	популярна в молекулярно-биологических исследованиях
Escherichia coli	4 639 221	4 406	вечный фаворит молекулярных биологов
Pseudomonas aeruginosa	6 264 403	5 570	самый большой прокариотический организм, чей геном секвенирован
Saccharomyces cerevisiae	12.1×10^6	5 885	дрожжи, первый секвенированный эукариотический геном
Caenorhabditis elegans	95.5×10^6	19 099	Червь
Arabidopsis thaliana	1.17×10^8	25 498	цветковое растение (покрытосемянное)
Drosophila melanogaster	1.8×10^8	13 601	плодовая мушка
Fugu rubripes	3.9×10^8	30 000	рыба-собака (fugu fish)
Человек	3.2×10^9	34 000?	
Пшеница	16×10^9	30 000	
Саламандра	10^{11}	?	
Psilotum nudum	10^{11}	?	травянистый папоротник — простое растение

участков ДНК лежит рядом с участками, кодирующими белки. Они содержат последовательности, которые служат в качестве сайтов связывания молекул, транскрибирующих ДНК, или последовательности, связывающие регуляторные молекулы, которые могут *блокировать* транскрипцию. Простой пример: в бактериальных геномах имеются соседние гены, которые кодируют несколько белков, катализирующих последовательные стадии одного биохимического пути. Они совместно включаются и выключаются, поскольку находятся под контролем общей регуляторной последовательности. Ф. Джэкоб, Дж. Монод и Е. Воллман назвали эти группы генов *оперонами*. Любой может легко понять выгодность параллельного механизма контроля их экспрессии.

У животных метилирование ДНК обеспечивает тканеспецифичную дифференциальную экспрессию генов в процессе развития. Продукты определенных генов запускают самоубийство клеток — процесс, названный *апоптозом*. Нарушения в апоптотическом механизме, ведущие с бесконтрольному росту, обнаружены в некоторых видах раковых опухолей, и стимуляция этих механизмов является основным подходом при лечении рака.

Вывод состоит в том, что редуцирование генетической информации до отдельных кодирующих последовательностей означает скрытие очень сложной природы взаимодействий между ними и игнорирование исторических и интегративных аспектов генома. Роббинс идеально описал данную ситуацию:

> ... Рассмотрим 3,2 Гигабайта человеческого генома в качестве эквивалента 3,2 Гигабайта файлов на накопительном устройстве (жестком диске) какого-то неизвестного компьютера. Получение последовательности эквивалентно получению образа содержимого жесткого диска. Понимание последовательности эквивалентно обратной инженерии такого неизвестного компьютера (всего «железа» и 3,2 Гб информации), т. е. пониманию всех технологий и идей, вложенных в создание данного компьютера.
>
> Обратная инженерия последовательности сложна потому, что окончательный образ накопительного устройства не будет копией каждого файла, а будет потоковым дампом байтов, расположенных в порядке, в котором они были введены в устройство. Более того, известно, что файлы могут быть фрагментированы. А еще некоторые устройства содержат стертые файлы или другой мусор. Как только мусор распознан и удален из рассмотрения, а фрагментированные файлы вновь собраны, обратная инженерия кодов может быть осуществлена с помощью лишь частичного (а иногда и ошибочного) понимания процессора, на котором эти коды работают. Фактически, выведение структуры и функции процессора является частью проекта, в то время как некоторое количество из 3,2 Гигабайта данных представляют собой двоичную спецификацию многочисленных параметров работы процессора. В дополнение к этому каждый может допускать, что огромная база данных также содержит код, полученный как результат миллионов усовершенствований и проверок, осуществленных, порой, с использованием клуджей (программ, которые теоретически не должны работать, а почему-то работают), «макаронных» программ, а также при участии предприимчивых спонсоров, которые восхищаются новыми идеями,

такими, как написание самомодифицирующегося кода и использование недокументированных системных тонкостей.

Р. Д. Роббинс, 1992, «Сложности в проекте „Геном человека“», Инженерия IEEE в медицине и биологии, 11, 25–34 (© IEEE, 1999).

Белки

База данных последовательностей аминокислот в белке напрямую связана с базой данных нуклеотидных последовательностей ДНК благодаря генетическому коду. Действительно, информация о последовательности нового белка в настоящее время определяется, скорее путем расшифровки последовательности ДНК, чем прямым секвенированием белков. Исторически так сложилось, что проблема определения химическими методами аминокислотной последовательности белков была решена до того, как был определен генетический код, и были развиты методы определения нуклеотидных последовательностей ДНК. Секвенирование инсулина, осуществленное Ф. Сенгером в 1955 г., было первым доказательством того факта, что белки имеют определенную белковую последовательность. До того момента это утверждение было гипотетическим.

Есть ли разница между аминокислотными последовательностями, определенными напрямую из белков или же путем расшифровки ДНК? Во-первых, предположим, что, возможно правильно определить белок-кодирующие участки среди потока информации ДНК. Программы распознавания, которые занимаются этим вопросом, подвержены трем типам ошибок: потеря истинной последовательности белков, неправильный сплайсинг гена, или же белок может быть описан не полностью. Существует несколько вариантов, усложняющих задачу: гены, кодирующие разные белки, могут перекрываться или же собираться из экзонов разными способами в разных тканях. И наоборот, некоторые генетические последовательности, которые, казалось бы, кодируют белки, могут быть дефектными или неэкспрессируемыми. *Белок, полученный по геномной последовательности, является гипотетическим объектом до тех пор, пока его существование не будет подтверждено экспериментом.*

Во-вторых, очень часто продуктом экспрессии гена является молекула, которая в дальнейшем должна быть преобразована в клетке. Процесс преобразования ведет к созданию зрелого белка, который значительно отличается от того, который был предсказан расшифровкой генной последовательности. Во многих случаях участки, утерянные вследствие посттрансляционной модификации (которая является молекулярным аналогом пирсинга), очень важны, и в отличие от пирсинга, эти участки несут функциональное значение. К посттрансляционным модификациям относят присоединение лигандов (например, ковалентносвязанной группы гема в цитохроме с), гликозилирование, метилирование, урезание белка, и многое другое. Образование дисульфидных мостиков (химические связи между цистеиновыми остатками) не могло быть распознано по аминокислотной последовательности. В некоторых случаях мРНК редактируется перед трансляцией так, что создаются аминокислотные последовательности, которые нельзя получить из гена.

Протеомы

Геном организма дает полный, но статический набор характеристик потенциальной жизни конкретного представителя. Уровень развития организма и его активность на молекулярном уровне в любой момент времени зависит в первую очередь от количества и распределения белков. Протеомный проект — это крупномасштабная программа, использующая интегральный подход для работы с экспрессией белков в биологических системах, дополняя и расширяя геномные проекты.

Какие типы данных мы бы хотели оценить, и какие продуманные экспериментальные методы существуют для этого? Основная цель — это пространственно-временное описание синтеза белков в организме. Уровни синтеза различных белков варьируются в разных тканях и типах клеток, а также для разных уровней активности. Доступны методы эффективного анализа транскрипционной экспрессии для множественных генов (см. табл. на с. 86 и цветную иллюстрацию V). Тем не менее, поскольку белки разрушаются с разной скоростью, необходимо изучать белки напрямую. Распределение белков в клетке — это кинетический баланс между скоростью белкового синтеза и белковой деградации. Двумерный электрофорез в полиакриламидном геле с высокой разрешающей способностью показывает экспрессию белкового содержимого в пробе. Масс-спектроскопическая техника идентифицирует белки, которые воздействовали на образец, и их посттрансляционные модификации.

Также возможно создавать белковые матрицы, для обнаружения белковых взаимодействий (см. рис. 2.1).

Применение этих методов дает общую картину белковой активности организма, в то время как геном дает набор потенциальных белков. Р. Симпсон провел аналогию: если геном — это список инструментов в оркестре, то протеом — это оркестр, играющий симфонию.

ДНК-микрочипы

ДНК-микрочипы — это способ проверки пробы ДНК одновременно на присутствие многих последовательностей. ДНК-чипы могут быть использованы при решении следующих задач. (1) Определения экспрессии разных белков путем обнаружения мРНК. (2) Определения генотипа путем обнаружения разных вариантов последовательностей гена, учитывая однонуклеотидные полиморфизмы (ОНП, SNP), но не сводя к ним. Можно определить просто их отсутствие или присутствие, или вычислить относительный избыток. Заметим, однако, что корреляция между избытком мРНК и соответствующего белка — неполная.

Чтобы определить экспрессию всех генов клетки необходимо измерить относительные количества многих мРНК. Гибридизация — это аккуратный и чувствительный способ для обнаружения, присутствует ли конкретная последовательность в пробе ДНК. Ключ к обширному анализу заключается в проведении многих гибридизационных экспериментов параллельно. Это то, что достигается с помощью микрочипов.

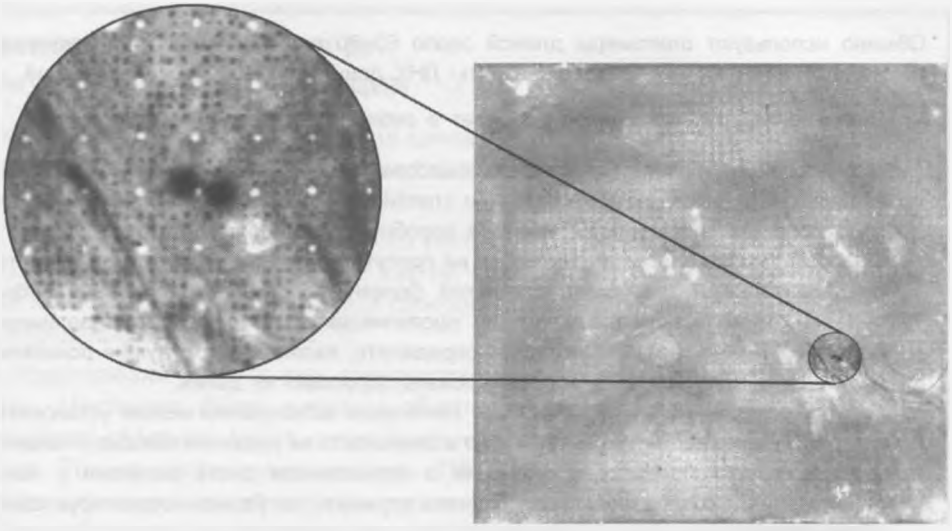


Рис. 2.1. Работа иммунной системы зависит главным образом от выбора правильной комбинации веществ в устойчивом комплексе антитело—антиген. Такую картину можно увидеть в теле позвоночного животного при эксперименте, в котором антитела находятся на поверхности фага или в заданной области. Показан случай, при котором примерно 1000–5000 белков человеческого мозга проверялись на сродство к антителам. Белки были выработаны специальными бактериальными клонами, видимыми как точки на мембране. Экспрессированные белки были получены при помощи клеточного лизиса. Результат обработали смесью фрагментов 12 антител (содержащих антигенсвязывающий сайт), и связанные антитела были визуализированы и обнаружены авторадиографией. Каждый клон отмечен дважды, чтобы способствовать идентификации нужных клеток; этим объясняется дублированное изображение. Развитие этого метода позволит провести обширный скрининг более широких типов белковых взаимодействий

Для проведения параллельного гибридизационного анализа большое количество разнообразных олигомеров ДНК крепится к определенным местам на неподвижную подложку в обычном двумерной матрице. Смесь, которую надо анализировать, готовят с использованием радиоактивных или флуоресцентных меток, чтобы можно было обнаружить гибриды. После совмещения чипа и смеси получается, что каждый элемент чипа, к которому прикрепился некоторый компонент смеси, несет радиоактивную или флуоресцентную метку.

ДНК-чип может содержать пробу, состоящую из 100 000 олигомеров. Заметим, что эта цифра больше чем общее количество генов даже в высших организмах. Размер точки может быть диаметром всего около ~150 мкм (0,15 мм). Сам же чип же обычно в поперечнике составляет несколько сантиметров.

Чтобы измерить экспрессию, берут образец олигомеров, который представляет собой кДНК или фрагменты кДНК (cDNA), отражающие мРНК различных генов.

Обычно используют олигомеры длиной около 50–80 пар оснований. Для анализа генотипа используют геномные фрагменты ДНК длиной 500–5000 пар оснований.

Применение ДНК-микрочипов включает в себя следующие области.

- *Исследование строения клетки и процессов, происходящих в ней.* Экспрессия, зависящая от изменений в состоянии клетки, может пояснить механизмы таких процессов как споруляция или смена аэробного метаболизма на анаэробный.
- *Диагностика заболеваний.* Проверка на присутствие мутаций может подтвердить предварительный диагноз генетической болезни, включая обнаружение заболеваний, характеризующихся поздно проявляющимися симптомами (например, болезнь Хантингтона). Это позволит определить, являются ли будущие родители носителями гена, который мог бы угрожать здоровью их детей.
- *Генетическая предрасположенность.* Некоторые заболевания нельзя установить с полной уверенностью по генотипу, но возможность их развития связана с генами и их экспрессией. Человек, знающий о повышенном риске развития у него заболевания, может в некоторых случаях улучшить ситуацию, корректируя свой образ жизни.
- *Выбор лекарств.* Обнаружение генетических факторов, которые определяют реакцию на медикаменты, указывает на то, что для некоторых пациентов выбранное лечение может быть неэффективным или даже вызвать серьезные неблагоприятные реакции.
- *Классификация заболеваний.* Разные виды лейкемии характеризуются разной экспрессией генов. Знание о конкретном типе заболевания важно для выбора оптимального лечения.
- *Выбор мишеней для создания медикаментов.* Белки, для которых показана усиленная транскрипция при развитии некоторой болезни, могут изучаться как мишень для фармакологического вмешательства (при условии, что найдется доказательство того, что усиление транскрипции существенно для развития заболевания).
- *Сопротивляемость патогенам.* Сравнение генотипов и экспрессии у штаммов бактерий, восприимчивых или устойчивых к антибиотикам, указывает на то, что белки вовлечены в механизм сопротивляемости организма.

Для биоинформатики ДНК-чипы являются пока всего лишь одним из перспективных направлений для получения информации. Это влечет за собой типичные проблемы создания эффективных архивов и систем для нахождения информации. Преимущество ДНК-микрочипов состоит в том, что получаемая информация настолько нова, что эта область не была загромождена форматами данных, основанными на устаревшем оборудовании и программах. (Следует, однако, иметь в виду, что данные, полученные с помощью таких массовых экспериментов, как экспрессионные чипы, несут в себе весьма значительный уровень информационного шума. — *Прим. ред.*)

Отслеживание передачи генетической информации

То, как наследственная информация хранится, передается и осуществляется, — наверно, фундаментальная проблема биологии. Существует три основных типа генетических карт (см. врезку, с. 90):

1. Генетические карты сцепленности генов.
2. «Бэндовые» схемы хромосом — карты чередования темных и светлых полос на хромосоме при специальном способе их окрашивания.
3. Последовательности ДНК.

Они представляют собой три совершенно разных типа данных. Гены, описанные Менделем, были в целом абстрактными объектами. Хромосомы — физические объекты, их отличительными признаками являются «бэндовые» модели. Только последовательностям ДНК непосредственно определена роль хранения наследственной информации в физической форме.

Очень большим достижением прошлого столетия в биологии было установление связи между этими тремя типами данных. Первые шаги — они были гигантским прогрессом — доказали, что для каждой хромосомы карты являются одномерными множествами, и, безусловно, они коллинеарны. Любой школьник сейчас знает, что гены вытянуты вдоль хромосомы, и что каждый ген соответствует последовательности ДНК. Но доказательства этих постулатов заработали большое количество Нобелевских премий.

Расщепление длинной молекулы ДНК — например, ДНК целой хромосомы — на фрагменты удобного для клонирования и секвенирования размера нуждается в дополнительных картах для описания порядка фрагментов, так, что целая последовательность может быть восстановлена из последовательностей фрагментов. Рестрицирующая эндонуклеаза — фермент, разрезающий ДНК на отдельные последовательности, обычно около 6 пн длиной. Разрезание ДНК несколькими рестрицирующими ферментами с разной специфичностью образует набор перекрывающихся фрагментов. Основываясь на размерах фрагментов можно построить рестрикционную карту, основанную на порядке и расстоянии между сайтами расщепления рестрицирующих ферментов. Мутация в одном из этих сайтов расщепления изменит размеры фрагментов, произведенных соответствующим ферментом, и мутация будет локализована на карте.

Рестрицирующие энзимы могут производить довольно большие фрагменты ДНК. Разрезание ДНК на более маленькие части, которые затем будут клонированы и упорядочены по перекрываниям последовательности (как пример использование текста на с. 33), в результате дает лучший анализ ДНК, называемый картой контигов.

В прошлом связи между хромосомами, генами и последовательностями ДНК были существенны для определения молекулярных дефектов, подчеркивающих наследственные заболевания, такие как болезнь Хантингтона или кистозный фиброз. Секвенирование человеческого генома радикально изменило ситуацию. Тем не менее болезнь Хантингтона и кистозный фиброз являются

Генетические карты, хромосомные карты и карты последовательностей

1. *Генетическая карта* классически определена наблюдаемыми моделями наследования. Химически связанные группы и частоты рекомбинации могут определить: находятся ли гены на одной хромосоме, или на разных, а для генов одной хромосомы — как далеко они расположены. Принцип состоит в том, что хромосомы рекомбинируют в процессе мейоза посредством кроссинговера. Поэтому, два гена одной хромосомы, находящиеся очень далеко будут казаться не сцепленными. Единица длины генетической карты — Морган, определенный таким отношением, что 1 сМ отвечает 1% частоты рекомбинации. (Известно, что 1 сМ $\sim 10^6$ пн у человека, но значение варьирует в зависимости от расположения в геноме и расстояния между генами.)
2. *Карты объединения «бэндов» хромосом.* Хромосомы — физические объекты. Бэнды — видимые детали на них. Номенклатура такова: Во многих организмах хромосомы пронумерованы в порядке размера, 1 — самая большая. Два плеча человеческой хромосомы, разделенные центромерой, называются p (petite = короткий) плечом и q (= queue = хвостовой) плечом. Участки внутри хромосомы названы p1, p2, ... и q1, q2 ... наружу от центромеры. Последующие цифры показывают подразделение бэндов. Например определенные бэнды на q плече 15 хромосомы человека помечены 15q11.1, 15q11.2, 15q12. Первоначально были определены бэнды 15q11 и 15q12, затем при более детальном анализе15q11 было разделено на 15q11.1 и 15q11.2. Делеции, содержащие этот участок связаны с синдромами Прайдера—Вилли и Англемана. Эти синдромы имеют интересную особенность: клинические последствия зависят от того была пораженная хромосома унаследована от отца или от матери. Эти наблюдения генетического следа показывают, что генетическая информация в оплодотворенной яйцеклетке — не просто голые последовательности ДНК, внесенные родителями. Хромосомы отцовского и материнского происхождения имеют различные состояния метилирования, что сигнализирует о разной экспрессии генов (это явление называется импринтингом. — *Прим. ред.*). Процесс модификации ДНК, имеющий место на протяжении видоизменения в развитии, начинается уже в зиготе.
3. *Собственно последовательность ДНК.* Физически последовательность нуклеотидов в молекуле, из расчетов — цепочка типа A, T, G и C. Гены — это участки последовательности, во многих случаях прерванные некодирующими участками.

следствием дисфункции одного белка, кодируемого одним геном, и не связаны с окружающими факторами или стилем жизни. В противоположность много похожих расстройств, включая диабет и рак, вызываются поражением многих генов.

Данные болезни могут быть отнесены к дефектам белков, поэтому:

- Если мы знаем, какой белок вовлечен в заболевание, мы можем использовать рациональные подходы к лечению.
- Если мы знаем, что вовлечен ген, то мы можем разработать тесты для идентификации больных или носителей.

- Во многих случаях, знания о расположении хромосом в гене неважны для какой либо терапии или определения; они требуются только для идентификации гена, связывающего модели наследования и последовательность ДНК. (Это не справедливо для расстройств возникающих из хромосомных аномалий.)

Например, в случае серповидно-клеточной анемии, мы знаем, что участвует определенный белок. Это заболевание является следствием точечной мутации в гемоглобине. Мы можем перейти непосредственно к разработке лекарства. Последовательность ДНК нам нужна только для генетических тестов и консультаций. Напротив, если нам не известен ни белок, ни ген, мы должны как-нибудь перейти от фенотипа назад к гену — процесс, называемый позиционным клонированием обратной генетики. Использованное позиционное клонирование вызывает подобие каскада от генетической карты к хромосомной карте к последовательности ДНК. (Позже мы увидим, как недавние разработки замкнули этот процесс.)

Модели наследования определяют тип генетического дефекта, ответственного за состояние. Они показывают, например, что болезнь Хантингтона и кистозный фиброз вызываются единичными генами. Чтобы найти ген, связанный с кистозным фиброзом, было необходимо начать с генетической карты, используя модели сцепленности наследственности в пораженных семьях, чтобы локализовать пораженный ген в определенном регионе определенной хромосомы. Зная нужный регион хромосомы, далее можно искать ДНК этого региона, чтобы определить кандидатные гены, и, в конце концов, окончательно определить ген, ответственный за расстройство и секвенировать его (см. врезку, с. 92). В противоположность этому, многие расстройства не обнаруживают простого наследования, или, даже если вовлечен только один ген, наследственность создает только склонность к клиническим последствиям, проявление которых зависит от окружающих факторов. Полная последовательность генома человека и измерения моделей экспрессии будут необходимы для определения генетических составляющих в таких более сложных случаях.

Соответствие между картами

Сопоставленные генетические карты сцепленности генов могут быть сопоставлены с соответствующими хромосомными моделями с помощью наблюдения отдельных частей хромосом с делециями или транслокациями. Эти гены ответственны за фенотипические изменения, связанные с делецией, и должны быть расположены внутри этой делеции. Транслокации коррелируют с изменением в последовательности и рекомбинации.

Было создано множество методов для выявления соответствия между рисунком дифференциальной окраски хромосом с последовательностью генов:

- Во флуоресцентной гибридизации *in situ* (FISH) зонд соединяется с флуоресцентной меткой. Далее он гибридизуется с исследуемой ДНК. Локализация флоуресцентной метки на хромосоме фиксируется на фотографическом снимке (см. цветную иллюстрацию IV). Разрешение метода обычно $\sim 10^5$ пн, но с помощью некоторых современных техник удается улучшить

Идентификация гена кистозного фиброза

Кистозный фиброз — болезнь, известная по крайней мере со средних веков и изучаемая около 500 лет, обусловлена наследственной рецессивной аутосомной аномалией. Его симптомы включают закупорку кишок, пониженную плодовитость, включая анатомические аномалии (особенно у мужчин); и периодическое засорение и инфекции легких — основная причина смерти сейчас, так как существует эффективное лечение желудочно-кишечных симптомов. Около половины больных умирает до 25 лет, и немногие доживают до 50. Кистозный фиброз поражает 1/2500 часть особей Американского и Европейского населения. Приблизительно 1/25 белых и 1/65 афро-американцев являются носителями мутантного гена. Белок, дефектный при кистозном фиброзе, также действует как рецептор для проникновения в организм *Salmonella typhi* — микроба, вызывающего тиф. Повышенная устойчивость к тифу у гетерозигот — у которых не развивается кистозный фиброз, но которые являются носителями мутантного гена — возможно, объясняет, почему ген не был элиминирован из популяции.

Модель наследования показывает, что кистозный фиброз — эффект одного гена. Тем не менее, неизвестно, какой конкретно белок вовлечен. Он должен быть найден посредством генетического анализа.

Клинические наблюдения предоставили генетические маркеры, с полезной информативностью. Было известно, что проблемой был транспорт ионов хлора в эпителиальных тканях. В народе давно определено, что дети с чрезмерным содержанием солей в поту (что ощущалось, когда целуешь младенца в лоб) — жили недолго. Современные физиологические исследования показали, что эпителиальные ткани пациентов, больных кистозным фиброзом не могут реабсорбировать хлорид. При приближении к гену, подходящими проводниками были ожидаемое распределение его экспрессии между тканями и тип вовлеченного белка.

В 1989 г. ген кистозного фиброза был изолирован и секвенирован. Этот ген, называемый CFTR (cystic fibrosis transmembrane conductance regulator), кодирует белок из 1480 аминокислотных остатков, который в норме формирует cAMP-регулируемый эпителиальный Cl^- канал. Ген, включающий 24 экзона, занимает участок длиной около 250 тыс. пар нуклетидов. Для 70% мутантных аллелей характерна мутация — делеция трех пар оснований, удаляющая остаток 508Phe из белка. Эти мутации обозначаются как del508. Эффект этой делеции — дефектное перемещение белка, который уничтожается в эндоплазматическом ретикулуме до того, как транспортируется в мембрану клетки.

In vitro тест для кистозного фиброза основан на выделении эмбриональной ДНК. Праймеры для ПЦР предназначены давать продукт длиной 154 пн из нормальной аллели и 151 пн продукт из del508 аллели.

Клиницисты использовали преимущество того факта, что пораженные ткани дыхательных путей легко доступны для экспериментов с генной терапией. Генетически созданный аденовирус, несущий правильный ген CFTR и распыленный в дыхательных путях, может передать его эпителиальным тканям.

Позиционное клонирование генов: выделение гена кистозного фиброза

Процесс, в результате которого был выделен данный ген, называется обратной генетикой.

- Поиск маркера данного гена в геномах родственников, имеющих это заболевание, показал, что ген находится рядом с минисателлитной ДНК (VNTR), DOCR-917. Это соответствует участку q3 седьмой хромосомы в гибридных соматических клетках.
- Были найдены маркеры, лежащие еще ближе к исследуемому гену. Таким образом, он оказался заключенным между VNTR в онкогене MET и второй VNTR — D7S8. Исследуемый ген, лежит на расстоянии 1.3 сМ и 0.9 сМ от MET и D7S8, соответственно. Данный участок ДНК оценивается в $1–2 \cdot 10^6$ пн и может содержать от 100 до 200 генов.
- Наследственный паттерн добавочных маркеров позволил локализовать данный ген еще более точно: в районе $5 \cdot 10^5$ пн. Методика, названная «прыжки по хромосоме», сделала исследования данного региона более эффективными.
- Справа от маркеров был клонирован участок ДНК, длиной в $3*10^5$ пн. Далее в изолированных пробах этого участка были найдены активные гены, характеризующиеся наличием вышележащей последовательности CCGG. (На этом этапе использовалась рестриктирующая эндонуклеаза HpaII, режущая ДНК только в случае, если второй С последовательности CCGG не метилирован, т. е. когда ген находится в активном состоянии.)
- Идентификация генов путем определения последовательности.
- Проверка этих генов в животных обнаружила четыре кандидатных гена, ответственных за заболевание. Проверка данных генов по библиотеке кДНК из потовых желез пациентов, страдающих данным заболеванием, а также использование контроля из здоровых людей, позволила сократить число кандидатов до одного гена. Большинство больных кистозным фиброзом, характеризуются трехнуклеотидной делецией в данном гене, приводящей к потере аминокислоты 508Phe из соответствующего белка.

Доказательства корректной идентификации:

- 70% аллелей гена кистозного фиброза имеют данную делецию. У людей, не страдающих данным заболеванием, она не обнаружена.
- Экспрессия дикой аллели гена в клетках, изолированных из пациентов, восстанавливает нормальный транспорт Cl^-.
- Выключение гомолога этого гена в мышах вызывает появление фенотипа кистозного фиброза.
- Паттерн экспрессии данного гена совпадает с органами, в которых ожидается возникновение кистозного фиброза.
- Белок, кодируемый этим геном, содержит трансмембранный домен, что согласуется с данными об участии данного белка в мембранном транспорте.

его до $\sim 10^3$ пн. Синхронное использование двух зондов позволяет выявить взаимное расположение последовательностей и оценить расстояние между ними. Данный подход незаменим для изучения биологических видов с длительным периодом генерации, так как стандартные генетические подходы в данном случае неэффективны. FISH также может использоваться для детектирования хромосомных аберраций.

- *Гибриды соматических клеток* — клетки грызунов, содержащие несколько человеческих хромосом или их частей. (Такие фрагменты получаются при облучении человеческих клеток). Гибридизация флоуресцентного зонда с набором таких гибридов позволяет детектировать хромосому, содержащую последовательность, интересующую нас. Однако такой подход был вытеснен использованием клонов дрожжей, бактерий или фагов, содержащих фрагменты человеческой ДНК в искусственных хромосомах (YACs, BACs, PACs соответственно).

Конечно, в наше время, когда почти вся последовательность ДНК известна, такие подходы считаются устаревшими.

Генетические карты высокого разрешения

Некогда гены были единственными используемыми маркерами геномов. Теперь используются также маркеры, не влияющие на фенотип. Гены по сравнению с ними слишком разбросаны по геному, чтобы на их основе строить достоверные карты. Сейчас, когда мы можем идентифицировать любую последовательность ДНК, в качестве маркеров используются любые последовательности, полиморфные у разных индивидуумов, включая:

- *Локусы с варьирующим числом тандемных повторов* (VNTRs), иногда зовущиеся минисателлитами. VNTRs содержат участки длиной 10–100 пн, повторенные различное число раз. У разных индивидуумов VNTRs, построенные на одном и том же мотиве, могут содержать различное число повторов. Различия в длине этих фрагментов могут использоваться в качестве генетических маркеров. Наследование VNTRs, можно проследить и привязать к какому-либо фенотипу. VNTRs первыми, из генетических данных, были использованы для идентификации личности в криминалистике, а также в судебных разбирательствах по вопросам отцовства. Данный метод получил название фингерпринт (генетические отпечатки пальцев).

  Некогда VNTRs исследовались на основе полиморфизма длины рестрикционных фрагментов (RFLPs). VNTRs почти всегда содержат несколько сайтов рестрикции для одной и той же рестриктазы, по которым их можно аккуратно нарезать. Результаты можно разделить на геле и провести Саузерн-блоттинг. Обратите внимание: VNTRs — характеристики генома, а RFLPs – искусственная смесь рестрикционных фрагментов, полученная в лаборатории для идентификации VNTRs.

  Сейчас для измерения длины VNTRs все чаще используется ПЦР (Полимеразная Цепная Реакция), данный метод почти полностью заменил использование рестриктаз в этой области.

- *Полиморфизм коротких тандемных повторов* (STRPs), часто называемых микросателлитами. STRPs — участки из 2–5 пн, повторенных большое число раз (10–30). Существует много преимуществ использования STRPs, одно из которых — достаточно равномерное распределение по геному.

Нет причин, по которым такие маркеры должны лежать внутри экспрессирующихся генов, и чаще всего так и происходит. (исключением являются CAG повторы в генах болезни Хантингтона и некоторых других заболеваний).

Набор микросателлитных маркеров значительно упрощает идентификацию генов. Весьма интересно сопоставить современные проекты по поиску генов заболеваний, когда известна последовательность человеческой ДНК, с классическими работами, к примеру, с идентификацией гена кистозного фиброза (см. врезку, с. 96).

Различные дополнительные техники картирования более точно работают с ДНК, что позволяет сократить процесс идентификации генов:

- *Контиг, или непрерывная карта клонов* — серия перекрывающихся ДНК клонов хромосомы в известной последовательности, хранящиеся в дрожжевых клетках в YACs (Yeast Artificial Chromosomes), BACs (Bacterial Artificial Chromosomes). Такая карта является весьма хорошим отображением генома. В YACs человеческая ДНК интегрирована в маленькую дополнительную хромосому дрожжевой клетки. Такая хромосома может содержать до 10^6 пн, и весь человеческий геном можно уместить в 10 000 YACs. В BACs ДНК вставлена в плазмиду *E. coli*. (Плазмида — отдельная маленькая двуспиральная ДНК, являющийся дополнением к основному геному, чаще всего она кольцевая.) BAC может нести до 250 000 пн. Несмотря на то что это меньше, чем у YAC, BACs более предпочтительны; так как они стабильнее, и с ними проще работать.
- *Ярлык, определенный последовательностью* (sequence tagged site STS) — короткий, секвенированный участок ДНК, обычно 200–600 пн в длину, локализованный в строго определенной области генома. Он не обязан быть полиморфным. STS может быть нанесен на карту генома с помощью ПЦР и клеток, содержащих непрерывную карту клонов.

  Один из типов STS возникает из ярлыков экспрессируемых последовательностей (EST), коротких фрагментов кДНК (комплементарной ДНК, т. е. последовательностей, полученных из мРНК экспрессирующихся генов). Последовательности EST содержат только экзоны, сплайсированные вместе в последовательность, кодирующую белок. кДНК может быть картирована на хромосому с помощью FISH или локализованы в карте контигов.

Как карта контигов и ярлыки последовательностей могут облегчить идентификацию генов? Если вы работаете с организмом, для которого не известна полная последовательность генома, но для которого есть полная карта контигов для всех хромосом, то вы можете идентифицировать STS-маркеры, плотно сцепленные с интересующим вас геном, а затем локализовать эти маркеры на карте контигов.

Идентификация гена, ответственного за возникновение синдрома Берардинелли—Сейпа

Синдром Берардинелли—Сейпа (врожденная генерализованная липодистрофия) — это аутосомное рецессивное заболевание, симптомами которого являются: отсутствие жировой прослойки, сахарный инсулинрезистентный диабет, а также гипертрофия скелета.

Для определения гена, который участвует в возникновении этого синдрома, научная группа под руководством Дж. Магре получила ДНК у родственников больных и подвергла ее анализу сцепления и гомозиготному картированию с помощью полногеномной панели, содержащей порядка 400 микросателлитных маркеров с известным расположением, находящихся на расстоянии примерно 10 сМ друг от друга. В данной процедуре для сравнения целой ДНК больных людей с ДНК здоровых была использована фиксированная панель праймеров, специфичных для амплификации и анализа каждого маркера. Измерения показали длины повторов, ассоциированных с каждым микросателлитом. Для каждого микросателлита каждая обнаруженная длина является аллелью. Идентификация микросателлитных маркеров, которые тесно связаны с фенотипом, дает информацию о локализации искомого гена. При измерениях использованы коммерческие наборы праймеров, инструменты и приборы.

Два маркера в хромосомных бэндах 11q13 — D11S4191 и D11S987 — разделяются при заболевании, и некоторые больные, которые родились в близкородственных семьях, являлись гомозиготными по этим маркерам. Генотипирование и картирование с использованием дополнительных маркеров, локализовало искомый ген на 11-й хромосоме в области 2,5 млн пн (для обозначения млн пн часто употребляют термин мегабаза).

В рассматриваемом 2,5-мегабазном участке и в его окрестности находятся 27 генов. Секвенирование ДНК нескольких пациентов показало наличие трехэкзонной делеции в одном из этих генов. Эти последовательности сравнили с последовательностями родственников больных и обнаружили корреляцию между нарушениями в этом гене и возникновением болезни, чем было доказано участие данного гена в развитии синдрома. При этом ни один из оставшихся 26 генов не показали подобной корреляции.

В более ранних исследованиях было показано участие другого гена BSCL1, в 9q34, у других семей с тем же синдромом. Ген BSCL1 еще не был идентифицирован. Возможно, что нарушения в этих двух генах вызывают один и тот же эффект из-за того, что продукты этих генов участвуют в общем биохимическом пути, который блокируется при нарушении одного из генов.

Ген BSCL2 на хромосоме 11 содержит 11 экзонов общей длиной 14 000 пар оснований. Он кодирует белок сейпин, содержащий 398 аминокислотных остатков. В этом гене найдены нарушения в виде больших и малых делеций, а также в виде одиночных аминокислотных замен. Эти нарушения, вызывающие смещения рамки считывания или укорочение, равносильны утрате функционального белка, а замена Ala212 на Pro, вызываемая миссенс-мутацией, однозначно связана со стабильностью элементов его вторичной структуры.

У мыши и дрозофилы имеются гомологи сейпина. Функции этих гомологов остаются неизвестными, хотя они предположительно содержат трансмембранные

спирали. Высокий уровень экспрессии в мозге и семенниках дает основания судить об этиологии некоторых аспектов этого синдрома. Это может иметь отношение к ранним эндокринологическим исследованиям у больных, страдающих синдромом Берардинелли—Сейпа. Данные исследования показали наличие нарушений в выделении питуитарных гормонов гипоталамусом. Открытие белка с неизвестной функцией, участвующего в возникновении синдрома, предоставляет возможность исследовать, вероятно, новый биохимический путь.

Локализация генов в геноме

Компьютерные программы для анализа геномов определяют открытые рамки считывания (ORF). ORF — это районы ДНК, которые начинаются со старт-кодона (ATG) (в прокариотах старт-кодонами могут также быть GTG и CTG. — *Прим. ред.*) и заканчиваются стоп-кодоном. ORF — это потенциальный белок-кодирующий фрагмент.

Подходы к идентификации белок-кодирующих областей основаны на следующих подходах:

1. *Определение районов похожих на известные белок-кодирующие области из других организмов.* Эти районы могут кодировать аминокислотные последовательности, похожие на известные белки, или могут быть похожими на EST. Поскольку EST определены из мРНК, они соответствуют генам, о которых известно, что они экспрессируются. Необходимо секвенировать всего несколько сотен начальных нуклеотидов кДНК, чтобы получить достаточно информации для идентификации гена: определение гена по EST аналогично индексированию стихов или песен по первым строкам.
2. *Методы поиска и идентификации генов ab initio (от начала), на основе только знания последовательности.* Компьютерная аннотация генома является более точной и полной для бактерий, чем для эукариот. Бактериальные гены сравнительно легко аннотировать, поскольку они непрерывны — в них нет интронов, характерных для эукариот, а межгенные промежутки достаточно малы. В высших организмах идентификация генов сложнее. Идентификация экзонов — одна из проблем, тесно связанная с другой проблемой — альтернативным сплайсингом.

Процедура предсказания генов *ab initio* в эукариотических геномах имеет следующие особенности:

- Начальный экзон (5′) начинается со старта транскрипции, перед которым расположен сайт основного промотора (core promoter), такого как TATA-бокс, обычно расположенного в 30 пн перед геном. Он обычно не содержит стоп-кодонов в рамке и заканчивается непосредственно перед сигналом сплайсинга GT. (Иногда некодирующий экзон предшествует экзону, содержащему стартовый кодон.)
- Внутренние экзоны, также как и начальные не содержат стоп-кодоны в рамке. Они начинаются сразу после сайта сплайсинга AG и заканчи-

ваются непосредственно перед сайтом сплайсинга GT. Предшествующий интрон содержит так называемый сайт ветвления и полипиримидиновый тракт, которые взаимодействуют с аппаратом сплайсинга.

- Конечный экзон (3′) начинается сразу после сайта сплайсинга AG и заканчивается стоп-кодоном, после которого идет сайт полиаденилирования. (Иногда некодирующий экзон следует за экзоном, в котором локализован стоп-кодон).

Все кодирующие последовательности имеют неслучайные характеристики, частично связанные с предпочтением использования кодонов. Эмпирически найдено, что статистика гексануклеотидов лучше всего различает кодирующие и некодирующие районы. Начиная с набора генов, известных для данного организма, можно построить обучающую выборку для настройки параметров для данного генома.

Аккуратное предсказание генов является критическим компонентом анализа последовательностей геномов. Эта проблема в настоящее время находится в фокусе внимания текущих исследований.

Геномы прокариот

Большинство прокариотических клеток хранят свой генетический материал в виде одной большой кольцевой молекулы с характерной длиной 5 млн пн. Кроме того, они могут содержать плазмиды.

Белок-кодирующие участки в бактериальных геномах не содержат интронов. Во многих прокариотических геномах белок-кодирующие области организованы в опероны — тандемно (друг за другом) расположенные гены, транскрибирующиеся в виде одной молекулы мРНК, и находятся под общим транскрипционным контролем. Многие опероны в бактериальных геномах кодируют функционально связанные гены. Например, последовательные гены в триптофановом опероне *E. coli* кодируют ферменты, катализирующие последовательные реакции биосинтеза триптофана. В археях не столь часто наблюдается взаимосвязь генов в оперонах.

Типичный прокариотический геном содержит сравнительно мало некодирующей ДНК (в сравнении с эукариотами), распределенной вдоль генома. В *E. coli* только 11% ДНК представлено некодирующей последовательностью.

Геном бактерии *Escherichia coli*

E. coli, штамм K-12, уже давно является «рабочей лошадкой» в молекулярной биологии. Геном штамма MG1655, опубликованный в 1997 г. группой Ф. Блаттнера (F. Blattner) из Висконсинского университета, содержит 4 639 221 пар оснований в однотяжевой кольцевой молекуле ДНК, не содержащей плазмид. Приблизительно 89% последовательности несет информацию о белках или структурных РНК. Аннотация показывает наличие:

- 4285 белок-кодирующих генов
- 122 генов структурных РНК

- некодирующие повторяющиеся последовательности
- регуляторные элементы
- транскрипционные/трансляционные управляющие элементы
- транспозазы
- остатки профагов
- инсерционные элементы последовательности
- вставки нетипичных фрагментов, предположительно инородных элементов, принесенные с помощью горизонтального переноса.

Анализ последовательности генома заключался в идентификации и аннотации белок-кодирующих генов и других функциональных областей. В результате многих лет интенсивных исследований многие белки *E. coli* были известны еще до того, как секвенирование было завершено: 1853 белка были описаны до публикации геномной последовательности. По аналогии с гомологами, най-

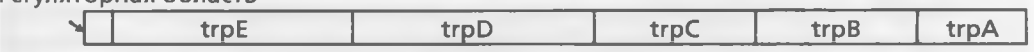


Рис. 2.2. Trp оперон в *E. coli* начинается с контрольной области, содержащей промотор, оператор и главные последовательности. Пять структурных генов кодируют белки, которые катализируют последовательные этапы в синтезе аминокислоты триптофан (tryptohpan) из ее предшественника — хоризмата (chorismate):

$$\text{Хоризмат} \xrightarrow[(1)]{} \text{Антранилат} \xrightarrow[(2)]{} \text{фосфорибозил-антранилат} \xrightarrow[(3)]{}$$

$$\xrightarrow[(3)]{} \text{Индолглицерол-фосфат} \xrightarrow[(4)]{} \text{индол} \xrightarrow[(5)]{} \text{Триптофан}$$

Стадии реакции: (1) trpE и trpD кодируют 2 компонента антранилатсинтазы. Этот тетраметрический фермент, состоящий из 2 копий каждой субъединицы, катализирует преобразование хоризмата в антранилат. (2) белок, кодируемый trpD, катализирует последующую фосфорибозилацию антранилата. (3) trpC кодирует другой бифункциональный фермент — фосфорибозилантранилатизомеразу — индолглицеролфосфатсинтазу. Он превращает фосфорибозилантранилат в индолглицеролфосфат через интермедиат фосфат карбоксифениламиндиоксирибулозы. (4 и 5) trpB и trpA кодируют соответственно α и β субъединицы третьего бифункционального фермента — триптофансинтазы (тетрамер $\alpha_2\beta_2$). Интермедиатиндол, выработанный α-субъединицей (реакция индолглицеролфосфат $\rightarrow$ индол), диффундирует без контакта с растворителем в активный сайт β-субъединицы внутри структуры этого фермента, где индол превращается в триптофан.

Обособленный ген trpR, не связанный напрямую с этим опероном, несет информацию о репрессоре trp. Репрессор способен связываться с операторной последовательность в ДНК (с в регуляторной области) только в том случае, если он связан с триптофаном. Связывание репрессора блокирует доступ РНК полимеразы к промотору в случае избытка триптофана. Дальнейший контроль транскрипции в зависимости от содержания триптофана вызван аттенюатор-элементом внутри лидирующей последовательности мРНК. Аттенюатор (1) состоит из лидерного пептида, ген которого содержит 2 последовательно расположенных парных trp кодона и (2) может альтернативно сворачиваться в разные по форме вторичные структуры, одна из которых завершает транскрипцию. Концентрация триптофана определяет количество аминоацилированной trp-тРНК, что влияет на темп продвижения рибосомы через последовательно расположенные trp кодоны. Потеря скорости (стайлинг) рибосомы на последовательно расположенных trp кодонах в ответ на низкий уровень триптофана препятствует формированию вторичной структуры мРНК, терминирующей транскрипцию.

денными в банках последовательностей, стало возможным приписать функции также другим генам. Чем уже область специфичности функций этих гомологов, тем более точно могло быть определено их распределение. В настоящее время более 60% белков могут быть приписаны хотя бы основные функции (см. таблицу на с. 101). Другие области генома, такие как регуляторные сайты или как мобильные генетические элементы, также опознаны на основе сходства с последовательностями известных гомологов из других организмов.

Распределение белок-кодирующих генов по геному *E. coli*, кажется, не следуют ни одним простым правилам, ни по положению на хромосоме ДНК, ни по ориентации. На самом деле сравнение штаммов показывает, что гены непостоянны.

Геном *E. coli* представляет собой относительно плотно упакованные гены. Гены, несущие информацию о белках или структурных РНК, занимают примерно 89% последовательности. Средний размер открытой рамки считывания (ORF) составляет 317 аминокислот. Если бы даже гены были распределены равномерно, то среднее межгенное расстояние составляло бы 130 пар оснований; наблюдаемое же среднее расстояние между генами — 118 пар оснований. Тем не менее расстояние между генами значительно варьируется. Есть большие межгенные участки. Они содержат регуляторных сигналы и повторяющиеся последовательности. Самый длинный межгенный участок (1730 пар оснований) содержит некодирующие повторяющиеся последовательности.

Приблизительно три четверти транскрипционных единиц содержат только 1 ген; оставшиеся содержат несколько последовательных генов или оперонов. Было оценено, что геном *E. coli* содержит 630–700 оперонов. Опероны варьируются в размере, хотя только некоторые содержат больше, чем 5 генов. Гены внутри (одного) оперона, как правило, имеют взаимосвязанные функции.

В некоторых случаях она и та же последовательность ДНК кодирует части более чем одной полипептидной цепи. Один ген несет информацию о τ- и γ-субъединицах ДНК полимеразы III. Трансляцию полного гена формирует τ-субъединица. Ψ-субъединица гомологична двум третям N-конца τ-субъединицы. Сдвиг рамки генетического кода на рибосоме на этой точке ведет к обрыву роста цепи в 50% случаев, вызывая в соотношении 1 : 1 выработку τ- и γ-субъединиц. Они не выглядят как перекрывающиеся гены, в которых все различные считывающие рамки несут информацию об экпрессированных белках.

В других случаях одни и те же полипептидные цепи фигурируют в нескольких ферментах. Белок, сам по себе функционирующий как липоатдегидрогеназа, также является субъединицей пируватдегидрогеназы, 2-оксоглутаратдегидрогеназы и глицинового расщепляющего комплекса.

Имея полный геном, мы можем протестировать белковый ассортимент *E. coli*. Самый большой класс белков — ферменты; их кодирующие последовательности занимают примерно 30% от всех генов. Многие ферментативные функции распределены по нескольким белкам. Некоторые из таких наборов функционально сходных ферментов близкородственные, вероятно появившиеся в результате дупликации либо в самой *E. coli*, либо в ее предшественнике. Другие наборы функционально сходных ферментов имеют очень несхожие

Распределение белков *E. coli* по 22 функциональным группам

Функциональный класс	*Number*	*%*
Регуляторная функция	45	1.05
Предполагаемые регуляторные белки	133	3.10
Структура клетки	182	4.24
Предполагаемые мембранные белки	13	0.30
Предполагаемые структурные белки	42	0.98
Фаги, транспозоны/длинные «прыгающие» гены, плазмиды	87	2.03
Транспортные и связывающие белки	281	6.55
Предполагаемые транспортные белки	146	3.40
Обмен энергии	243	5.67
Репликация, рекомбинация, модификация и репарация ДНК	115	2.68
Транскрипция, синтез РНК, метаболизм и модификация	55	1.28
Трансляция, посттрансляционное изменение белков	182	4.24
Процессы в клетке (включая адаптацию, защиту)	188	4.38
Биосинтез кофакторов, протестические группы и переносчики	103	2.40
Предполагаемые шапероны	9	0.21
Биосинтез нуклеотидов и метаболизм	58	1.35
Биосинтез аминокислот и метаболизм	131	3.06
Метаболизм жирных кислот и фосфолипидов	48	1.12
Катаболизм карбоновых соединений	130	3.03
Метаболизм центрального посредника (интермедиата)	188	4.38
Предполагаемые ферменты	251	5.85
Другие известные гены (генные продукты или известные по фенотипу)	26	0.61
Гипотетические, неклассифицированные, неизвестные	1632	38.06

[Blattner et al (1997) 'The complete genome sequence of *Escherichia coli* K12'; Science277, 1453–62].

между собой последовательности и различаются по специфичности, регуляции или внутриклеточной локализации.

Некоторые особенности набора ферментов *E. coli* дают гибкость, которая позволяет *E. coli* расти и конкурировать при различных условиях среды:

- Могут синтезироваться все компоненты белков и нуклеиновых кислот (аминокислоты и нуклеотиды) и кофакторы.

- У *E. coli* есть метаболическая гибкость: возможны как бескислородный так и кислородный рост с использованием различных путей получения энергии. *E. coli* может расти на многих различных источниках углерода и азота. Не все метаболические пути постоянно активны; альтернативы позволяют отвечать на изменения в условиях.
- Широкий круг транспортеров позволяют поглощать множество типов питательных веществ.
- Даже для специфических метаболитических реакций существует большое количество различных ферментов. Это предоставляет дополнительные возможности к способности подстраивать метаболизм к изменяющимся условиям. Комплексы регуляторных механизмов строят паттерн белковой экспрессии.
- Тем не менее *E. coli* не владеет полным набором ферментативных способностей. Она не может фиксировать CO_2 или N_2.

Здесь мы описали некоторые общие черты генома *E. coli* и его белкового набора. Исследования продвигаются в сторону динамических аспектов, к исследованию профилей экспрессии белков.

Геном архея *Methanococcus jannaschii*

С. Луриа предположил, что для определения общих черт живых организмов не обязательно изучать все живое, а можно найти наиболее отличающийся от нас организм и определить, что у нас с ним общего. Согласно этому предположению, необходимо найти организм, адаптированный к среде, наименее похожей на нашу.

В глубине океана были обнаружены места, настолько же отличающиеся от привычных нам, как описанные в научной фантастике. Гидротермальные выходы — это глубоководные вулканы, выпускающие горячую лаву и газы сквозь трещины в океаническом дне. Они создают экологические ниши для сообществ живых организмов, не связанных с поверхностью и использующих в качестве источников неорганических веществ минералы, выделяемые сквозь трещины в дне океана. Эти сообщества — единственные известные формы жизни, не зависящие прямо или косвенно от солнечного света как источника энергии.

Микроорганизм *Methanococcus jannaschii* был взят из гидротермального источника глубиной 2600 м на побережье в Мексике (Baja California) в 1983 г. Это термофильный организм, живущий при температурах 48-94°C, комфортно развивается при 85°C. Это облигатный анаэроб, он способен к самовоспроизведению из неорганических веществ. Общее уравнение его метаболизма — синтез метана из H_2 и CO_2.

M. jannaschii принадлежит к археям — одному из трех надцарств (бактерии, археи и эукариоты). К археям относится несколько групп прокариот, включая организмы, адаптированные к экстремальным условиям среды, таким как высокие температуры и давление или высокая концентрация соли.

Геном *M. jannaschii* был секвенирован в 1995 г. в Институте исследования геномов (TIGR — The Institute for Genomic Research). Это был первый секве-

нированный геном архея. Он включает большую хромосому, образованную кольцевой двухтяжевой молекулой ДНК длиной 1664976 пн и 2 экстрахромосомных элемента по 58407 и 16550 пн соответственно. Из 1743 найденных кодирующих регионов 1682 располагаются на хромосоме и 44 и 12 — на большом и малом экстрахромосомных элементах соответственно. Некоторые РНК содержат интроны. Как и в других прокариотических геномах, здесь немного некодирующей ДНК.

По-видимому, *M. jannaschii* удовлетворяет цели С. Лурии найти самого отдаленного от нас из ныне живущих организмов. Сравнение его генома с другими показывает, что относительно других форм жизни он — очень дальний родственник. Только для 38% открытых рамок считывания удается определить функцию на основе сходства с белками из других организмов (Теперь это можно сделать для более 50% ORF.) Однако ко всеобщему удивлению, археи по некоторым свойствам ближе к эукариотам, чем бактерии! Они, образно говоря, представляют собой сложную смесь. У архей белки, участвующие в транскрипции, трансляции и регуляции, более походят на эукариотические, а участвующие в метаболизме белки — на бактериальные.

Геномы наиболее просто организованных организмов: *Mycoplasma genitalium*

Mycoplasma genitalium — патогенная бактерия, вызывающая негонококковый уретрит.

Ее геном был секвенирован в 1995 г. при сотрудничестве групп из TIGR, Университета Джонса Хопкинса (The Johns Hopkins University) и Университета Северной Каролины (The University of North Carolina). Он представляет из себя одну молекулу ДНК, содержащую 580 070 пн. Это наименьший из известных сегодня геномов. Таким образом, *M. genitalium* наиболее близка к *самому малому* организму, способному к независимой жизни (вирусам же необходимы клеточные структуры хозяина).

Этот геном плотный в кодирующих участках. Для 468 генов были найдены соответствующие им белки; 85% всей последовательности является кодирующей. Средняя длина кодирующего региона — 1040 пн. Как и у других бактерий, кодирующие регионы не содержат интронов. Дальнейшее сжатие генома достигается перекрыванием генов. По-видимому, многие перекрывания возникли при утере стоп-кодонов.

Часть генов *M. genitalium* кодируют белки, важные для ее независимого воспроизведения (участвующие в репликации ДНК, транскрипции, трансляции и другие), рибосомальные и транспортные РНК. Другие гены нужны только для патогенной деятельности. Они кодируют белки адгезины, участвующие в связывании с инфецированной клеткой, другие молекулы для защиты от иммунной системы хозяина — множество транспортных белков. Как адаптация к паразитическому образу жизни широко распространена потеря ферментов, участвующих в метаболизме, в том числе участвующих в биосинтезе аминокислот; в самом деле, одна аминокислота отсутствует во всех

белках *M. genitalium* (на самом деле это неверно. — *Прим. ред.*) (см. Интернет-задание 2.7).

Геномы эукариот

Обнаружение совершенно нового мира с неожиданными феноменами — такие события в науке встречаются довольно редко. Геном эукариот открывает перед нами такой сложный мир (см. таблицу).

Описание эукариотического генома

Умеренные повторы

- Функции которых известны
  - распространенные семейства генов, например, актин, глобин
  - тандемные повторы генов (tandem gene family arrays)
    - гены рРНК (250 копий)
    - гены тРНК (у человека 50 сайтов с 10–100 копиями в каждом)
    - гистоновые белки у многих видов
- Функции которых неизвестны
  - короткие распространенные элементы (SINEs — short interspersed elements) например, Alu (некоторая функция в регуляции генов) длиной 200–300 пн. Сотни тысяч копий (300 000 Alu) располагаются отдельно (не в тандемных повторах)
  - длинные распространенные элементы (LINEs — long interspersed elements) длиной в 1–5 тпн. в геноме присутствует 10–10 000 копий

Частые повторы

- Минисателлиты
  - состоят из повторяющихся участков по 14–500 пн
  - 1–5 тпн в длину
  - большое разнообразие
  - разбросаны по всему геному
- Микросателлиты
  - состоят из повторяющихся участков длиной до 13 пн
  - в длину порядка сотен тпн
  - около 10^6 копий на геном
  - составляют большую часть гетерохроматина вокруг центромеры
- Теломеры
  - состоят из малых повторяющихся участков (обычно 6 пн: TTAGGG в геноме человека, TTGGGG у *Paramecium*, TAGGG у трипаносом, TTTAGGG у *Arabidopsis*)
  - 250–1000 повторов на конце каждой хромосомы

В эукариотических клетках абсолютное большинство ДНК находится в ядре в виде нуклеопротеидных структур — хромосом. Каждая хромосома содержит одну двухтяжевую молекулу ДНК. Небольшое количество ДНК есть

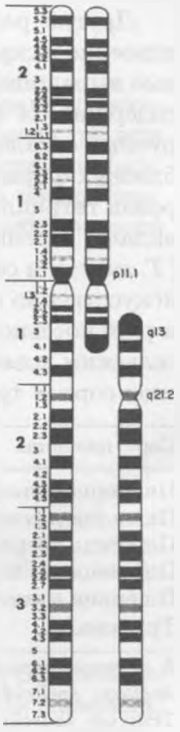


Рис. 2.3. *Слева* — 2-я хромосома человека. *Справа* — хромосомы из шимпанзе. (из Yunis, J. J., Sawer, J. R., и Dunham, K. (1980). The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee, *Science* 208, 1145–8. С разрешения © 1980 American Association for the Advancement of Science.)

и в органеллах — митохондриях и хлоропластах. Эти органеллы происходят от паразитов, живших когда-то внутри клеток.

Геномы органелл обычно имеют вид кольцевой двухтяжевой ДНК, но иногда могут быть линейными или состоять из множества кольцевых ДНК. Генетический код органелл отличается от такового у ядерных генов.

Ядерные геномы разных видов могут быть очень разными по размеру (см. с. 83). Корреляция между размером генома сложностью организма — довольно сильно неоднозначна. Она определенно не поддерживает того предвзятого мнения, что человек стоит на вершине развития. Во многих случаях, разницу в размерах генома определяет количество простых повторяющихся участков, которых часто относят к «ДНК-хламу» («junk DNA'»). (Сидни Бреннер предложил удобный признак «хлама», чтобы отличить его от «мусора» («garbage»): мусор вы выбрасываете, хлам вы держите рядом.)

В дополнение к вариации содержания ДНК, эукариоты различаются количеством хромосом и распределением генов по ним. Некоторые различия в распределении включают транслокации, или хромосомные фрагментации или соединения. Например, у человека 23 пары хромосом; у шимпанзе — 24. 2-я человеческая хромосома соответствует слиянию 12-й и 13-й хромосомы шимпанзе (см. рис. 2.3). Сложность в коньюгации хромосом во время митоза в зиготе после такого события может вносить вклад в репродуктивную изоляцию, требуемую для разделения видов.

Другие различия в хромосомном наборе отражают такие события, как удвоение и скрещивание. Пшеница однозернянка (*Triticum monococcum*), впервые выращиваемая как культура на Среднем Востоке 10 000–15 000 лет назад, содержит 14 пар хромосом. Пшеница двузернянка (*T. dicoccum*), культивируемая с палеолита, и пшеница твердая (*T. turgidum*) слились с гибридами близких форм пшеницы однозернянки с другими дикими растениями, сформировав тетраплоидные виды. Дополнительные скрещивания с разными дикими видами пшеницы дали гексаплоидные формы, такие как пшеница спельта (*T. spelta*), и современная пшеница мягкая (*T. aestivum*). *Triticale* — это гибрид, искусственно созданный путем скрещивания пшеницы твердой (*T. turgidum*) и ржи посевной (*Secale cereale*). Это сильное растение, созданное современным сельским хозяйством и в основном используемое для корма скота. Большинство сортов тритикале гексаплоидны.

Сорт пшеницы	Классификация	Хромосомный набор
Пшеница однозернянка	*Triticum monococcum*	AA
Пшеница двузернянка	*Triticum turgidum*	AABB
Пшеница твердая	*Triticum turgidum*	AABB
Пшеница спельта	*Triticum spelta*	AABBDD
Пшеница мягкая	*Triticum aestivum*	AABBDD
Тритикале	*Triticosecale*	AABBRR

A = геном исходной диплоидной пшеницы или близких форм, B = геном дикого злака *Aegilops speltoides* или *Triticum speltoides* или близких форм, D = геном дикого злака *Triticum tauschii* или близких форм, R = геном ржи *Secale cereale*.

Все эти виды до сих пор культивируются — некоторые в совсем небольших масштабах — и имеют индивидуальное использование в кулинарии. Пшеница спельта (*farro*, итал.) используется в кулинарии как основа знаменитого супа; макароны изготовляются из твердой пшеницы; хлеб — из *T. aestivum*.

Даже внутри отдельных хромосом в эукариотах довольно часто присутствуют целые семейства генов. Некоторые члены семейства являются *паралогами* — близкими генами, которые были дуплицированы в пределах одного генома и во многих случаях разошлись, чтобы осуществлять раздельные функции в видах-потомках. Изменения в экспрессии могут предшествовать развитию новых функций. (*Ортологи*, наоборот, — это гомологи в разных видах, и часто осуществляют одну функцию. Например, человеческие α и β глобины — это паралоги, а миоглобины человека и лошади — ортологи.)[1)] Другие схожие последовательности могут быть псевдогенами, которые могут появиться из-за дупликации, или ретротранспозиции из матричной РНК, за которой следуют накопление мутаций до потери функции. Хорошим примером является человеческий кластер генов глобинов (см. табл. на с. 107).

[1)] Более строгие (биологически) определения. Ортологами называются гены (и их продукты), которые разошлись в результате видообразования. Паралогами называются гены, которые разошлись в результате внутригеномной дупликации. — *Прим. ред.*

Кластер глобиновых генов человека

Гены и псевдогены гемоглобина человека расположены в кластерах на хромосомах 11 и 16. В нормальном взрослом организме синтезируются в основном три типа цепей глобина: α- и β-цепи, которые собираются в гемоглобиновые тетрамеры $\alpha_2\beta_2$, и миоглобин — мономерный белок, находящийся в мускулах. Другие гемоглобины, кодируемые в других генах, синтезируются на эмбриональной и постэмбриональной стадиях развития.

α-Генный кластер на хромосоме 16 имеет длину около 28 тпн. Он содержит три функциональных гена: ζ- и 2α-гены, идентичные в кодирующей части; три псевдогена: $\psi\zeta$, $\psi\alpha_1$ и $\psi\alpha_2$; а также еще один гомологичный ген, функция которого неясна — θ_1. β-генный кластер на хромосоме 11 имеет длину около 50 тпн. Он состоит из пяти функциональных генов: ε, два γ-гена (G_γ и A_γ), которые различаются на одну аминокислоту, гены δ и β, а также один псевдоген — $\psi\beta$. Ген миоглобина не связан ни с одним из этих кластеров.

Все гены гемоглобина и миоглобина человека имеют одну и ту же интрон/экзонную структуру. Они содержат три экзона, разделенные двумя интронами:

E = экзон, I = интрон. Длины участков на рисунке соответствуют гену β-глобина человека. Эта экзон/интронная структура консервативна в большинстве экспрессируемых генов глобина позвоночных, включая α- и β-цепи гемоглобина и миоглобин. Гены глобина растений имеют один дополнительный интрон, гены глобина *Paramecium* имеют на один интрон меньше, а гены глобина насекомых не содержат ни одного. Ген человеческого неврoглобина, недавно открытого гомолога, экпрессируемого в мозге на нижних уровнях, содержит 3 интрона, как и ген глобина растений.

Гемоглобиновые гены и псевдогены распределены по своим хромосомам, что отражает их эволюцию через дупликацию и расхождения.

Экспрессия этих генов строго следует стадиям развития организма. На эмбриональной стадии (до 6 недель после оплодотворения) синтезируются 2 цепи гемоглобина — ζ и ε, которые формируют тетрамер $\zeta_2\varepsilon_2$. После 6 недель с момента оплодотворения и до 8 недель после рождения, основной вид гемоглобина — постэмбриональный гемоглобин $\alpha_2\gamma_2$. Гемоглобин взрослого организма — $\alpha_2\beta_2$.

Талассемия — генетическая болезнь, связанная с дефектом или потерей генов гемоглобина. У большинства европеоидов имеются четыре гена нормального α гемоглобина взрослого организма, две аллели каждого из генов α_1 и α_2. Поэтому α-талассемия может иметь различную степень клинического проявления, в зависимости от того, сколько генов экспрессируют нормальные α-цепи. В норме только делеции, оставляющие менее двух активных генов, проявляются симптоматически. При генетических дефектах наблюдаются и делеции обоих генов (процесс который может происходить в местах совместного расположения генов и повторяющихся последовательностей, которые помогают кроссинговеру); и потеря стоп-кодона, что приводит к созданию протяженной нестабильной цепи.

β-Талассемия вызвана обычно точечными мутациями, включающими миссенс-мутации (замена аминокислоты, мутации с изменением смысла) и нонсенс-мутации (бессмысленные мутации, меняющие кодирующий кодон на стоп-кодон) приводящие к преждевременной терминации и к появлению процессированного белка, мутации в сплайс-сайтах или мутации в регуляторных участках. Определенные делеции, как нормального стоп-кодона, так и межгенного участка между δ- и β-генами, порождают δ-β-слитный белок.]

Геном *Saccharomyces cerevisiae* (пекарские дрожжи)

Дрожжи — один из самых простых известных эукариотических организмов. Клетки дрожжей, как и наши, состоят из ядра и других специализированных внутриклеточных компартментов. Расшифровка генома, которая была сделана благодаря очень эффективно работающему международному консорциуму, включающему ~100 лабораторий, была завершена в 1992 г. Геном дрожжей содержит 12 057 500 пар оснований ядерной ДНК, распределенной по 16 хромо-

сомам. Хромосомы различаются по размеру независимо от порядка важности, от 1352 тыс. пар оснований (тпн) четвертой хромосомы до 230 тпн первой.

Геном дрожжей содержит 5885 генов по расчетам кодирующих белки, ~140 генов для рибосомных РНК, 40 генов для малых ядерных РНК, и 275 генов транспортных РНК. По двум причинам геном дрожжей в кодирующих участках плотней, чем известные геномы более сложных эукариотов *Caenorhabditis elegans, Drosophila melanogaster* и человека: Во-первых, интроны относительно редки и относительно малы. Только 231 ген дрожжей содержит интроны; во-вторых, там меньше повторяющихся последовательностей по сравнению с более сложными эукариотами.

Считается, что ~150 млн лет назад произошла дупликация всего генома у дрожжей. За этим последовала транслокация кусочков дуплицированной ДНК и потеря одной из копий большинства (~92%) генов.

Из 5885 генов, которые потенциально кодируют белки, 3408 соответствуют известным белкам. Еще около 1000 кодируют белки, схожие с известными белками в других видах. Остальные ~800 сходны с открытыми рамками из других геномов, соответствующих неизвестным белкам. Многие из этих гомологов появляются у прокариот. Только у приблизительно одной трети белков дрожжей есть идентифицируемые гомологи в геноме человека.

При сборе сведений о генах имеет смысл группировать их функции по основным категориям. Следующая классификация дрожжевых белков взята из: `http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/`:

- Метаболизм
- Энергия
- Клеточный рост, клеточное деление, синтез ДНК
- Транскрипция
- Синтез белков
- Функции белков
- Пассивный транспорт
- Клеточный транспорт и транспортные механизмы
- Клеточный биогенез
- Межклеточное взаимодействие /передача сигналов
- Защита клеток, клеточная смерть и старение
- Ионный гомеостаз
- Клеточная организация
- Мобильные элементы, вирусные и плазмидные протеины
- Неклассифицированные

Дрожжи — это модель для тестирования методов определения функций продуктов генов. Поиск гомологов дает исчерпывающие результаты и сейчас продолжает использоваться. Существуют коллекции мутантов содержащих нокаут для каждого гена. (Уникальная последовательность — «штрих-код», вставленная в каждый мутант, облегчает идентификацию того мутанта, который проявляется при определенных условиях.) Открыты принципы клеточной локализации и экспрессии. На основе различных типов измерений, включаю-

щих в себя измерения, основанные на активации транскрипции пар протеинов, способных образовывать димеры, создаются каталоги межбелковых взаимодействий.

Геном *Caenorhabditis elegans*

Нематода *Caenorhabditis elegans* вошла в биологические исследования в 1960-х гг. благодаря Сиднею Бренеру, который предложил ее как организм, с одной стороны, достаточно сложный, но с другой — достаточно простой, на котором можно было осуществить полный анализ на клеточном уровне его развития и нервной системы.

Геном *C. elegans* был полностью получен в 1998 г. Это был первый геном многоклеточного организма, где была расшифрована полная последовательность ДНК. Геном *C. elegans* содержит ~97 млн пн ДНК, размещенной в парных хромосомах I, II, III, IV, V и X. Хромосома Y отсутствует. Пол у *C. elegans* определяется генотипом XX, самооплодотворяющийся гермафродит, и генотипом XO, мужской пол.

Геном *C. elegans* примерно в восемь раз длиннее генома дрожжей, количество предсказанных генов — 19 099 — примерно в три раза больше, чем для дрожжей. Плотность генов сравнительно низкая для эукакриотов, ~1 ген/5 тпн ДНК. Экзоны покрывают ~27% генома; гены содержат в среднем по 5 интронов. Около 25% генов находятся в кластерах родственных генов.

Многие белки *C. elegans* сходны с белками других форм жизни. Другие, по-видимому, специфичны для нематод. У 42% белков найдены гомологи вне типа, 34% гомологичны белкам других нематод и 24% не имеют гомологов вне вида *C. elegans*. Многие из белков были классифицированы по структуре и функциям.

Были определены некоторые виды генов РНК. Геном *C. elegans* содержит 659 генов тРНК, почти половина из них (44%) находятся на хромосоме X.

Распределение генов в *C. elegans*

Хромосома	Размер (млн пн)	Число кодирующих генов	Плотность кодирования (тпн/ген)	Число генов тРНК
I	7.9	2803	5.06	13
II	8.5	3259	3.65	6
II	7.6	2508	5.40	9
IV	9.2	3094	5.17	7
V	9.8	4082	4.15	5
X	10.1	2631	6.54	3

***C. elegans*:**
20 наиболее распространенных белковых доменов

Тип домена	Число
7-Трансмембранный хеморецептор	650
Домен эукариотической протеин-киназы	410
Два домена, цинковые пальцы типа C4	240
Коллаген	170
7-Трансмембранный рецептор (семейство родопсинов)	140
Цинковый палец типа C2H2	130
Лектин типа C	120
РНК-распознающий мотив	100
Цинковые пальцы типа C3HC4	90
Протеиновая тирозин-фосфатаза	90
Анкириновый повтор	90
WD домен G-бета повтора	90
Гомеобоксный домен	80
Ионный канал переносчика нейросигналов	80
Цитохром P-450	80
Консервативная C-терминальная хеликаза	80
Короткая цепь алкогольдегидрогеназы	80
UPD-глюкорозил- и UDP-глюкозил-трансферазы	70
EGF-подобный домен	70
Суперсемейство иммуноглобулинов	70

Источник: статья консорциума по секвенированию *C. elegans* от 11 декабря 1998, *Science*.

Сплайсосомные молекулы РНК появляются в виде множественных копий, часто идентичных. (Сплайсосомы — органеллы, конвертирующие пре-мРНК-транскрипты в зрелую мРНК путем вырезания интронов.) Рибосомальная РНК появляется в виде последовательно соединенного массива на конце хромосомы I. 5S РНК появляются в тандемном массиве у хромосомы V. Некоторые РНК-гены находятся в интронах генов, кодирующих белки.

Геном *C. elegans* содержит множество повторяющихся последовательностей. Приблизительно 2,6% генома состоит из тандемных повторов. Почти 3,6% генома содержит инвертированные повторы; они расположены преимущественно внутри экзонов, реже между генами. Повторы гексамера TTAGGC появляются многократно. Присутствуют также некоторые простые дупликации, включающие от сотен до десятков тысяч тпо.

Геном *Drosophila melanogaster*

Drosofila melanogaster, плодовая муха, используется как объект для подробных исследований в генетике и в биологии развития. Генетическая последовательность дрозофиллы была получена в результате сотрудничества Celera Genomics и Berkeley Drosophila Genome Project и опубликована в 1999 г.

Хромосомы *D. melanogaster* представляют собой нуклеопротеиновые комплексы. Примерно третья часть генома содержится в гетерохроматине, сильно скрученных и компактных (и потому контрастно окрашиваемых) областях, расположенных по бокам центромеров. Другие две трети представлены эухроматином, менее раскрученной, мало компактной формой. Большая часть активных генов находится в эухроматине. Гетерохроматин *D. melanogaster* содержит множество тандемных повторов последовательности AATAACATAG и сравнительно мало генов.

Всего хромосомная ДНК содержит примерно ~180 млн пн. Последовательность эухроматиновой части, состоит из 120 млн пн.

Геном распределен между пятью хромосомами: три большие аутосомы, Y-хромосома и пятая маленькая хромосома, содержащая всего ~1 млн пн эухроматина. Количество генов у мухи — 13 601, что примерно вдвое больше, чем у дрожжей, но, что весьма удивительно, меньше, чем у *C. elegans*. Средняя плотность генов в эухроматиновой последовательности — 1 ген/9 тпн; что намного ниже типичной плотности у прокариот — 1 ген/1 тпн.

Несмотря на то что насекомых нельзя назвать близкими родственниками млекопитающих, геном мушки используется для исследования заболеваний человека. Он содержит гомологи 289 человеческих генов, имеющих отношение к различным болезням, таким как рак, сердечно-сосудистые, неврологические, эндокринные, почечные заболевания, заболевания кровеносной системы, нарушения обмена веществ. Некоторые из этих гомологов у человека и мухи выполняют разные функции. Другие гены, связанные с болезнями человека, могут быть встроены в геном мушки и исследованы на ней. К примеру, человеческий ген спинномозговой атаксии типа 3, экспрессированный в мушке, приводит к сходной дегенерации нервных клеток. Болезни Паркинсона и малярии уже моделируют на мушке.

Некодирующие области генома *D.melanogaster* должны содержать участки, контролирующие пространственные модели развития во времени (сегментацию эмбриона). Биология развития очень активно изучается на мушке. Таким образом, это организм, на котором изучение геномики развития показывает себя крайне информативным.

Геном *Arabidopsis thaliana*

Растение *Arabidopsis thaliana* является очень далеким родственником других высших эукариот, для которых получены геномные последовательности. Это делает его удобным объектом для сравнительного анализа с целью установления между ними общих и специфических особенностей.

Геном *Arabidopsis thaliana* содержит ~125 млн пн, из которых 115,4 млн пн были объявлены прочитанными в 2000 г. в рамках проекта «Геном

Arabidopsis». Имеется 5 пар хромосом, содержащих 25 498 предполагаемых генов. Геном сравнительно компактен, содержит в среднем 1 ген на 4,6 тпн. Эта цифра — что-то среднее между прокариотами и *Drosophila*, и, при грубой оценке, она оказывается близкой к аналогичной для *C. elegans*. Гены *Arabidopsis* относительно коротки. Типичный экзон состоит из 250 пн, а интроны сравнительно малы — 170 пн в среднем. Для генов растений характерно GC-обогащение кодирующих областей.

Геном *Arabidopsis thaliana*

	Хромосома					
	1	2	3	4	5	Сумма
Длина (пн)	29 105 111	19 646 945	23 172 617	17 549 867	25 353 409	115 409 949
Число генов	6 543	4 036	5 220	3 825	5 874	25 498
Плотность (тпн/ген)	4.0	4.9	4.5	4.6	4.4	
Средняя длина гена	2 078	1 949	1 925	2 138	1 974	

Большинство белков *Arabidopsis* имеет гомологов в белковых последователях животных, но некоторые уникальные системы, например, отвечающие за производство клеточной стенки или фотосинтез, встречаются только у растений. Многие белки, встречающиеся у животных, широко дивергировали со времен последнего общего предка. Еще одно типичное отличие между растениями и животными состоит в том, что у растений 25% клеточных ядер содержат сигнальные последовательности, управляющие их транспортом в органеллы — митохондрии и хлоропласты, в то время как у животных всего 5% генов управляют транспортом в митохондрии.

В геноме *Arabidopsis* встречаются следы крупномасштабных и небольших дупликаций. 58% генома содержат 24 дуплицированных сегмента длиной не менее 100 тпн. Гены в этих сегментах обычно *не* могут быть идентифицированы как гомологи. Тем не менее 4140 генов (17% генов) встречаются в 1528 дуплицированных группах генов, содержащих до 23 членов. На основании этого примера дупликации появилось предположение, что предок *Arabidopsis* претерпел удвоение всего генома, и образовался тетраплоидный вид. На то, что это произошло давно — по оценкам, 112 млн лет назад — указывают очень большие потери одной из копий и широкая дивергенция генов в оставшихся копиях.

Высшее растение — единое целое, состоящее из трех геномов — ядерного, митохондриального и хлоропластного. Геномы органелл значительно меньше ядерного.

Распределение генов *A. thaliana* между ядром и органеллами

	Ядро	Хлоропласт	Митохондрия
Размер (тпн)	125 100	154	367
Гены белков	25 498	79	58
Плотность (тпн/ген)	4.5	1.2	6.25

Многие гены, кодирующие белки, транспортируемые в органеллы, вероятно, были перенесены в ядро из органелл, где находились изначально.

Геном *Homo sapiens* (геном человека)

Замечание

«Лица, пытающиеся определить мотивы этого произведения, будут преследоваться по закону; лица, пытающиеся отыскать здесь мораль, будут выпровожены; лица, пытающиеся увидеть здесь заговор, будут застрелены.»

Марк Твен
(Приключения Геккельберри Финна. Предисловие)

В феврале 2001 г. Международный консорциум по секвенированию генома человека и Celera Genomics независимо опубликовали последовательность генома человека. Эта последовательность длиной ~3.2 млрд пн в тридцать раз больше геномов *C. elegans* и *D. melanogaster*. Одна из причин такой несоразмерности состоит в том, что кодирующие последовательности составляют лишь 5% человеческого генома; повторы составляют более 50% генома. Пожалуй, наиболее удивительным фактом оказалось малое число идентифицированных генов. Обнаружение лишь порядка 30 000–40 000 генов говорит о том, что альтернативный сплайсинг вносит значительный вклад в разнообразие наших белков. По оценкам, ~35% генов подвержены альтернативному сплайсингу. (По последним оценкам геном человека содержит около 25 тыс. генов. — *Прим. ред.*)

Геном человека распределен по 22 парам хромосом, плюс X и Y хромосомы. Длина ДНК, содержащейся в аутосомах, колеблется от 279 млн пн до 48 млн пн. X-хромосома содержит 163 млн пн, а Y-хромосома — только 51 млн пн.

Экзоны человека из белок-кодирующих генов малы по сравнению с экзонами в других известных эукариотических геномах. Интроны относительно длинные. В результате многие гены, кодирующие белки, имеют большую длину. К примеру, ген дистрофина, кодирующий белок длиной 3685 аминокислот, имеет длину более 2,4 млн пар оснований.

Белок-кодирующие гены

Анализ набора белков в геномной последовательности человека сопровождается трудностями из-за проблем с надежным обнаружением генов и альтер-

нативными путями сплайсинга. Международным консорциумом установлено, что общее число генов в геноме человека — около 32 тыс.

Основными функциональными группами являются:

Функции	Число	%
Связывающие нуклеиновые кислоты	2207	14.0
ДНК-связывающие	1656	10.5
Белки репарации ДНК	45	0.2
Факторы репликации ДНК	7	0.0
Факторы транскрипции	986	6.2
Связывающие РНК	380	2.4
Структурные белки рибосом	137	0.8
Факторы трансляции	44	0.2
Связывающие факторы транскрипции	6	0.0
Регуляторы клеточного цикла	75	0.4
Шапероны	154	0.9
Двигательные	85	0.5
Связывающие актин	129	0.8
Защитные (иммунные) белки	603	3.8
Ферменты	3242	20.6
Пептидазы	457	2.9
Эндопептидазы	403	2.5
Протеинкиназы	839	5.3
Протеинфосфатазы	295	1.8
Активаторы ферментов	3	0.0
Ингибиторы ферментов	132	0.8
Ингибиторы апоптоза	28	0.1
Сигнальные	1790	11.4
Рецепторы	1318	8.4
Трансмембранные рецепторы	1202	7.6
Рецепторы, связанные с G-белком	489	3.1
Обонятельные рецепторы	71	0.4
Запасные белки	7	0.0
Белки адгезии	189	1.2
Структурные белки	714	4.5
Структурные белки цитоскелета	145	0.9
Переносчики	682	4.3
Ионный канал	269	1.7
Переносчики нейромедиаторов	19	0.1
Связывающие или переносящие лиганды	1536	9.7
Переносчики электронов	33	0.2
Цитохром P450	50	0.3
Белки, подавляющие опухолеобразование	5	0.0
Неклассифицированные	4813	30.6
Всего	15683	100.0

Источник: `http://www.ebi.ac.uk/proteome/` [under Functional classification of *H. sapiens* using Gene Ontology (GO): General statistics (InterPro proteins with GO hits)].

Структурная классификация по наиболее часто встречающемуся типу:

Наиболее часто встречающиеся типы белковых доменов

Белок	Число
Иммуноглобулины и домены главного комплекса гистосовместимости	591
Цинковый палец, тип C2H2	499
Протеинкиназы эукариот	459
Надсемейство родопсинподобных GPCR	346
Активный сайт семейства серин/треонин-протеинкиназ	285
EGF-подобный домен	259
РНК-связывающий участок RNP-1 (мотив узнавания РНК)	214
WD-40 бета-повторы G-белков	196
Домен гомологичный Src 3 (SH3)	194
Плекстрин-подобный домен (PH)	188
Семейство EF-hand	185
Гомеобокс-домен	179
Каталитический домен тирозин-киназы	173
Иммуноглобулин типа V	163
RING-палец	159
Экстенсин, обогащенный пролином	156
Домен III фибронектина	151
Анкириновые повторы	135
Бокс KRAB	133
Подтип иммуноглобулина	128
Кадхериновый домен	118
Домен PDZ (или DHR, или GLGF)	117
Повтор доменов, обогащенных лейцином	113
Серинпротеазы, семейство трипсинов	108
Надсемейство Ras GTP	103
Домен, гомологичный Src 2 (SH2)	100
Домен BTB/POZ	99
TPR-повторы	92
Надсемейство AAA АТФаз	92
Сайт гидроксилирования аспарагина и аспарагиновой кислоты	91

Источник: `http://www.ebi.ac.uk/proteome/`

Повторяющиеся последовательности

Повторяющиеся последовательности занимают более 50% генома:

- Перемещаемые элементы или рассеянные повторы — почти половина всего генома! Они включают: повторы типа LINE и SINE.
- Псевдогены ретропозиции
- Простые разбросанные повторения коротких олигомеров типа мини- и микросателлитов. Трехнуклеотидные повторы, такие как CAG, соответствующие глутаминовым повторам в белке, вовлечены во множество болезней.

Типы мобильных элементов в человеческом геноме

Элемент	Размер (пн)	Количество копий	Доля в геноме (%)
Короткие рассеянные ядерные элементы (SINE)	100–300	1 500 000	13
Длинные рассеянные ядерные элементы (LINE)	6 000–8 000	850 000	21
Длинные терминальные повторы	15 000–110 000	450 000	8
Остатки транспозонов ДНК	80–3 000	300 000	3

- Сегментарные дупликации блоков размером приблизительно 10–300 тпн. Межхромосомные дупликации, появляющиеся в негомологичных хромосомах, иногда во множестве сайтов. Некоторые внутрихромосомные дупликации, включающие в себя расположенные близко повторяющиеся участки длиной много тысяч пар оснований очень схожей структуры, вовлечены в генетические заболевания. Например, синдром Шарко—Мари типа 1А, прогрессирующая периферийная нейропатология, возникающая в результате дупликации участка, содержащего ген периферийного миелинового белка 22.
- Блоки тандемных повторов, включающие семейства генов.

РНК

Гены РНК в человеческом геноме включают[1]:

1. 497генов тРНК. Один особенно большой кластер содержит 140 генов тРНК внутри участка на 6-й хромосоме длиной 4 млн пн.
2. Гены 28S и 5.8S рибосомной РНК имеются в 150–200 копиях участка длиной 44 тпн Гены 5S рибосомной РНК также появляются в виде тандемных последовательностей содержащих 200–300 генов, наибольшая из которых находится на 1-й хромосоме.
3. Малые ядерные РНК (мяРНК, snRNA) включают 2 семейства молекул, которые разрезают и обрабатывают рРНК.
4. Сплайсосомные мяРНК включают U1, U2, U4, U5, U6 мяРНК, многие из которых появляются в кластерах тандемных повторов с практически одинаковыми или инвертированными последовательностями.

[1]К этому списку следует добавить по крайней мере несколько сотен микро-РНК — малых РНК, принимающих участие в регуляции экспрессии генов. В настоящее время этот тип объектов привлекает особое внимание исследователей. — *Прим. ред.*

WEB-РЕСУРСЫ: ИНФОРМАЦИЯ О ГЕНОМЕ ЧЕЛОВЕКА

Интерактивный доступ к аминокислотным и нуклеотидным последовательностям:
`http://www.ensembl.org/`

Изображения хромосом, карты, локусы:
`http://www.ncbi.nlm.nih.gov/genome/guide/`

Генетическая карта 99:
`http://www.ncbi.nlm.nih.gov/genemap99/`

Обзор генетической структуры человека:
`http://hgrep.ims.u-tokyo.ac.jp`

Банк СНП:
`http://snp.cshl.org/`

Генетические заболевания человека:
`http://www.ncbi.nlm.nih.gov/Omim/`
`http://www.geneclinics.org/profiles/all.html`

Этические, правовые, социальные вопросы:
`http://www.nhgri.nih.gov/ELSI/`

Однонуклеотидные полиморфизмы (SNP, СНП)[1)]

Однонуклеотидный полиморфизм (или СНП) — это генетическая изменчивость между особями. Условия возникновения СНП ограничены одной начальной парой, в которой может возникнуть замена, вставка или делеция. Злокачественная анемия — пример заболевания, вызванного специфичным СНП: замена A→T в β-глобиновом гене вызывает замену Glu→Val, делая поверхность гемоглобина способной к «слипанию», что ведет к полимеризации бескислородной дезокси-формы.

СНП, рассредоточенные по всему геному, встречаются в среднем один раз на 2000 пар. (По современным данным, они встречаются в среднем чаще, чем 1 раз на 500 нуклеотидов. — *Прим. ред.*) Несмотря на то что они вызываются мутациями, многие позиции, содержащие СНП, имеют низкие уровни мутации, и могут быть использованы в качестве стабильных маркеров для картирования генов. Консорциум из четырех центров академических исследований по геному

[1)] В русской литературе правильнее было бы использовать аббревиатуру ОНП, однако все специалисты (в том числе и в России) используют термин СНП (в разговоре «снип»)

и 11 частных компаний создает высококачественную карту человеческих СНП с высокой плотностью размещения и с возможностью общего доступа. Последняя опубликованная версия содержит 1.42 млн СНП.

Не все СНП связаны с заболеваниями. Многие из них встречаются внутри нефункциональных областей (хотя плотность СНП внутри областей, содержащих гены, выше средней[1]). Некоторые СНП, которые происходят внутри экзонов, вызывают замену на синонимичный кодон или замену, незначительно влияющую на функциональность белка. Другие типы СНП могут вызвать изменения в белке более серьезного уровня, чем локальные: (1) замена значащего кодона на стоп-кодон, что приводит к преждевременной остановке синтеза белка; (2) делеция или вставка приводит к сдвигу рамки считывания.

A, B и O аллели, определяющие группу крови, обусловлены СНП-заменами. Они кодируют родственные белки, которые присоединяют различные сахаридные остатки к антигену на поверхности эритроцитов.

Аллель	Последовательность	Сахарид
A	`\dots{}gctggtgacccctt\dots{}`	N-ацетилгалактозамин
B	`\dots{}gctcgtcaccgcta\dots{}`	Галактоза
O	`\dots{}cgtggt-acccctt\dots{}`	—

Аллель O прошла мутацию, вызвавшую сдвиг рамки считывания, и не производит активного фермента. Эритроциты группы 0 не содержат ни A-, ни B-антиген. Поэтому люди с группой крови 0 являются универсальными донорами при переливании крови. Потеря активности белка, по всей видимости, не влечет за собой каких-либо неблагоприятных последствий. Действительно, люди с группой крови 0 обладают повышенной устойчивостью к оспе.

Сильная зависимость между заболеванием и специфичными СНП может быть использована в лечебной практике, так как это сравнительно легко обнаружить у больных. Но если заболевание происходит из-за дисфункции специфичного белка, то может быть много сайтов мутации, которые могут вызвать эту инактивацию. Конкретный сайт может доминировать, если (1) все носители гена являются потомками одной особи, у которой произошла мутация, и/или (2) , если заболевание возникает скорее в результате получения, чем в результате потери какой-либо специфической особенности, такой как способность гемоглобина серповидных клеток полимеризоваться, и/или (3) , если частота мутации в каком-то сайте излишне высока (пример — Glu380→Arg мутация в гене роста фибробласта FGFR3), что ассоциируется с ахондроплазией; одно из проявлений — низкий рост).

[1] Это скорее связано с тем, что области, содержащие гены, более детально исследованы, и поэтому в них *обнаружено* больше СНП, а протяженные межгенные области менее изучены и поэтому в них *видели* меньше СНП. Однако это не означает, что там их меньше на *самом деле*. Это является хорошим примером того, что банки данных зачастую не дают *представительной выборки*, и поэтому далеко не всегда на основе массового анализа данных можно делать правильные статистические выводы. — *Прим. ред.*

Напротив, многие независимые мутации были определены в генах BRCA1 и BRCA2, в прилегающих локусах, связанных с повышенной расположенностью к раннему появлению рака груди и яичников. В то же время продукт нормального гена подавляет опухоль. Мутанты, являющиеся результатом вставки или делеции, вызывающие сдвиг рамки считывания, обычно не производят белок или производят неактивный белок. Но эта зависимость не может быть выведена априори вне зависимости от того, повлияет ли возникновение замены в генах BRCA1 или BRCA2 на степень риска.

Лечение заболеваний, вызванных дефектными белками или отсутствием белка, включает:

1. *Обеспечение организма нормальным белком.* Мы упоминали инсулин при лечении диабета и фактор VIII при гемофилии. Другой пример — управление человеческим гормоном роста у пациентов, страдающих от его полного или частичного отсутствия. Использование рекомбинантных белков снижает риск передачи СПИДа через переливание крови или развития болезни Кретцфельда—Якоба, возникающей при изоляции гормона роста от рецепторов гипофиза.
2. *Изменение образа жизни, что делает данную функцию не обязательной.* Фенилкетонурия (PKU) — генетическое заболевание, вызванное дефицитом фенилаланингидроксилазы — фермента, катализирующего превращение фенилаланина в тирозин. Накопление большого количества фенилаланина вызывает проблемы индивидуального развития, включающие задержку умственного развития. Избежать симптомы можно диетой, бедной фенилаланином. В США и многих других странах юридически разрешена проверка новорожденных на уровень фенилаланина в крови.
3. Генная *терапия* для возмещения белка при его отсутствии в стадии разработки.

Другие методы использования СНП в медицине отражают связь между генотипом и реакцией на терапию (фармакогеномика). Например, СНП в гене N-ацетилтрансферазы NAT-2 связан с периферийной невропатией — слабость, нечувствительность и боль в ладонях, руках, ногах и ступнях как побочный эффект лечения изониазидом (гидразид изоникотиновой кислоты) при туберкулезе. Пациентам с этим СНП прописывают альтернативное лечение.

Генетическое разнообразие в антропологии

Данные СНП имеют важное применение в антропологии, давая ключ к пониманию исторических изменений размеров популяций, а также моделей миграций.

Степень генетического разнообразия представляют в терминах численности популяции. Основателем называют исходный набор особей, от которых и происходит вся популяция. Это могут быть либо первоначальные колонисты, как, например, полинезийцы, которые впервые поселились в Новой Зеландии,

либо просто выжившие особи в популяции, которая находилась на грани вымирания. Гепарды сейчас являются примером популяции, которая (по оценкам) 10 000 лет назад была в условиях резкой нехватки ресурсов. Все ныне живущие гепарды настолько же близки родственно, как брат и сестра.

Экстраполяция митохондриальной ДНК у различных людей одного возраста и ее показанная изменчивость предполагает существование некоторой матери-предка, которая жила 140 000–200 000 лет назад. Называя ее Евой, мы предполагаем, что она была первой женщиной. Но исследования окаменелостей доказывают, что люди существовали намного раньше. Митохондриальная Ева была всего лишь основательницей выжившей ветви популяции, которая находилась на грани вымирания.

Специфические для каждой популяции данные CHII дают информацию о миграциях. Митохондриальные последовательности при этом представляют информацию о предках женского пола, а последовательности из Y-хромосомы — о предках мужского. Например, было предположено, что население Исландии, впервые заселенной более чем 1100 лет назад, происходит от скандинавских мужчин и женщин как из Скандинавии, так и с Британских островов. В свою очередь, средневековые исландские летописи указывают на вражеские набеги на поселения, расположенные на Британских островах.

Просто потрясающее сходство между последовательностями ДНК людей и их языковыми семьями было открыто Л. Л. Кавали-Сфорза и его коллегами. Эти исследования оказались полезными при установлении связи между языками северо-американских индейцев. Они указали на то, что баски, известные как лингвистически изолированная популяция, были, кроме того, генетически изолированы.

Во время исследований изолированных популяций антропологическая генетика предоставила ценные данные для медицины, ибо картирование генов, отвечающих за болезни, гораздо легче, если фоновые изменения генома незначительны. В генетически изолированные популяции в Европе включают, помимо басков, также финнов, исландцев, валлийцев (жителей Уэлса) и лопарей. В Исландии имеются хорошие генеалогические и медицинские записи. Правительство Исландии в 1998 г. издало закон, разрешающий создание базы данных, содержащей медицинские записи, историю семей и генетические последовательности 275 000 жителей. Компания Decode Genetics намерена сотрудничать со всеми производителями лекарственных препаратов, применяя эти данные.

Генетическое разнообразие и идентификация личности

Вариация наших последовательностей ДНК дала нам аналог индивидуальных отпечатков пальцев, что очень полезно для идентификации и для определения родства, в частности (но не ограничиваясь этим) для решения вопросов отцовства. Использование анализа ДНК как улики в расследовании криминальных дел сейчас прочно укоренилось.

Техника анализа «генетических отпечатков пальцев» изначально базировалось на примере VNTR, но сейчас настолько распространила свое влияние,

что допускает анализ других признаков, включая последовательность митохондриальной ДНК .

У большинства людей митохондрии генетически идентичны (в пределах одного организма. — *Прим. ред.*); это явление называется гомоплазией. У некоторых индивидуумов митохондрии могут содержать разные последовательности ДНК; это гетероплазия. Такое несоответствие в последовательности гена может усложнить уже изученное модельное наследование болезни.

Самый известный человек с гетероплазией — русский царь Николай II. После революции 1917 г. царь и его семья были вывезены в ссылку в Екатеринбург в центральной России. В ночь с 16 на 17 июля 1918 г. царь, царица Александра, по меньшей мере трое из пятерых их детей, врач и трое слуг, которые сопровождали семью, были убиты, а их тела захоронены в секретной могиле. Когда останки были обнаружены, исследование костей и зубов позволило предположить, а анализ последовательностей ДНК подтвердил, что это останки царской семьи. Подлинность останков царицы была доказана сопоставлением последовательности ее митохондриальной ДНК и ДНК родственника по материнской линии принца Филиппа, канцлера Кэмбриджского университета, графа Эдинбургского — внучатого племянника царицы.

Однако сравнение последовательностей михотондриальной ДНК предполагаемых останков царя с последовательностями ДНК от двух родственников по материнской линии показало расхождение в основании с номером 16 169: у царя в этой позиции был цитозин (С), а у родственников — тимидин (Т). Некоторые люди утверждали, что никаких сомнений допускать нельзя. Последующие анализы показали, что у царя была гетероплазия: Т-минорный компонент в положении 16 169. Чтобы подтвердить идентичность и не оставить никаких вопросов, тело великого князя Георгия, брата царя, было эксгумировано, и у него была показана та же редкая гетероплазия.

Генетический анализ одомашнивания крупного рогатого скота

Животноводство — неотъемлемая часть человеческой культуры. Анализ последовательностей ДНК проливает свет на историю животноводства и генетическое разнообразие разводимых в настоящее время животных.

Современный домашний крупный рогатый скот включает в Западной Европе и Северной Америке вид *Bos Taurus* — корову, а в Африке и Индии — зебу *Bos indicus*. Внешнее очевидное различие этих видов — это горб у зебу. Ранее было широко распространено мнение, что одомашнивание их произошло один раз и чуть более чем 8–10 тыс. лет назад, а затем эти два вида последовательно разошлись (произошла дивергенция).

Анализ последовательностей митохондриальных ДНК европейских, африканских и азиатских коров показал, однако, что (1) все европейские и африканские породы в большей степени родственны друг с другом, чем с индийскими породами, и (2) эти две группы разошлись более чем 200 тыс. лет назад, т. е. предположительно два разных вида были недавно и независимо

одомашнены. Сходство в физическом облике африканских и индийских зебу (а также другие похожие признаки на молекулярном уровне; например, VNTR-маркеры в ядерной ДНК) может быть связано с импортом крупного рогатого скота из Индии в восточную Африку.

Эволюция геномов

Доступность полной информации о геномных последовательностях переориентировало направление научных исследований. Основным направлением при анализе геномов стало распознавание «интересных событий». Фоновая скорость мутаций в кодирующих последовательностях отражена в синонимичных заменах нуклеотидов — изменения в кодонах, которые не меняют кодируемую аминокислоту. Приняв это как данность, можно искать случаи, когда намного больше скорость несинонимичных замен нуклеотидов — изменения кодонов, которые приводят к мутациям в кодируемом белке (следует, однако, отметить, что синонимичные замены не обязательно селективно нейтральны).

Имея две выровненные последовательности генов, мы можем рассчитать K_a = число несинонимичных замен и K_s = число синонимичных замен (расчет не столь прост, потому что нужно оценивать возможные множественные замены одного нуклеотида и внести соответствующую поправку). Высокое значение отношения K_a/K_s свидетельствует о позитивном отборе, что связано, возможно, с изменением функции.

Сравнительная геномика поднимает совершенно новые вопросы:

- Какие *гены* различают фенотип особей? Какие гены уникальны для каждого конкретного фенотипа особи? Варьирует ли их положение в геноме от особи к особи?
- Какие гомологичные *белки* различают фенотип особей? Какие белки уникальны для каждого конкретного фенотипа особи? Изменяется ли суммарное действие этих белков от особи к особи? Изменяются ли механизмы контроля нормальной экспрессии этих белков?
- Какие *биохимические функции* различают фенотип особей? Какие биохимические функции уникальны для каждого конкретного фенотипа особи? Изменяется ли суммарное действие этих биохимических функций от особи к особи? И если две особи имеют некоторую функцию и в одной есть ответственный за нее белок, а в другой — его гомолог, то несет ли этот белок ту же функции во втором организме?

Подобные вопросы правомочны и для различных видов внутри каждого таксона.

Наше внимание, в первую очередь, привлекает геном человека. Но способность взаимодействия различных геномов крайне важна.

М. А. Андрадэ, С. Озунис, С. Сандер, Дж. Тамамес и А. Валенсия сравнили набор белков из различных видов трех основных доменов: *Haemophilus influenzae* — представителя бактерий (*Bacteria*), *Methanococcus jannaschii* — архей (*Archaea*), *Saccharomyces cerevisiae* (дрожжи) — эукариот (*Eukarya*). Они

предложили классификацию белков, включающую в себя следующие основные категории: энергетика клетки, информация, коммуникации и регуляции.

Главные функциональные классы белков:

- Энергетика клетки
  - Биосинтез кофакторов, аминокислот
  - Центральный и опосредованный метаболизм
  - Энергетический обмен
  - Жирные кислоты и фосфолипиды
  - Биосинтез нуклеотидов
  - Транспорт
- Информация
  - Репликация
  - Транскрипция
  - Трансляция
- Коммуникация и регуляция
  - Регуляторные функции
  - Клеточная оболочка/клеточная стенка
  - Клеточные процессы

Число генов в трех видах:

Вид	Число генов
Haemophilus influenzae	1680
Methanococcus jannaschii	1735
Saccharomyces cerevisiae	6278

Есть ли общие белки для общих функций? В классе белков клетки, выполняющие энергетические функции, распределены по трем классам. В классе коммуникации белки уникальны в каждом домене. В классе регуляции и информации археи (бактерии) имеют некоторые общие белки с бактериями, другие — с эукариотами.

Анализ общих функций среди всех доменов жизни привели к вопросу, возможно ли определить минимальный организм, т. е. организм с минимальным геном, согласующимся с основой жизни — центральной догмой ДНК → РНК → белок (т. е. догмой, отрицающей существование безбелковых форм жизни, основанных исключительно на РНК). Минимальный организм должен обладать способностью к воспроизведению, но не обязательно должен конкурировать в способности роста и размножения с другими организмами. Можно предположить, что минимальный организм должен усваивать питательную среду, обеспечивая организму биосинтез, а также обеспечивать ответ на стресс, в том числе и восстановление поврежденной ДНК.

Самый малый самостоятельный организм *Mycoplasma genitalium* имеет 468 предсказанных кодирующих последовательностей. В 1996 г. А. Р. Мушегян и Е. В. Кунин сравнили геномы *M. genitalium* и *H. influenzae*. (В то время это были единственными полностью секвенированные бактериальные геномы.) Последний общий предок далеко разошедшихся бактерий жил 2 млрд лет назад. Из 1703 белок-кодирующих генов в *H. influenzae* 240 гомологичны белкам в *M. genitalium*. Мушегян и Кунин сделали вывод, что все они должны быть незаменимыми, но могут быть недостаточными для самостоятельной жизни: некоторые важнейшие функции могут осуществляться неродственными белками из двух организмов. Например, этот набор из 240 белков не обеспечивает всех важнейших метаболических путей, для которых привлекаются дополнительно 22 фермента из *M. genitalium*. Наконец, идентификация функционального избытка и генов, специфичных для паразитов, дает список из 256 генов как необходимый и достаточный минимальный набор.

Из чего состоит предложенный минимальный геном? Функциональные классы включают:

- Трансляция, в том числе синтез белка
- Репликация ДНК
- Рекомбинация и репарация — вторая функция основных белков, вовлеченных в репликацию ДНК
- Аппарат транскрипции
- Белки-шапероны
- Промежуточный метаболизм — гликолитический путь
- Нет систем биосинтеза нуклеотидов, аминокислот и жирных кислот
- Структуры экспорта белка
- Ограниченный набор белков, участвующих в транспорте метаболитов[1)]

Подчеркивается, что жизнеспособность организма с этими белками не доказана. Кроме того, даже если эксперименты докажут, что какой-то минимальный набор генов (предсказанный или некоторый другой) — необходимый и достаточный, то это никак не связано с вопросом об идентификации дополнительных генов общего предка *M. genitalium* и *H. influenzae*, или более ранних клеточных форм жизни. Только для 71% из предложенного набора из 256 белков распознаны гомологии среди белков эукариот и архей.

Тем не менее определение функций, несомненно, общий подход при изучении всех форм жизни, что позволяет нам исследовать экологические ниши, в которых каждая жизненная форма сформировала эти функции аналогичными путями. Катализируют ли гомологичные белки из различных видов похожие реакции? Анализы геномов выявили семейства белков с гомологией у архей, бактерий и эукариот. Допущение состоит в том, что семейства белков развились из индивидуального гена-предшественника в ходе видообразования и повторения событий, хотя это могло быть эффектом горизонтального переноса. Задача — составить карту общих функций и общих белков.

[1)] Следует отметить, что, скажем, отсутствие биосинтеза какого-либо важного метаболита, неизбежно требует наличия соответствующего транспортера. — *Прим. ред.*

Несколько тысяч белковых семейств установлены путем анализа гомологии с археями, бактериями и эукариотами. Разные виды содержат различное количество этих общих семейств: бактерии *Aquifex aeolicus* — 83% белков имеют гомологов среди архей и эукариот, а *Borrelia burgdorferi* — только 52% белков имеют гомологов среди архей и эукариот. Геномы архей имеют немного больший процент белков (62–71%), гомологичных белкам бактерий и эукариот. Но только 35% белков дрожжей гомологичны белкам бактерий и архей.

Правильно ли, что общий набор белков выполняет общий набор функций? Среди белков минимального набора установленного из *M. genitalium* только около 30% имеют гомологии во всех известных геномах. Другие важные функции должны осуществляться при помощи неродственных белков или в некоторых случаях с применением неузнанных гомологов. Семейства белков, в которых гомологи осуществляют общие функции в археях, бактериях и эукариотах, вовлечены в трансляцию:

Функциональные классы белков	Количество семейств, появляющихся во всех известных геномах
Трансляция, включая структуру рибосом	53
Транскрипция	4
Репликация, рекомбинация, репарация	5
Базовый метаболизм	9
Клеточные процессы (шапероны, секреция, клеточное деление, формирование клеточной стенки)	9

Мы видим, что эволюция открыла очень широкие возможности белков к приспособлению с целью выполнения определенных функций. Это должна быть наиболее консервативная часть из множества синтезируемых белков.

WEB-РЕСУРСЫ: WEB-РЕСУРСЫ: БАЗЫ ДАННЫХ С ВЫРОВНЕННЫМИ СЕМЕЙСТВАМИ ГЕНОВ

Pfam: Protein families database:
`http://www.sanger.ac.uk/Software/Pfam/`

COG: Clusters of orthologous groups:
`http://www.ncbi.nlm.nih.gov/COG/`

HOBACGEN: Homologous Bacterial Genes Database:
`http://pbil.univ-lyon1.fr/databases/hobacgen.html`

HOVERGEN: Homologous Vertebrate Genes Database:
`http://pbil.univ-lyon1.fr/databases/hovergen.html`

TAED: The Adaptive Evolution Database:
`http://www.sbc.su.se/$\sim$liberles/TAED.html`

Многие другие сайты содержат данные об индивидуальных семействах генов.

Пожалуйста, передайте гены: горизонтальный перенос генов

Узнав о том, что трипсин в организме *S. griseus* более близок трипсину из организма быка, чем к другим протеиназам микроорганизмов, Брайан Хартли сказал в 1970 г., что «...бактерия, должно быть, заражена коровой». Это был пример межвидового, или горизонтального, переноса генов — бактерия, «берущая» ген из субстрата, на котором она росла и который образовался в результате деятельности организма другого вида. Более ранними примерами были классические эксперименты по трансформации пневмококков, проводившиеся О. Эвери, К. Маклеодом и М. Маккарти, где было определено, что ДНК является генетическим материалом. В общем случае горизонтальный перенос генов — это приобретение генетического материала одним организмом от другого (обычно естественным путем, но иногда и в результате лабораторных процедур) способами, отличными от репликации или дупликации. Известно несколько механизмов горизонтального переноса генов, таких как прямое поглощение (как в экспериментах по трансформации пневмококков) или через вирусные переносчики.

Анализ генных последовательностей показал, что горизонтальный перенос не такое редкое явление, хотя встречается оно в основном в генах микроорганизмов. Оно требует от нашего мышления некоторого отхода от обычных «клоновых», или родительских, моделей наследственности. Факты, подтверждающие горизонтальный перенос, включают: во-первых, различия среди эволюционных деревьев, составленных из разных генов, и, во-вторых, прямые сравнения генных последовательностей у разных видов:

- В *E. coli* 755 открытых рамок считывания (всего 547,8 тпн, $\sim$ 18% генома) появились благодаря горизонтальному переносу после дивергенции от потомков организма *Salmonella* 100 млн лет назад.
- В эволюции микроорганизмов горизонтальный перенос более распространен среди «рабочих» генов (которые ответственны за «домашнее хозяйство», например биосинтез), чем среди информационных генов (которые ответственны за организационные процессы, такие как транскрипция и трансляция). Например, *Bradyrhizobium japonicum*, азотфиксирующая бактерия, состоящая в симбиозе с высшими растениями, имеет два гена глутамин-синтетазы. Один — такой же, как и у других родственных бактерий; другой — на 50% совпадает с аналогичным геном высших растений. Рубиско (рибулозо-1,5-дифосфаткарбоксилаза/оксигеназа), фермент, который первым фиксирует диоксид углерода в цикле Кальвина (один из этапов темновой фазы фотосинтеза), встречается в бактериях, митохондриях и пластидах водорослей, является следствием генной дупликации. Многие гены из бактериофагов, появившиеся в геноме *E. coli*, демонстрируют нам дальнейшие примеры и указывают на механизм переноса.

Однако феномен горизонтального переноса известен не только у прокариот, но и у эукариот. Последние получили свои информационные гены в основном из организма, родственного *Methanococcus*, а «рабочие» гены — в основном из протеобактерий с небольшим «вкладом» цианобактерий и метанпродуцирующих бактерий. Почти все информационные гены из *Methanococcus* такие же,

как аналогичные гены у дрожжей. Однако горизонтальный перенос наблюдался не только у далеких предков. Было высказано предположение (которое сейчас оспаривается), что многие бактериальные гены вошли в геном человека, а по меньшей мере восемь человеческих генов появилось в геноме организма *M. tuberculosis*.

Из полученных данных напрашивается вывод, что существует модель «всеобщего организма», общего генетического «рынка», и даже «Всемирной ДНК-паутины», из которой организмы загружают гены по собственному желанию! Как это может согласовываться с фактическим разнообразием видов, которое продолжает существовать? Традиционное объяснение заключается в том, что биосфера содержит экологические «ниши», к которым приспосабливаются отдельные виды. Из разнообразия ниш следует разнообразие видов. Но это объяснение опирается на устойчивость нормальной наследственности, необходимой для поддержания «здоровья» видов. Почему «всеобщий организм» не разрушает границы между видами подобно тому, как глобальный доступ к поп-культуре угрожает разрушить границы между национальными и этническими культурами? Возможный ответ — благодаря информационным генам, которые изначально менее подвержены горизонтальному переносу, определяющему сходство видов.

Интересно, что хотя важность горизонтального переноса генов ясна и очевидна, данный факт был забыт на долгое время как редкий и не играющий особой роли. Причина этого «умственного дискомфорта» лежит на поверхности: перенос генов от родителей к потомкам лежит в основе дарвинской модели биологической эволюции посредством селекции (дифференцированного воспроизведения) родительских фенотипов с изменением частоты встречаемости генов в следующем поколении. Приобретение потомками генов от кого-то еще, кроме родителей, ассоциировалось с моделью Ламарка и другими не очень принятыми альтернативами дарвинской модели. Мнение, что эволюционное дерево является организующим принципом биологического родства, глубоко укоренилось в сознании ученых: они проявляют «экологическое» рвение в своей приверженности к деревьям даже тогда, когда деревья не являются подходящей моделью для выявления родства. Возможно, было бы хорошо напомнить, что Дарвин ничего не знал о генах; и механизм, который создает разнообразие признаков в ходе селекции, был для него загадкой. Может быть, он бы легче согласился с горизонтальным переносом генов, чем его последователи!

Сравнительная геномика эукариот

Сравнение геномов дрожжей, плодовой мушки, червя и человека выявило 1308 групп белков, которые присутствуют во всех организмах. Они формируют консервативную часть белков, которые отвечают за основные функции: метаболизм, репликация и репарация ДНК, трансляция.

Эти белки состоят из отдельных доменов, включающих в себя однодоменные белки, олигомерные белки и модульные белки с множеством доменов (самый большой из них, мышечный белок титин, содержит 250–300 доменов).

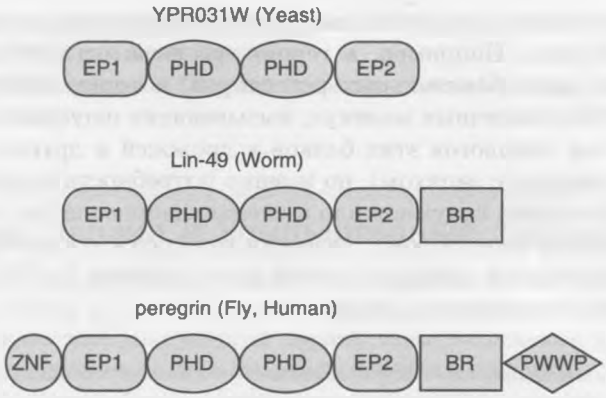


Рис. 2.4. Эволюция путем приращения доменов, а именно молекул, родственных перегрину (человеческому белку, который, возможно, отвечает за регуляцию транскрипции). Гомолог *C. elegans*, Lin-49, важен для нормального развития червя. Функция гомолога дрожжей неизвестна. Белки содержат следующие домены: ZNF = C_2H_2 — цинковый домен (не путать с ацетиленом; C — цистеин, H — гистидин); EP1 и EP2 — ретрансляторы поликомба 1 и 2, PHD — растительный гомеодомен (репрессорный домен, содержащий $C_4H_3C_3$), BR — бромовый домен; PWWP — домен, содержащий последовательность Pro–Trp–Trp–Pro

Белки плодовой мушки и червя построены из структур, в которых в три раза больше доменов, чем в белках дрожжей. Человеческие белки содержат, в свою очередь, в два раза больше доменов, чем белки мухи и червя. Большинство этих доменов присутствуют также в бактериях и археях, но некоторые специфичны только для позвоночных (возможно, найдены только в них) (см. таблицу ниже). Они включают в себя белки, которые ответственны за процессы, наблюдаемые у позвоночных (например, защитные и иммунные белки, а также белки в нервной системе); один из них — фермент рибонуклеаза, по-видимому, не делающий ничего специфически-позвоночного.

Распределение вероятных гомологов предсказанных белков человека

Только позвоночные	22%
Позвоночные и другие животные	24%
Животные и другие эукариоты	32%
Эукариоты и прокариоты	21%
Нет гомологов у животных	1%
Только прокариоты	1%

Для создания новых белков не часто происходит изобретение новых доменов. Обычно создаются все более сложные комбинации уже существующих доменов. Общий механизм состоит в приросте доменов на концах исходных белков (см. рис. 2.4). Этот процесс может происходить независимо и разнонаправленно в разных ветвях живой природы.

Дупликация генов, за которой следует их дивергенция, — механизм создания семейств белков. Например, в геноме человека есть 906 генов (вместе с псевдогенами) для обонятельных рецепторов, которые могут связываться с примерно 10 000 различных молекул, вызывающих ощущение запаха. Было показано наличие гомологов этих белков у дрожжей и других грибов (некоторые из них связаны с запахом), но именно потребность позвоночных в высокоразвитом обонянии преумножило и специализировало это семейство. 80% генов обонятельных рецепторов у человека находятся в кластерах. Сравните с этим очень маленький размер кластера генов глобина (с. 107), которому не требуется такое огромное разнообразие.

Литература

C. elegans sequencing consortium (1999) 'How the worm was won. The *C. elegans* genome sequencing project', *Trends in Genetics* 15, 51–58. [Описание проекта, в котором было изобретено высокоскоростное секвенирование ДНК, и получен результат — первый секвенированный геном многоклеточного организма.]

Doolittle, W. F. (1999) 'Lateral genomics', *Trends in Cell Biology* 9, M5–8. [Как горизонтальный перенос генов переворачивает традиционные взгляды эволюции.]

Fitch, W. M. (2000). Homology: A personal view of some of the problems. *Trends in Genetics* 16, 227–31. [Вдумчивое исследование понятия гомологии.]

Koonin, E. V. (2000) 'How many genes can make a cell: the minimal-gene-set concept', *Annual Review of Genomics and Human Genetics* 1, 99–116. [Краткое изложение работы по сравнительной геномике.]

Kwok, P.-Y. and Gu, Z., (1999) 'SNP libraries: why and how are we building them?', *Molecular Medicine Today* 5, 538–43. [Прогресс и актуальность баз данных единично-нуклеотидных полиморфизмов.]

Публикации по расшифровке генома человека, появившиеся в специальных выпусках журналов Nature от 15 февраля 2001 г., где изложены результаты поддерживаемого общественностью Проекта Генома человека, и Science от 16 февраля

WEB-РЕСУРСЫ: БАЗЫ ДАННЫХ ПО ГЕНОМАМ

Списки полных геномов:

`http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html http://www.ebi.ac.uk/genomes/mot/index.html http://pir.georgetown.edu/pirwww/search/genome.html`

Базы данных по отдельным организмам:

`http://www.unl.edu/stc-95/ResTools/biotools/biotools10.html http://www-fp.mcs.anl.gov/$\sim$gaasterland/genomes.html http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-db.html http://www.bioinformatik.de/cgi-bin/browse/Catalog/Databases/Genome\_Projects/`

2001 г., где изложены результаты, полученные Celera Genomics. Это исторические выпуски, важная веха в истории науки.

Выпуск мая 2001 г. журнала Genome Research, часть 11, номер 5, посвященный Геному человека.

Упражнения, задачи и компьютерные задания

Упражнение 2.1. Частоты встречаемости оснований в геноме *E. coli* составляют: A = T = 49.2%; G = C = 50.8%. Дана случайная последовательность из 4 639 221 нуклеотида, для которой справедливы написанные выше равенства. Сколько раз (в среднем) в этой последовательности встретится участок CTAG?

Упражнение 2.2. Геном *E. coli* содержит некоторое число пар ферментов, которые катализируют одну и ту же реакцию. Как это обстоятельство скажется на экспериментах по выявлению функции фермента методом генного нокаута (делеции или инактивации отдельных генов)?

Упражнение 2.3. Какие из категорий, использующихся для классификации функций белков дрожжей (см. с. 109), подойдут для классификации белков из прокариотического генома?

Упражнение 2.4. Что произошло раньше: полет человека на Луну или открытие глубоководных гидротермальных источников? Выскажите свое предположение перед тем, как обратиться к справочной литературе.

Упражнение 2.5. Синдром Гарднера — это состояние, при котором большое число полипов (выростов) развивается в нижней части желудочно-кишечного тракта, что при отсутствии лечения неизбежно приводит к раку. В каждом рассматриваемом случае один из родителей пациента также страдал этим заболеванием. Каким путем наследуется это заболевание?

Упражнение 2.6. Ген ретинобластомы передается вместе с геном эстеразы D, с которым он тесно сцеплен. Однако каждый из двух аллелей эстеразы D может передаваться с любым из аллелей ретинобластомы. Как можно показать, что ретинобластома не прямое следствие фенотипа эстеразы D?

Упражнение 2.7. Если все соматические клетки организма имеют одну и ту же последовательность ДНК, то почему необходимо иметь библиотеки кДНК из различных тканей?

Упражнение 2.8. Предположим, вы пытаетесь идентифицировать ген, вызывающий болезнь человека. Вы находите генетический маркер на расстоянии 0,75 сМ от гена болезни. Сколько пар оснований приблизительно насчитывает область, внутри которой может быть расположен ген, который вы ищите? Примерно сколько генов может содержать этот участок?

Упражнение 2.9. Наследственная зрительная нейропатия Лебера (Leber), LHON — это состояние, которое может вызвать потерю центрального зрения, возникающее вследствие мутаций в митохондриальной ДНК. Вам надо проконсультировать женщину с нормальной митохондриальной ДНК и мужчину с LHON, которые собираются пожениться. Какие сведения вы можете им сообщить о риске развития этого заболевания у их будущих детей?

Упражнение 2.10. Недостаточность глюкозо-6-фосфатдегидрогеназы — моногенный рецессивный, сцепленный с Х-хромосомой дефект, от которого страдают миллионы людей. Клиническая картина включает в себя гемолитическую

анемию и стойкую неонатальную желтуху. Этот ген не был элиминирован из популяции, потому что дает устойчивость к малярии. В этом случае общее знание метаболических путей выявило белок, вызывающий болезнь. Если дана аминокислотная последовательность белка, как вы определите место (места) расположения в хромосоме соответствующего гена?

Упражнение 2.11. До того как ДНК была признана носителем генетического материала, природа гена в биохимических деталях была темным вопросом. В 1940-х гг. Дж. Бидл (G. Beadle) и Э. Тэтам (E. Tatum) заметили, что единичные мутации могут вырезать отдельные этапы в биохимических путях. На основании этого они предположили гипотезу: один ген — один фермент. На копии иллюстрации (рис. 2.1) нарисуйте линии, связывающие гены с пронумерованными стадиями в последовательности реакций метаболического пути. В какой степени гены триптофанового оперона удовлетворяют гипотезе один ген — один фермент и в какой степени они представляют исключения?

Упражнение 2.12. Эта иллюстрация показывает пятую хромосому человека (*слева*) и соответствующую ей хромосому шимпанзе. На копии этой иллюстрации покажите, какие участки обнаруживают инверсию узора полос.

(Перепечатано с разрешения Yunis, J. J., Sawyer, J. R., and Dunham, K. (1980). 'The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee', ***Science*** **208, 1145–48. © 1980 American Association for the Advancement of Science.)**

Упражнение 2.13. Опишите в общих чертах, как изменится картина FISH на фотографии IV, если измененный участок двадцатой хромосомы будет не удален, а перемещен на другую хромосому.

Упражнение 2.14. Путем горизонтального переноса в геном *E. coli* в течение 14.4 млн лет были внесены 755 ORF, что является причиной *E. coli* дивергенции от *Salmonella*. Оцените среднюю скорость горизонтального переноса в (парах нуклеотидов)/год. Каков процент известных нам генов, внесенных в геном *E. coli* посредством горизонтального переноса?

Упражнение 2.15. В каком смысле геном живого организма напоминает базу данных? Какоие из следующих отличительных черт являются общими у реально существующих геномов и компьютерных баз данных? Какие отличительные черты реально существующих геномов отсутствуют у баз данных? Каких особенности баз данных не наблюдается у реально существующих геномов?

(а) Используется как хранилище информации.
(б) Самотолкование.
(в) Различные копии не идентичны.
(г) Ученые могут обнаружить ошибки.

(д) Ученые могут исправить ошибки.
(е) Имеется планомерная и организованная ответственность за собирание и распространение информации.

Задача 2.1. Какие экспериментальные факты показывают, что генетическая карта, соответствующая каждой хромосоме линейно упорядочена?

Задача 2.2. Для геномов *M. genitalium* и *H. influenzae* чему равны (а) генетическая плотность в генах/тпн, (б) средний размер гена в пн, (в) число генов? Какой фактор способствует такому сильному уменьшению размера генома в *M. genitalium* по сравнению с *H. influenzae*?

Задача 2.3. Рассчитано, что иммунная система человека продуцирует около 10^{15} антител. Возможно ли для такого большого числа белков каждому из них быть полностью закодированным индивидуальным геном, полученным с помощью дупликации гена и дивергенции? Размер типичного гена иммуноглобулина IgG равен примерно 2000 пн.

Интернет-задание 2.1. На фотокопиях рис. 1.2–1.4 отметьте позиции видов, для которых полные геномные последовательности известны. (`http://www.ebi.ac.uk/genomes/`)

Интернет-задание 2.2. Какие существуют различия между стандартным генетическим кодом и митохондриальным генетическим кодом позвоночных?

Интернет-задание 2.3. Каково хромосомное расположение человеческого гена миоглобина?

Интернет-задание 2.4. Сколько вхождений тетрануклеотида CTAG в геном *E. coli*? Является ли этот тетрануклеотид пере- или недопредставленным по отношению к ожидаемому, с учетом частот встречаемости нуклеотидов в геноме? (См. упр. 2.1.)

Интернет-задание 2.5. Постройте гистограмму общего числа полных геномных последовательностей в каждом году, начиная с 1995.

Интернет-задание 2.6. (а) Как много предсказанных ORF на хромосоме X дрожжей? (б) Как много тРНК генов?

Интернет-задание 2.7. Какая аминокислота полностью отсутствует в белках *M. genitalium*? Как генетический код *M. genitalium* отличается от стандартного кода?

Интернет-задание 2.8. У человека 1 сМ $\sim 10^6$ пн. Сколько приблизительно комплементарных пар на 1 сМ приходится у дрожжей?

Интернет-задание 2.9. Клетки спермы активно плавают и содержат митохондрии. При оплодотворении содержимое клетки спермы полностью переходит в ядро. Как получается, что митохондриальная ДНК полностью унаследована от матери?

Интернет-задание 2.10. Во врезке на с. 107 показаны дупликации и дивергенции, приводящие к человеческим α- и β-кластерам гена глобина. (а) У каких видов, тесно связанных с предками человека, такие дивергенции имеют место? (б) У каких видов, близкородственных предкам человека, возникает специфический паттерн экспрессии гемоглобина при развитии ($\zeta_2\epsilon_2$ = эмбриональный, $\alpha_2\gamma_2$ = эмбриональный, $\alpha_2\beta_2$ = взрослый)?

Интернет-задание 2.11. Языковые группы более тесно коррелированы с различными митохондриальными ДНК человека или последовательностями Y-хромосомы? Предложите объяснение для наблюдаемого результата.

Интернет-задание 2.12. Мутацией, приводящей к серповидно-клеточной анемии, является единственная замена основания A на T, приводящая к замене Glu → Val в β-цепи гемоглобина. Замена оснований происходит в последовательности 5′-GTGAG-3′ (нормальная) → GTGTG (мутантная). Какая рестриктаза может быть использована для обнаружения этого различия? Какова ее специфичность?

Интернет-задание 2.13. Какие мутации являются основной причиной фенилкетонурии (PKU)?

Интернет-задание 2.14. Найдите три примера мутаций в CFTR-гене (связанных с цистозным фиброзом), которые приводят к ухудшению функционирования, но не к полной блокировке хлоридного канала. Каковы клинические симптомы этих мутаций?

Интернет-задание 2.15. Приведите пример генетической болезни, которая: (а) аутосомально доминантна, (б) аутосомально рецессивна (иной, чем цистический фиброз), (в) X-сцепленная доминантна, (г) X-сцепленная рецессивна, (д) сцеплена с Y, (е) является следствием ненормальной митохондриальной ДНК (иная, чем Leber's Hereditary Optic Neuropathy).

Интернет-задание 2.16. (а) Определите штат в США, в котором новорожденные дети регулярно тестируются на гомоцистинурию. (б) Определите штат США, в котором новорожденные дети нерегулярно тестируются на гомоцистенурию. (в) Определите штат США, в котором новорожденные дети регулярно тестируются на биотинидазу. (г) Определите штат США, в котором новорожденные дети нерегулярно тестируются на биотинидазу. (д) К каким клиническим последствиям приводит отказ детектировать гомоцистенурию или дефицит биотинидазы?

Интернет-задание 2.17. (а) Какова нормальная функция белка, отсутствующая в болезни Менке? (б) Имеется ли гомолог этого гена в геноме *A. thaliana*? (в) Если да, то какова функция этого гена в *A. thaliana*?

Интернет-задание 2.18. Мышечная дистрофия Дюшена (DMD) — это X-сцепленная наследственная болезнь, приводящая к прогрессирующей мышечной слабости. Больные DMD обычно теряют способность ходить с 12 лет. Продолжительность их жизни не более 20–25 лет. Мышечная дистрофия Беккера (BMD) не так строго зависит от одного гена. К обеим болезням обычно приводят делеции в единственном гене дистрофине. Для DMD характерно полное отсутствие функционального белка, а для BMD характерен искаженный белок, сохраняющий свою функцию. Некоторые из делеций в случае BMD длиннее, чем другие, вызывающие BMD. Какие составляющие двух классов делеций приводят к этим двум болезням?

Интернет-задание 2.19. Какая хромосома коровы содержит участок, гомологичный региону 8q21.12 человеческой хромосомы?

3 Архивы и извлечение информации

Введение

Эта глава дает навыки извлечения информации, которые позволяют эффективно использовать базы данных. Это облегчит освоение базовых навыков работы с компьютером и их дальнейшее развитие. На самом деле, во многих базах данных включены системы обучения, которые облегчают изучение их возможностей.

Оглавление базы данных и терминология поисковых систем

Индекс — это набор указателей к информации в базе данных. Во время поиска в Web (WWW) или в специализированной базе данных молекулярной биологии, вы вводите один или несколько поисковых терминов, а программа проверяет их по своим таблицам индексов. Модель такова, что вся база данных состоит из записей — отдельных последовательных кусков информации. Программа извлечения информации представляет пользователю записи с содержанием, отвечающем запросу. Пример простейшей парадигмы — это когда вы вводите термин «лошадь» и программа выдает вам список записей, содержащей термин «лошадь».

Полный поиск в Web предоставил бы вам информацию о многих различных аспектах, связанных с лошадьми, таких как молекулярная биология, разведение, скачки, стихи о лошадях, большинство из которых вам совершенно не нужно. Для того чтобы поиск был удачным, недостаточно упомянуть, что вам нужно. Вы должны убедиться, что желаемые ответы не окажутся погребенными в массе ненужной ерунды. (Конечно, под ерундой мы имеем ввиду то, что могло бы заинтересовать других людей.)

Чтобы сфокусироваться на нужной информации, механизм ее изъятия выбирает множественные поисковые термины или ключевые слова. Поиск «лошадь_печень_алкогольдегидрогеназа» приведет вас к ответам, которые имеют отношение к этому ферменту. Поиск обнаружит информацию, содержащую все четыре ключевых слова, которые вы внесли: «и лошадь, и печень, и алкоголь, и дегидрогеназа». В ответах вы не найдете стихов о лошадях. (Если, конечно, в стихах не содержится всех четырех ключевых слов.)

Можно запрашивать различные логические комбинации индексовых терминов. Например, если поисковый механизм не имеет понятия о различном написании в Англии и в Америке, то было бы полезно указать в поиске оба варианта написания этого слова. (Необходимо избежать неправильного истолкования вашего термина, чтобы система не предоставила вам исключительно документы, написанные международными комитетами или полуграмотными эмигрантами.)

Если вы хотите узнать о других дегидрогеназах, вы можете запросить «не алкогольдегидрогеназу». Предоставляется информация, содержащая термин «дегидрогеназа», но не содержащая термин «алкоголь». Вы найдете там информацию о лактат-дегидрогеназе, малатдегидрогеназе и т. д. Вы опустите ссылки на обзоры, где сравниваются алкогольдегидрогеназа и другие дегидро-

геназы, а также встречаются перечни других дегидрогеназ. А эта информация могла бы вам пригодиться.

Многие поисковые механизмы базы данных позволяют использовать сложные логические выражения и помогают правильно написать сами термины. Составление таких выражений — это упражнения в заданной теории, которые помогают с черчением диаграмм Венна. Хотя логика поиска не зависит от программы, используемой в поиске по базе данных, различные программы требуют различных синтаксисов для использования тех же самых условий. Например, если вы ищете «не алкогольдегидрогеназу», возможно, нужно ввести «дегидрогеназа—алкоголь» или «дегидрогеназа!алкоголь».

Специализированные базы данных, включая молекулярно-биологические, предлагают пути получения информации, с тем чтобы разделить различные категории информации. Это важно. В настоящее время активно работающие ученые в области биомедицины известны: E(lizabetta) Coli, (John D.) Yeast, (Patrice) Rat, а так же многие Rabbits, несколько Crystals и Blots. Если вы хотите найти работы, опубликованные этими учеными, было бы наивно ограничиться обычным поиском в базе данных молекулярной биологии, введя только их фамилии. Многие базы данных обладают отдельными указателями и поисковой системой в различных категориях информации. Это позволяет вам находить работы, автором которых является E. Coli.

Некоторые категории, такие как таксономия, обладают контролируемыми словарями. Нередко они предстают перед пользователем в свернутом окне. Для того чтобы найти «глобин немлекопитающего» и выбрать сравнительно мало терминов о глобинах немлекопитающих, а не многочисленную терминологию о глобинах, включая человеческие гемоглобины, где просто не используют термин «млекопитающий», необходима система выбора, которая «понимает» таксономическую иерархию.

Техническая проблема, которая часто создает трудности — это как вводить термины, содержащие такие нестандартные знаки, как акценты или умляуты, греческие буквы и, как уже упоминалось, различия между американской и британской орфографиями. Специализированные базы данных, такие как NCBI's ENTREZ, могут справляться с различиями в американской и британской орфографии с помощью словаря синонимов. Программы, которые указываются в Web, часто не могут справиться с подобным заданием. Поэтому игнорируйте акцентировку и надейтесь на лучшее.

Какие еще вопросы могут возникать

Роясь в базах данных, вы вряд ли с первого раза найдете то, что нужно. Обычно приходится обновлять поиск на основе тех результатов, которые вы изначально получили. Большинство программ извлечения информации позволяет производить последовательный поиск с измененным набором поисковой терминологии или логических связей. Соответственно, как только вы нашли то, что искали, вы захотите продолжить поиск, чтобы найти родственный материал. Если вы нашли генетическую последовательность, возможно, вы захотите узнать о наличии гомологичных генов в других организмах или о том,

возможно ли получить пространственную структуру соответствующего белка. Или вы захотите узнать о работах, опубликованных об этой последовательности.

Для этих дополнительных поисков вам понадобятся связи между вводами в той же самой или другой базах данных. Это особый пример того, как «рыться» в электронных библиотеках — пример трудной проблемы.

Чтобы найти гомологичные гены, вам понадобятся связи с терминами в той же самой базе данных (в данном случае в базе данных генетических последовательностей). Чтобы найти структуры или библиографические ссылки на ген, вам понадобятся связи между различными базами данных (от базы данных генетических последовательностей до базы данных пространственных структур или библиографические базы данных). Взаимодействие баз данных в молекулярной биологии становится все более эффективным, так что эти операции сейчас уже довольно легко выполнимы, в то время как раньше приходилось проводить отдельные поиски в изолированных базах данных. Обратите внимание, что это обобщение изначальной модели базы данных как набора независимо вводимых терминов, которые могут быть выбраны только по их обозначенному содержанию.

Анализ полученных данных

В некоторых случаях для завершения исследовательской работы, вы захотите запустить программу, подавая на вход полученные результаты. Например, если вы установили последовательность интересующего белка и хотите произвести поиск с помощью PSI-BLAST. В целом, проблема — не только в поиске в базе данных; раньше вам пришлось бы отдельно вручную ввести полученную последовательность в программу. Теперь же, как и при поиске во множестве других баз данных, системы получения данных в молекулярной биологии часто предоставляют удобные средства для начала работы и представления результатов, что значительно облегчает работу.

Архивы

Несмотря на то что наши знания о биологических последовательностях и структурах очень далеки от полных, полученная информация уже весьма внушительна, причем она увеличивается чрезвычайно быстро благодаря труду многих ученых. Хранение и распространение информации производятся особыми организациями, создающими базы данных.

Архивы данных по биоинформатике первоначально создавали исследовательские группы, и мотивацией к этому был научный интерес. Вместе с ростом требований к техническому оснащению к персоналу также стали предъявляться повышенные требования, прежде всего это касается навыков работы на компьютере. Поэтому-то стало возможным создание специальных крупномасштабных национальных и даже интернациональных проектов. Любой, кто проследит за всей историей этих проектов, будет впечатлен их ростом

от маленьких, скромных и плохо финансируемых проектов, осуществляемых узким кругом людей, до многонационального крупномасштабного, совершившего по-настоящему научный переворот, и выполнявшего роль силового рычага на бизнес и политические взгляды.

Первые архивы данных, относящихся к биологическим макромолекулам, включали:

- Последовательности нуклеиновых кислот, вплоть до последовательностей геномов
- Аминокислотные последовательности белков
- Структуры белков и нуклеиновых кислот
- Кристаллические структуры малых молекул
- Функции белков
- Данные по экспрессии генов
- Научные публикации

Базы данных последовательностей нуклеиновых кислот

Мировой архив последовательностей нуклеиновых кислот возник благодаря партнерству трех организаций: Национального центра биотехнологической информации (National Center for Biotechnology Information) (США), Библиотеки данных Европейского института биоинформатики (EMBL Data Library (European Bioinformatics Institute)) (Великобритания) и Банка данных ДНК Японского национального института генетики (DNA Data Bank of Japan (National Institute of Genetics)). Эти учреждения ежедневно обмениваются информацией. Первоначальные данные были идентичны, несмотря на то что формат хранения и характер аннотаций несколько различаются. В эти базы поступают, хранятся и распространяются данные о последовательностях ДНК и РНК, полученные из проектов исследований геномов, научных публикаций и заявок на патенты. Чтобы эти фундаментальные данные находились в свободном доступе, научные журналы требуют при публикации статьи внесения новых нуклеотидных последовательностей в базу данных. Такие же условия действуют для публикаций по аминокислотным последовательностям и структурам белков и нуклеиновых кислот.

Как уже говорилось, базы данных последовательностей нуклеиновых кислот — это набор записей. Каждая из них имеет вид текстового файла, содержащего данные и аннотации, относящиеся к конкретной последовательности. Многие записи могут быть собраны из нескольких научных статей, описывающих пересекающиеся фрагменты полной последовательности.

Записи в базе данных имеют жизненный цикл. Из-за желания части пользователей получать быстрый доступ к данным, новые записи становятся доступными до того, как аннотация закончена и выполнена проверка. Записи проходят следующие стадии:

Неаннотированная → Предварительная → Непроверенная → Стандарт

Изредка случается, что запись «умирает» — удаляется из базы, так как была установлена ее неверность.

Пример последовательности ДНК из базы данных EMBL, включающая аннотацию, данные последовательности, — ген ингибитора бычьего панкреатического трипсина (приведена только часть записи, пропущена большая часть собственно последовательности).

Ген ингибитора бычьего панкреатического трипсина (последовательность ДНК из базы данных EMBL)

The EMBL data library entry for the bovine pancreatic trypsin inhibitor gene

```
ID   BTBPTIG    standard; DNA; MAM; 3998 BP.
XX
AC   X03365; K00966;
XX
DT   18-NOV-1986 (Rel. 10, Created)
DT   20-MAY-1992 (Rel. 31, Last updated, Version 3)
XX
DE   Bovine pancreatic trypsin inhibitor (BPTI) gene
XX
KW   Alu-like repetitive sequence; protease inhibitor;
KW   trypsin inhibitor.
XX
OS   Bos taurus (cattle)
OC   Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC   Theria; Eutheria; Artiodactyla; Ruminantia; Pecora; Bovidae.
XX
RN   [1]
RP   1-3998
RA   Kingston I.B., Anderson S.;
RT   "Sequences encoding two trypsin inhibitors occur in strikingly
RT   similar genomic environments";
RL   Biochem. J. 233:443-450(1986).
XX
RN   [2]
RA   Anderson S., Kingston I.B.;
RT   "Isolation of a genomic clone for bovine pancreatic trypsin
RT   inhibitor by using a unique-sequence synthetic dna probe";
RL   Proc. Natl. Acad. Sci. U.S.A. 80:6838-6842(1983).
XX
DR   SWISS-PROT; P00974; BPT1_BOVIN.
XX
CC   Data kindly reviewed (08-DEC-1987) by Kingston I.B.
XX
FH   Key             Location/Qualifiers
FH
FT   misc_feature    795..800
FT                   /note="pot. polyA signal"
FT   misc_feature    835..839
FT                   /note="pot. polyA signal"
FT   repeat_region   837..847
FT                   /note="direct repeat"
FT   misc_feature    930..945
FT                   /note="sequence homologous to Alu-like
FT                   consensus seq."
FT   repeat_region   1035..1045
FT                   /note="direct repeat"
FT   misc_feature    2456..2461
FT                   /note="pot. splice signal"
```

```
FT   CDS             2470..2736
FT                   /note="put. precursor"
FT   misc_feature    2488..2489
FT                   /note="pot. intron/exon splice junction"
FT   misc_feature    2506..2507
FT                   /note="pot. intron/exon splice junction"
FT   CDS             2512..2685
FT                   /note="trypsin inhibitor (aa 1-58)"
FT   misc_feature    2698..2699
FT                   /note="pot. exon/intron splice junction"
FT   misc_feature    3690..3695
FT                   /note="pot. polyA signal"
FT   misc_feature    3729..3733
FT                   /note="pot. polyA signal"
XX
SQ   Sequence 3998 BP; 1053 A; 902 C; 892 G; 1151 T; 0 other;
     aattctgata atgcagagaa ctggtaagga gttctgattg ttctgcttga ttaaatgggt
     tgtaacagga tagtgtcttg tcctgatcct agcattcata tggtgtgtgt tctggggcaa
     gtcatctgca gtttcttcac ctgaacaggg ggaccaggtt acatgagttt cttaaaagat
     taccagtcat gagtatgaag agtttacact ttcctgatca atgacgtcca tttcccatca

                                3720 nucleotides deleted...

     gccaggtcaa actttggggt gtgttatttc cctgaatt
//
```

Таблица особенностей (строки, начинающиеся с FT) — это составная часть аннотации записи, которые описывают свойства специфичных участков, отдельных кодирующих последовательностей (CDS). Это сделано для того, чтобы компьютерные программы могли пользоваться этими файлами, например, чтобы перевести кодирующий участок в аминокислотную последовательность — нужен более контролируемый формат и более узкий набор символов. Развитие определенных наборов символов, общих словарей и хранилищ для ключевых слов и списков особенностей также важны для установления ссылок между различными банками данных.

Особенность может указывать на участок последовательности, который:

- Выполняет функцию или влияет на функцию
- Взаимодействует с другими молекулами
- Участвует в репликации
- Участвует в рекомбинации
- Является повторяющимся элементом
- Имеет элементы вторичной или третичной структуры
- Изменен или исправлен

Геномные базы данных

Хотя геномные последовательности представлены в виде записей в стандартных архивах последовательностей нуклеиновых кислот, многие геномные проекты имеют специальные базы данных, которые сочетают геномные последовательности и их описание с другими установленными данными для этих видов.

WEB-РЕСУРСЫ: ССЫЛКИ НА СПЕЦИАЛЬНЫЕ БАЗЫ ДАННЫХ ПО ОРГАНИЗМАМ[1]

http://www.unl.edu/stc-95/ResTools/biotools/biotools10.html

http://www-fp.mcs.anl.gov/$\sim$gaasterland/genomes.html

http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-db.html

http://www.bioinformatik.de/cgi-bin/browse/Catalog/Databases/Genome_Projects/

[1] Безусловно, к этому списку следует добавить Human genome browser (http://genome.ucsc.edu/), который дает очень удобный доступ к геному человека и базе данных по сравнительной геномике Vista (http://genome.lbl.gov/vista/index.shtml).

Базы данных белковых последовательностей

Большинство данных по аминокислотным последовательностям сейчас появляется путем транслирования последовательностей нуклеиновых кислот. The Swiss Institute of Bioiformatic (Швейцарский институт биоинформатики) в сотрудничестве с базой данных EMBL снабжают аннотированными аминокислотными последовательностями базу данных SWISS-PROT. Другая база данных аминокислотных последовательностей PIR включает группы Национального центра биотехнологической информации (National Center for Biotechnology Information) (Джорджтаунский университет, Вашингтон, США), Мюнхенского информационного центра белковых последовательностей (Munich Information Center for Protein Sequence(MIPS)) (Мюнхен, Германия), и Японской международной базы данных белков (Japan International Protein Information Database (Tsucuba, Япония)).

Формат записей аминокислотных последовательностей белков в базе данных PIR на примере бычьего панкреатического ингибитора трипсина — показан в таблице (на с. 142). (Ср. с данными из базы SWISS-PROT — Интернет-задание 3.1.)

PIR entry for the amino acid sequence of Bovine pancreatic trypsin inhibitor

```
ENTRY            TIBO  #type complete
TITLE            basic proteinase inhibitor precursor - bovine
ALTERNATE_NAMES  aprotinin; basic pancreatic trypsin inhibitor; BPTI;
                 cationic kallikrein inhibitor; inhibitor IV
ORGANISM         #formal_name Bos primigenius taurus #common_name cattle
   #cross-references taxon:9913
DATE             24-Apr-1984 #sequence_revision 22-Jul-1994 #text_change
                 16-Jun-2000
ACCESSIONS       S00277; A30333; S10546; S02486; S28197; A90162; A92023;
                 A90736; A90927; A34658; A93977; S10062; A01205
REFERENCE        S00274
   #authors      Creighton, T.E.; Charles, I.G.
   #journal      J. Mol. Biol. (1987) 194:11-22
   #title        Sequences of the genes and polypeptide precursors for two
```

```
                bovine protease inhibitors.
    #cross-references MUID:87283904
    #accession   S00277
       ##molecule_type DNA; mRNA
       ##residues 1-100 ##label CR2
       ##cross-references GB:M20934; GB:X05274; NID:g162767;
                PIDN:AAD13685.1; PID:g162769

                          12 additional references deleted...

 COMMENT          Basic proteinase inhibitor is an intracellular polypeptide
                  found in many tissues, probably located in granules of
                  connective tissue mast cells.
 GENETICS
    #introns      34/1; 98/1
 CLASSIFICATION   #superfamily basic proteinase inhibitor; animal
                  Kunitz-type proteinase inhibitor homology
 KEYWORDS         serine proteinase inhibitor
 FEATURE
   1-20                  #domain signal sequence #status predicted #label
                         SIG\
   21-35                 #domain propeptide #status predicted #label PRO\
   36-100                #product basic proteinase inhibitor #status
                         experimental #label MAT\
   40-90                 #domain animal Kunitz-type proteinase inhibitor
                         homology #label BPI\
   40-90,49-73,65-86     #disulfide_bonds #status experimental\
   50                    #inhibitory_site Lys (trypsin, chymotrypsin,
                         kallikrein, plasmin) #status experimental
 SUMMARY          #length 100 #molecular_weight 10903

 SEQUENCE
                5        10        15        20        25        30
     1 M K M S R L C L S V A L L V L L G T L A A S T P G C D T S N
    31 Q A K A Q R P D F C L E P P Y T G P C K A R I I R Y F Y N A
    61 K A G L C Q T F V Y G G C R A K R N N F K S A E D C M R T C
    91 G G A I G P W E N L

 PDB structures most related to TIBO:
      1CBWD (36-93) 100.0%; 1BZ5E (36-93) 100.0%; 9PTI (36-91) 100.0%
      1BZXI (36-93) 100.0%; 1B0CB (36-93) 100.0%; 1CBWI (36-93) 100.0%

                          17 lines, containing 51 additional PDB entries, deleted...

ALIGNMENTS containing TIBO:
     FA2061 basic proteinase inhibitor - 328.8 1.0
     SA0572 basic proteinase inhibitor superfamily 328.8
     MO1603 basic proteinase inhibitor - 1561.0 1.0
Associated Alignments:
     DA1053 animal Kunitz-type proteinase inhibitor homology

Link to iProClass (Superfamily classification and Alignment):
     iProClass Report for TIBO at PIR.

                          17 lines, containing 51 additional PDB entries, deleted...

ALIGNMENTS containing TIBO:
     FA2061 basic proteinase inhibitor - 328.8 1.0
     SA0572 basic proteinase inhibitor superfamily 328.8
     MO1603 basic proteinase inhibitor - 1561.0 1.0
Associated Alignments:
     DA1053 animal Kunitz-type proteinase inhibitor homology

Link to iProClass (Superfamily classification and Alignment):
     iProClass Report for TIBO at PIR.
```

Информацию о лигандах, дисульфидных мостиках, объединениях субъединиц, посттрансляционных модификациях, гликозилировании, выполненных редактированиях мРНК и т. д. невозможно получить из генной последовательности. Например, руководствуясь только генной информацией, нельзя узнать, что человеческий инсулин — это димер, связанный дисульфидными мостиками. Базы данных белковых последовательностей собирают эту дополнительную информацию из литературы и добавляют подходящие аннотации.

Базы данных, близкие SWISS-PROT

ENZYME DB и PROSITE (набор мотивов) — это две базы данных, очень близкие SWISS-PROT.

ENZYME DB дает следующую информацию о ферментах:

- EC-номер: цифровая идентификация, которую назначает Комиссия по ферментам, утвержденная Международным объединением биохимии и молекулярной биологии (Ensyme Commission, International Union of Biochemistry and Molecular Biology; см. `http://www.chem.qmw.ac.uk/iubmb/enzyme/`)
- Рекомендуемое название
- Альтернативные названия (если есть)
- Каталитическая активность
- Кофакторы (если есть)
- Ссылки на SWISS-PROT и другие базы данных
- Названия болезней, связанных с нехваткой этого фермента, если таковые известны.

Первые две буквы каждой строки обозначают, что содержит эта строка. Например, ID = идентификационный номер, DE = описание = официальное название, AN = альтернативное название, CA = каталитическая активность, CF = кофакторы, CC = комментарии, DR = ссылки на базы данных (SWISS-PROT).

В базе данных PROSITE содержатся распространенные паттерны аминокислотных остатков ряда белков. Подобный паттерн (иначе мотив, подпись (signature), отпечаток (fingerprint) или образец) обычно возникает в семействе белков из-за требований к строению связывающих сайтов. Его консервативность влияет на скорость эволюции белка. Часто с их помощью можно определить отдаленное родство, что невозможно сделать на основе сравнения последовательностей. Для пирофосфатазы паттерн, построенный на основе консенсусной последовательности, выглядит следующим образом: `D-[SGN]-D-[PE]-[LIVM]-D-[LIVMGC]`. Три консервативных остатка аспарагиновой кислоты (D) связывают двухвалентные катионы металлов.

PIR и связанные с ним базы данных

PIR произошел от самой первой базы данных последовательностей, разработанной Маргарет О. Дэйхофф — один из самых первых ученых в биоинформатике — в Национальной биомедицинской исследовательской организации (НБИО, the National Biomedical Research Foundation, NBRF) Джорджтаунского университета, США. В 1988 г. НБИО объединилась с MIPS и Японской

международной базой данных белков (ЯМБДБ, the Japan International Protein Information Database, JIPID) для образования PIR-International.

PIR объединяет несколько белковых баз данных:

- PIR-PSD: основная база данных белковых последовательностей.
- iProClass: классификация белков на основе их строения и функций.
- ASDB: базы данных аннотаций и подобий. Каждая запись содержит ссылку на список гомологичных последовательностей.
- P/R-NREF: обширное неизбыточное (без повторяющихся документов) собрание более 800 000 белковых последовательностей из всех возможных источников
- NRL3D: база данных, содержащая последовательности и аннотации ко всем белкам с известной кристаллической структурой, хранящимся в Protein Data Bank
- ALN: база данных выравниваний белковых последовательностей
- RESID: база данных ковалентных модификаций белков (напомним, что важные особенности строения белковой молекулы, например дисульфидные мостики, невозможно определить только по ее первичной структуре, и они не будут включены в базы данных аминокислотных последовательностей, полученных только на основании трансляции нуклеотидной последовательности)

Разработчиками базы данных PIR была также создана Интегрированная среда анализа последовательностей (the Integrated Environment for Sequence Analysis, IESA) — Web-сайт поиска информации и проведения расчетов.

Интернет-сервер базы данных PIR (`http://pir.georgetown.edu`) предоставляет богатый выбор доступных средств поиска информации:

- поиск записей в базе данных по: сходству последовательностей, информации в самой последовательности или в аннотации, мотивам, профилям или

Пример из базы данных ENZYME DB

```
ID   1.14.17.3
DE   PEPTIDYLGLYCINE MONOOXYGENASE.
AN   PEPTIDYL ALPHA-AMIDATING ENZYME.
AN   PEPTIDYLGLYCINE 2-HYDROXYLASE.
CA   PEPTIDYLGLYCINE + ASCORBATE + O(2) = PEPTIDYL(2-HYDROXYGLYCINE) +
CA   DEHYDROASCORBATE + H(2)O.
CF   COPPER.
CC   -!- PEPTIDYLGLYCINES WITH A NEUTRAL AMINO ACID RESIDUE IN THE
CC   PENULTIMATE POSITION ARE THE BEST SUBSTRATES FOR THE ENZYME.
CC   -!- THE ENZYME ALSO CATALYZES THE DISMUTATION OF THE PRODUCT TO
CC   GLYOXYLATE AND THE CORRESPONDING DESGLYCINE PEPTIDE AMIDE.
DR   P10731, AMD_BOVIN;  P19021, AMD_HUMAN;  P14925, AMD_RAT;
DR   P08478, AMD1_XENLA;  P12890, AMD2_XENLA;
```

скрытым марковским моделям (гл. 5), структурной или функциональной классификации или встречаемости в заданных геномах
- статистический анализ и классификация записей
- классификация семейств геномов, полезная для создания аннотаций
- проведение анализа, включая парное и множественное выравнивание, поиск последовательностей по подобию
- разнообразные ссылки на другие базы данных, включая библиографические.

Базы данных структур

Базы данных структур предназначены для архивирования, аннотирования и распространения наборов координат атомов. Наиболее авторитетная база данных макромолекулярных биологических объектов — это Protein Data Bank (PDB). В нем содержится информация о структурах белков, нуклеиновых кислот и некоторых углеводов. Основанная Уолтером Гамильтоном мл. в Брукхэвенской национальной лаборатории, Лонг-Айленд, США в 1971 г., PDB ныне курируется Исследовательским объединением структурной биоинформатики (ИОСБ, Research Collaboratory for Structural Bioinformatics, RCSB) на базе Рутгерского университета, Нью-Джерси, Компьютерного центра в Сан-Диего, Калифорния, и Национального института стандартов и технологий в Мерилэнде, США. Основной сайт Protein Data Bank: `http://www.rcsb.org`. Официальные зеркальные сайты существуют в Европе, Сингапуре, Японии и Бразилии, а также в других регионах мира.

Домашняя страница PDB содержит ссылки на непосредственно сами данные, учебный и вспомогательный материалы, включая краткий обзор новостей (PDB NEWSLETTER), инструменты добавления новых записей и специализированные средства поиска структур.

В таблице показана часть записи в PDB о структуре тиоредоксина *E. coli*. В ней содержится информация:

- О белке: из какого организма он был получен.
- В каком растворителе находится белок и ссылки на публикации, описывающие определение структуры.
- Особенности эксперимента определения структуры, а также данные о качестве, такие как ренгенографическое разрешение и стереохимическая статистика.
- Аминокислотная последовательность
- Дополнительно включенные в структуру молекулы, включая кофакторы, ингибиторы и молекулы воды
- Элементы вторичной структуры: α-спиралей и β-тяжей (β-листов)
- Расположение дисульфидных мостиков
- Координаты атомов.

PDB запись 2TRX, *E. coli* тиоредоксин

```
HEADER    ELECTRON TRANSPORT                      19-MAR-90   2TRX
COMPND    THIOREDOXIN
SOURCE    (ESCHERICHIA $COLI)
AUTHOR    S.K.KATTI,D.M.LE*MASTER,H.EKLUND
REVDAT  2   15-JAN-93 2TRXA   1       HEADER COMPND
REVDAT  1   15-OCT-91 2TRX    0
JRNL        AUTH   S.K.KATTI,D.M.LE*MASTER,H.EKLUND
JRNL        TITL   CRYSTAL STRUCTURE OF THIOREDOXIN FROM ESCHERICHIA
JRNL        TITL 2 $COLI AT 1.68 ANGSTROMS RESOLUTION
JRNL        REF    J.MOL.BIOL.                   V. 212   167 1990
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   A.HOLMGREN,B.-*O.SODERBERG,H.EKLUND,C.-*I.BRANDEN
REMARK   1  TITL   THREE-DIMENSIONAL STRUCTURE OF ESCHERICHIA COLI
REMARK   1  TITL 2 THIOREDOXIN-*S=2= TO 2.8 ANGSTROMS RESOLUTION
REMARK   1  REF    PROC.NAT.ACAD.SCI.USA         V.  72  2305 1975
REMARK   1  REFN   ASTM PNASA6  US ISSN 0027-8424                 040
REMARK   1 REFERENCE 2
REMARK   1  AUTH   B.-*O.SODERBERG,A.HOLMGREN,C.-*I.BRANDEN
REMARK   1  TITL   STRUCTURE OF OXIDIZED THIOREDOXIN TO 4.5 ANGSTROMS
REMARK   1  TITL 2 RESOLUTION
REMARK   1  REF    J.MOL.BIOL.                   V.  90   143 1974
REMARK   1  REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070
REMARK   1 REFERENCE 3
REMARK   1  AUTH   A.HOLMGREN,B.-*O.SODERBERG
REMARK   1  TITL   CRYSTALLIZATION AND PRELIMINARY CRYSTALLOGRAPHIC
REMARK   1  TITL 2 DATA FOR THIOREDOXIN FROM ESCHERICHIA $COLI B
REMARK   1  REF    J.MOL.BIOL.                   V.  54   387 1970
REMARK   1  REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070
REMARK   2
REMARK   2 RESOLUTION. 1.68 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT. BY THE RESTRAINED LEAST-SQUARES PROCEDURE OF J.
REMARK   3  KONNERT AND W. HENDRICKSON AS MODIFIED BY B. FINZEL
REMARK   3  (PROGRAM *PROFFT*).  THE R VALUE IS 0.165 FOR 25969
REMARK   3  REFLECTIONS IN THE RESOLUTION RANGE 8.0 TO 1.68 ANGSTROMS
REMARK   3  WITH FOBS .GT. 3.0*SIGMA(FOBS)
REMARK   3
REMARK   3  RMS DEVIATIONS FROM IDEAL VALUES (THE VALUES OF
REMARK   3      SIGMA, IN PARENTHESES, ARE THE INPUT ESTIMATED
REMARK   3      STANDARD DEVIATIONS THAT DETERMINE THE RELATIVE
REMARK   3      WEIGHTS OF THE CORRESPONDING RESTRAINTS)
REMARK   3    DISTANCE RESTRAINTS (ANGSTROMS)
REMARK   3      BOND DISTANCE                          0.015(0.020)
REMARK   3      ANGLE DISTANCE                         0.035(0.030)
REMARK   3      PLANAR 1-4 DISTANCE                    0.055(0.050)
REMARK   3    PLANE RESTRAINT (ANGSTROMS)              0.021(0.020)
REMARK   3    CHIRAL-CENTER RESTRAINT (ANGSTROMS**3)   0.131(0.150)
REMARK   3    NON-BONDED CONTACT RESTRAINTS (ANGSTROMS)
REMARK   3      SINGLE TORSION CONTACT                 0.165(0.500)
REMARK   3      MULTIPLE TORSION CONTACT               0.174(0.500)
REMARK   3      POSSIBLE HYDROGEN BOND                 0.180(0.500)
REMARK   3    CONFORMATIONAL TORSION ANGLE RESTRAINT (DEGREES)
REMARK   3      PLANAR (OMEGA)                          4.0(3.0)
REMARK   3      STAGGERED                              16.3(15.0)
REMARK   3      ORTHONORMAL                            11.7(20.0)
REMARK   3    ISOTROPIC THERMAL FACTOR RESTRAINTS (ANGSTROMS**2)
REMARK   3      MAIN-CHAIN BOND                        1.38(1.000)
REMARK   3      MAIN-CHAIN ANGLE                       2.28(1.000)
REMARK   3      SIDE-CHAIN BOND                        1.97(1.000)
```

```
REMARK   3      SIDE-CHAIN ANGLE                              3.27(1.500)
REMARK   4
REMARK   4 THERE ARE TWO MOLECULES IN THE ASYMMETRIC UNIT.  THEY HAVE
REMARK   4 BEEN ASSIGNED CHAIN INDICATORS *A* AND *B*.  THEY HAVE BEEN
REMARK   4 REFINED INDEPENDENTLY WITHOUT IMPOSING NON-CRYSTALLOGRAPHIC
REMARK   4 SYMMETRY RESTRAINTS.
REMARK   5
REMARK   5 IN ADDITION TO THE METAL COORDINATION SPECIFIED ON CONECT
REMARK   5 RECORDS BELOW, THERE ARE BONDS TO OD1 AND OD2 OF ASP 10 IN
REMARK   5 A SYMMETRY-RELATED MOLECULE.  DUE TO SOME LIMITATIONS OF
REMARK   5 PROTEIN DATA BANK FORMAT, THESE BONDS CANNOT BE PRESENTED
REMARK   5 ON CONECT RECORDS.
REMARK   6
REMARK   6 CORRECTION. CORRECT CLASSIFICATION ON HEADER RECORD AND
REMARK   6  REMOVE E.C. CODE.  15-JAN-93.
SEQRES   1 A  108  SER ASP LYS ILE ILE HIS LEU THR ASP ASP SER PHE ASP
SEQRES   2 A  108  THR ASP VAL LEU LYS ALA ASP GLY ALA ILE LEU VAL ASP
SEQRES   3 A  108  PHE TRP ALA GLU TRP CYS GLY PRO CYS LYS MET ILE ALA
SEQRES   4 A  108  PRO ILE LEU ASP GLU ILE ALA ASP GLU TYR GLN GLY LYS
SEQRES   5 A  108  LEU THR VAL ALA LYS LEU ASN ILE ASP GLN ASN PRO GLY
SEQRES   6 A  108  THR ALA PRO LYS TYR GLY ILE ARG GLY ILE PRO THR LEU
SEQRES   7 A  108  LEU LEU PHE LYS ASN GLY GLU VAL ALA ALA THR LYS VAL
SEQRES   8 A  108  GLY ALA LEU SER LYS GLY GLN LEU LYS GLU PHE LEU ASP
SEQRES   9 A  108  ALA ASN LEU ALA
SEQRES   1 B  108  SER ASP LYS ILE ILE HIS LEU THR ASP ASP SER PHE ASP
SEQRES   2 B  108  THR ASP VAL LEU LYS ALA ASP GLY ALA ILE LEU VAL ASP
SEQRES   3 B  108  PHE TRP ALA GLU TRP CYS GLY PRO CYS LYS MET ILE ALA
SEQRES   4 B  108  PRO ILE LEU ASP GLU ILE ALA ASP GLU TYR GLN GLY LYS
SEQRES   5 B  108  LEU THR VAL ALA LYS LEU ASN ILE ASP GLN ASN PRO GLY
SEQRES   6 B  108  THR ALA PRO LYS TYR GLY ILE ARG GLY ILE PRO THR LEU
SEQRES   7 B  108  LEU LEU PHE LYS ASN GLY GLU VAL ALA ALA THR LYS VAL
SEQRES   8 B  108  GLY ALA LEU SER LYS GLY GLN LEU LYS GLU PHE LEU ASP
SEQRES   9 B  108  ALA ASN LEU ALA
FTNOTE   1
FTNOTE   1 RESIDUES PRO A 76 AND PRO B 76 ARE CIS PROLINES.
FTNOTE   2
FTNOTE   2 RESIDUES HIS A 6, LEU A 7, ILE A 23, ASP A 47, GLU A 48,
FTNOTE   2 LEU A 58, LEU A 80, HIS B 6, ASP B 47, LEU B 58, AND
FTNOTE   2 LEU B 80 HAVE BEEN MODELED AS TWO CONFORMERS.
FTNOTE   3
FTNOTE   3 RESIDUES 11 - 21 IN CHAIN B ARE DISORDERED.
HET      CU     109       1     COPPER ++ ION
HET      CU     109       1     COPPER ++ ION
HET     MPD     601       8     2-METHYL-2,4-PENTANEDIOL
HET     MPD     602       8     2-METHYL-2,4-PENTANEDIOL
HET     MPD     603       8     2-METHYL-2,4-PENTANEDIOL
HET     MPD     604       8     2-METHYL-2,4-PENTANEDIOL
HET     MPD     605       8     2-METHYL-2,4-PENTANEDIOL
HET     MPD     606       8     2-METHYL-2,4-PENTANEDIOL
HET     MPD     607       8     2-METHYL-2,4-PENTANEDIOL
HET     MPD     608       8     2-METHYL-2,4-PENTANEDIOL
FORMUL   3   CU    2(CU1 ++)
FORMUL   4  MPD    8(C6 H14 O2)
FORMUL   5  HOH   *140(H2 O1)
HELIX    1 A1A SER A   11  LEU A   17  1 DISORDERED IN MOLECULE B
HELIX    2 A2A CYS A   32  TYR A   49  1 BENT BY 30 DEGREES AT RES 39
HELIX    3 A3A ASN A   59  ASN A   63  1
HELIX    4 31A THR A   66  TYR A   70  5 DISTORTED H-BONDING C-TERMINS
HELIX    5 A4A SER A   95  LEU A  107  1
HELIX    6 A1B SER B   11  LEU B   17  1 DISORDERED IN MOLECULE B
HELIX    7 A2B CYS B   32  TYR B   49  1 BENT BY 30 DEGREES AT RES 39
HELIX    8 A3B ASN B   59  ASN B   63  1
HELIX    9 31B THR B   66  TYR B   70  5 DISTORTED H-BONDING C-TERMINS
```

```
HELIX   10 A4B SER B   95  LEU B  107  1
SHEET    1 B1A 5 LYS A   3  THR A   8  0
SHEET    2 B1A 5 LEU A  53  ASN A  59  1  O  VAL A  55   N  ILE A   5
SHEET    3 B1A 5 GLY A  21  TRP A  28  1  N  TRP A  28   O  LEU A  58
SHEET    4 B1A 5 PRO A  76  LYS A  82 -1  O  THR A  77   N  PHE A  27
SHEET    5 B1A 5 VAL A  86  GLY A  92 -1  N  GLY A  92   O  LYS A  82
SHEET    1 B1B 5 LYS B   3  THR B   8  0
SHEET    2 B1B 5 LEU B  53  ASN B  59  1  O  VAL B  55   N  ILE B   5
SHEET    3 B1B 5 GLY B  21  TRP B  28  1  N  TRP B  28   O  LEU B  58
SHEET    4 B1B 5 PRO B  76  LYS B  82 -1  O  THR B  77   N  PHE B  27
SHEET    5 B1B 5 VAL B  86  GLY B  92 -1  N  GLY B  92   O  LYS B  82
TURN     1 T1A THR A   8  SER A  11      III (TYPE I IN MOLECULE B)
TURN     2 T2A ALA A  29  CYS A  32      I
TURN     3 T3A TYR A  49  LYS A  52      II
TURN     4 T4A GLY A  74  THR A  77      VIB (INCLUDES CIS PRO 76)
TURN     5 T5A LYS A  82  GLU A  85      I'
TURN     6 T1B THR B   8  SER B  11      I (TYPE III IN MOLECULE A)
TURN     7 T2B ALA B  29  CYS B  32      I
TURN     8 T3B TYR B  49  LYS B  52      II
TURN     9 T4B GLY B  74  THR B  77      VIB (INCLUDES CIS PRO 76)
TURN    10 T5B LYS B  82  GLU B  85      I'
SSBOND   1 CYS A   32    CYS A   35
SSBOND   2 CYS B   32    CYS B   35
CRYST1  89.500   51.060   60.450  90.00 113.50  90.00 C 2          8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.011173  0.000000  0.004858        0.00000
SCALE2      0.000000  0.019585  0.000000        0.00000
SCALE3      0.000000  0.000000  0.018039        0.00000
ATOM      1  N   SER A   1      21.389  25.406  -4.628  1.00 23.22
ATOM      2  CA  SER A   1      21.628  26.691  -3.983  1.00 24.42
ATOM      3  C   SER A   1      20.937  26.944  -2.679  1.00 24.21
ATOM      4  O   SER A   1      21.072  28.079  -2.093  1.00 24.97
ATOM      5  CB  SER A   1      21.117  27.770  -5.002  1.00 28.27
ATOM      6  OG  SER A   1      22.276  27.925  -5.861  1.00 32.61
ATOM      7  N   ASP A   2      20.173  26.028  -2.163  1.00 21.39
ATOM      8  CA  ASP A   2      19.395  26.125  -0.949  1.00 21.57
ATOM      9  C   ASP A   2      20.264  26.214   0.297  1.00 20.89
ATOM     10  O   ASP A   2      19.760  26.575   1.371  1.00 21.49
ATOM     11  CB  ASP A   2      18.439  24.914  -0.856  1.00 22.14
ATOM     12  CG  ASP A   2      19.199  23.629  -0.576  1.00 23.23
ATOM     13  OD1 ASP A   2      20.107  23.371  -1.387  1.00 22.71
ATOM     14  OD2 ASP A   2      18.905  22.959   0.420  1.00 23.61

                              ...protein atoms deleted

ATOM    844  N   ALA A 108      41.357  21.341   9.676  1.00 42.93
ATOM    845  CA  ALA A 108      42.151  20.619  10.674  1.00 46.31
ATOM    846  C   ALA A 108      42.632  19.312  10.013  1.00 48.21
ATOM    847  O   ALA A 108      41.703  18.483   9.767  1.00 49.54
ATOM    848  CB  ALA A 108      41.441  20.369  11.988  1.00 46.65
ATOM    849  OXT ALA A 108      43.857  19.249   9.766  1.00 49.19
TER     850      ALA A 108

                              ...protein atoms deleted

ATOM    844  N   ALA A 108      41.357  21.341   9.676  1.00 42.93
ATOM    845  CA  ALA A 108      42.151  20.619  10.674  1.00 46.31
ATOM    846  C   ALA A 108      42.632  19.312  10.013  1.00 48.21
ATOM    847  O   ALA A 108      41.703  18.483   9.767  1.00 49.54
ATOM    848  CB  ALA A 108      41.441  20.369  11.988  1.00 46.65
ATOM    849  OXT ALA A 108      43.857  19.249   9.766  1.00 49.19
TER     850      ALA A 108
```

```
                ...second chain, and methane-pentane diol molecules deleted

HETATM 1749  O   HOH   401      30.339  33.478  16.727  1.00 17.61
HETATM 1750  O   HOH   402      29.396  44.583   6.834  0.95 17.71

                          ...72 additional water molecules deleted
```

Белковая база данных PDB частично дублирует информацию других баз данных. В Кембриджском архиве кристаллических структур (CCDC) хранятся данные о структурах малых молекул; олигонуклеотиды, представленные там, есть и в PDB. Эта информация крайне полезна при изучении конформаций компонентов биологических макромолекул, а также, при исследование взаимодействий лиганда с макромолекулой. База данных структур нуклеиновых кислот (NDB) университета Ратгерса в Нью-Брансуике, Нью-Джерси (США) дополняет PDB. Банк BioMagResBank факультета биохимии в Висконсинском университете, Мэдисон, Висконсин (США) содержит белковые структуры, определенные при помощи ядерного магнитного резонанса (ЯМР, NMR).

В архивах собраны не только результаты определения структуры, но и те измерения, на основе которых были получены эти данные. Новые структуры, попадающие в базу данных PDB, определены при помощи рентгеноструктурного анализа, а в BioMagResBank — при помощи ЯМР.

Каждой структуре в PDB дается четырехзначный идентификационный номер. Первый знак в этом идентификаторе — это всегда цифра от 1 до 9. Не ищете в этом какого-то определенного смысла. Во многих случаях имеется несколько структур одного и того же белка, полученных при разных условиях связывания с лигандом, находящихся в разных кристаллических формах, полученных с использованием более удачных кристаллов или сделанных при помощи новых, более точных методов. Например, по меньшей мере, есть 4 структуры миоглобина из спермы кита, полученных в ходе разных кристаллографических экспериментов.

Получить структурные данные легко, если вы знаете идентификационный номер. Вы получите резюме по данной структуре, если на домашней странице базы данных PDB (RCSB) введете идентификатор (PDB ID) и нажмете кнопку Explore (исследовать). На рис. 3.1 показано как выглядит такое резюме для тиоредоксина с идентификатором 1TRX. На странице есть следующие ссылки:

- на базу данных PubMed, где можно найти статью с описанием структуры,
- на рисунки с изображением структуры (во многих случаях для просмотра необходимо поставить на компьютер специальные программы визуализации),
- на файл с самой структурой,
- на список родственных структур, причем родство может определяться при помощи нескольких разных классификаций,
- на ресурсы, содержащие другую информацию о данной структуре,
- на последовательность и ее связь со вторичной структурой,

PDB
PROTEIN DATA BANK

Structure Explorer - 2TRX

Summary Information

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details
Crystallization Info
SearchLite SearchFields

Title: **Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution.**
Compound: **Thioredoxin**
Authors: **S. K. Katti, D. M. LeMaster, H. Eklund**
Exp. Method: **X-ray Diffraction**
Classification: **Electron Transport**
Source: **Escherichia coli**
Primary Citation: **Katti, S. K., LeMaster, D. M., Eklund, H.: Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution. *J Mol Biol* 212 *pp.* 167 (1990) [Medline]**

Deposition Date: **19-Mar-1990** *Release Date:* **15-Oct-1991**

Resolution [Å]: **1.68** *R-Value:* **0.165**
Space Group: **C 2**
Unit Cell: *dim [Å]:* *a* **89.50** *b* **51.06** *c* **60.45**
angles [°]: *alpha* **90.00** *beta* **113.50** *gamma* **90.00**

Polymer Chains: **A, B** *Residues:* **216**
Atoms: **1842**
HET groups:

ID	*Name*	*Formula*
CU	COPPER (II) ION	CU_1
MPD	2-METHYL-2,4-PENTANEDIOL	$C_6H_{14}O_2$

© RCSB

Рис. 3.1. Резюме по структуре TRX, тиоредоксин из *E. coli*.

- на некоторые подробности о форме кристалла и методов, с помощью которых этот кристалл был получен.

Все хорошо, если знаешь идентификатор, но если нет, то как же найти структуру в этом случае? На домашней странице PDB есть простой инструмент SearchLite, позволяющий искать по ключевым словам. Если вы введете coli thioredoxin, то получите 20 структур молекулы и ее мутантов, включая 1TRX, но вы получите и нуклеазу стафилококка, потому что в файле со структурой нуклеазы упоминается слово тиоредоксин в заголовке. Полученная таким поиском информация легко позволит вам выбрать структуры для просмотра или анализа, в соответствие с вашими интересами и задачами.

По базе данных PDB также можно произвести более комплексный поиск. База данных структур макромолекул Европейского биоинформационного института (EBI) предлагает весьма полезный список программ для поиска

и просмотра PDB, включая такой поисковый инструмент как OCA. OCA — это база данных, предназначенная для просмотра и навигации в структурах и функциях белков, приводящая суммарную информацию из различных баз данных. Первоначально OCA была разработана Дж. Прилуски (J. Prilusky) и поддержана EBI, на Web-сайте которого она доступна, но также ее можно найти на множестве других Web-сайтов. (Имя OCA, переводится с испанского, как коза, и имеет такое же отношение к PDB, как компьютер HAL Э. С. Кларка в фильме «2001» к IBM.)

Другой полезный ресурс, доступный на Web-сайте EBI, — это база данных предполагаемых четвертичных структур биологически активных форм белков (PQS). Часто асимметричная структура, содержащаяся в файле, является только частью активного комплекса, или наоборот один файл содержит несколько копий активной единицы. Во многих случаях непросто определить, как соотносятся расшифрованная структура с активной формой молекулы.

Индикаторы качества структуры

Рентгеноструктурный анализ позволяет рассчитать позиции и вероятное нахождение атомов в молекуле, так называемый *B-фактор*. Основной особенностью экспериментальных данных (абсолютных значений коэффициентов ряда Фурье функции электронной плотности) является то, что все атомы вносят вклад во все рефлексы. Трудно оценить ошибки при расчете положений отдельных атомов.

Успешность определения кристаллической структуры зависит от упорядоченности различных частей молекулы (расстояние, через которое точно повторяется данная элементарная ячейка кристалла).

Степень упорядоченности зависит от разрешения экспериментальных данных. Разрешение — показатель качества данных рентгеноструктурного анализа структуры; оно равно отношению числа экспериментально определяемых параметров к числу рефлексов. При расчете структур малых органических молекул или минералов это отношение обычно приближается к 10. Но для белковых молекул обычные величины разрешения такие:

	Низкое разрешение			...	Высокое	
Разрешение, Å	4.0	3.5	3.0	2.5	2.0	1.5
Отношение наблюдений к вычисленным параметрам	0.3	0.4	0.6	1.1	2.2	3.8

(Разрешение показывает, насколько мелкие детали могут быть определены; следовательно, чем меньше величина разрешения, тем выше разрешение.)

Помимо неупорядоченности ошибки структурных определений также могут быть вызваны, как ошибками в данных, так и ошибками в расшифровке структуры.

Многие кристаллографы публикуют расшифрованную структуру вместе с экспериментальными данными. Это позволяет подробнейшим образом проверить результаты. Но все-таки во многих случаях экспериментальные дан-

ные оказываются недоступными. Как же в такой ситуации можно оценить качество работы по определению кристаллической структуры? Важным ключом к решению этой проблемы является B-фактор; высокое значение этого фактора в целом участке молекуле говорит о том, что этот участок не был хорошо распознан. Это чаще всего связано с нарушениями упорядоченности в кристалле. Существуют программы, которые позволяют определить стереохимических маргиналов — атомы и части молекулы, отклоняющиеся от стандартных значений, рассчитанных на основании хорошо расшифрованных молекул. Анализ файла из базы данных PDB можно провести на сервере по адресу `www.cmbi.kun.nl/gv/pdbreport`, который позволит узнать о возможных трудных местах в структуре и определить маргиналов.

Но несмотря на относительную легкость обнаружения странных значений, сложно решить, являются ли они истинными, но необычными особенностями структуры, либо результатом ошибок при построении модели или следствием неупорядоченности в кристалле. Для надлежащей оценки необходимы экспериментальные данные. Для фиксации реальных ошибок может понадобиться участие опытных кристаллографов. Итог однозначен: структурные факторы должны складываться в архивы и быть доступны.

Ядерный магнитный резонанс (ЯМР)

Ядерный магнитный резонанс — это второй основной метод определения структуры макромолекул. С его помощью получают структуры в целом верной топологии, но не настолько точные, как при хорошем рентгеноструктурном анализе, и, следовательно, менее полезные для изучения тонких деталей структур. (Следует, однако, иметь в виду, что ЯМР-структуры являются структурами белка в растворе — воде, в то время как рентгеноструктурные данные соответствуют структуре в кристалле, что не совсем одно и то же. — *Прим. ред.*) Кристаллографы описывают только одну или небольшое количество структур. ЯМР-спектроскописты создают обычно семейство из $\sim$ 10–20 родственных структур (или даже больше), рассчитанных по одним и тем же экспериментальным данным. Сравнение внутри такого набора указывает на точность. Области, в которых локальные вариации структуры малы, хорошо определяются из данных. Это приблизительный эквивалент кристаллографического B-фактора.

Классификации белковых структур

Несколько Web-сайтов предлагают иерархические классификации всего PDB в соответствии с характером укладки белка:

- SCOP: Structural Classification of Proteins (Структурная классификация белков)
- CATH: Class/Architecture/Topology/Homology (Класс/Архитектура/Топология/Гомология)
- DALI: Основана на выделении схожих структур из отдаленных объектов
- CE: База данных структурных выравниваний

WEB-РЕСУРСЫ: WEB-РЕСУРСЫ СТРУКТУРЫ БЕЛКОВ И НУКЛЕИНОВЫХ КИСЛОТ

Домашняя страница белкового банка данных Protein Data Bank (PDB):
`http://www.rcsb.org`

Домашняя страница базы данных структур макромолекул на EBI:
`http://msd.ebi.ac.uk/`

Домашняя страница BioMagResBank:
`http://www.bmrb.wisc.edu/`

Поиск белковых банков данных:
Домашняя страница SCOP (Structural classification of proteins — Структурная классификация белков):
`http://scop.mrc-lmb.cam.ac.uk/scop/`

Перечень браузеров:
`http://pdb-browsers.ebi.ac.uk/browse\_it.shtml`

OCA:
`http://oca.ebi.ac.uk/oca-bin/ocamain`

База данных четвертичных структур белков:
`http://pqs.ebi.ac.uk/`

Сообщения/отчеты о качестве структур:
`http://www.cmbi.kun.nl/gv/pdbreport`

Это основные полезные Web-сайты с данными о белковых структурах. Например, SCOP предоставляет возможность поиска по ключевому слову для идентификации структуры, навигацию вверх и вниз по иерархии, генерирование рисунков, доступ к аннотационным записям в PDB и ссылки на соответствующие базы данных.

Специализированные, или локальные, базы данных

Многие ученые или группы выбирают, аннотируют и объединяют данные, сосредоточенные на отдельных темах, и делают ссылки на информацию на интересующую тему.

Например, ресурс по белку протеинкиназе — специализированная компиляция, включающая последовательности, структуры, информацию о функциях, лабораторные методы, список заинтересованных ученых, сервис для анализа, доску объявлений и ссылки.

База данных HIV-протеазы хранит структуры протеиназ вируса иммунодефицита человека 1-го типа (HIV-1), протеаз вируса иммунодефицита человека 2-го типа, протеаз вируса иммунодефицита обезьян и их комплексов; и предоставляет сервис для анализа и ссылки на другие Web-сайты с информацией, имеющей отношение к СПИДу. Эти базы данных содержат некоторые кристаллические структуры, не размещенные в PDB.

VIPER (Virus Particle ExploreR — исследователь вирусных частиц) имеет дело с кристаллическими структурами икосаэдрических вирусов.

В иммунологии:

- IMGT (ImMunoGeneTics), международная иммуногенетическая база данных — высококачественная интегральная база данных, специализирующаяся на молекулах иммуноглобулинов (Ig), рецепторов Т-клеток (TcR) и главного комплекса гистосовместимости (MHC) всех видов позвоночных. Сервер IMGT предоставляет открытый доступ ко всем иммуногенетическим данным. В настоящее время IMGT включает две базы данных: IMGT/LIGM-DB — полная база данных генов иммуноглобулинов и рецепторов Т-клеток человека и других позвоночных с трансляцией для полностью аннотированных последовательностей и IMGT/HLA-DB — база данных главного комплекса гистосовместимости человека
- KABAT — база данных по белковым последовательностям иммунологической направленности Северо-Западного Университета (США)
- MHCPEP (Major Histiocompability Complex Binding Peptides Database) — база данных по белкам, связывающимся с главным комплексом гистосовместимости Института Уолтера и Элизы Холл (Мельбурн, Австралия).

Базы данных по экспрессии и протеомике

Вспомним главное: ДНК делает РНК, которая делает белок. Геномные базы данных содержат последовательности ДНК. В экспрессионных базах записаны данные по уровням экспрессии мРНК, определенные обычно с помощью EST (короткие терминальные последовательности кДНК, синтезированные с мРНК), описывающих паттернов транскрипции генов (современные EST — не обязательно терминальные. — *Прим. ред.*). Протеомные базы данных содержат данные по белкам, описывая паттерны трансляции генов.

Сравнение профилей экспрессии дает ключ к (1) функции и механизму действия продуктов генов, (2) тому, как организмы координируют контроль над метаболическими процессами в различных условиях, например дрожжи в аэробном и анаэробном состоянии, (3) вариациям в активизации генов на разных стадиях клеточного цикла или развития организма, (4) механизмам бактериальной устойчивости к антибиотикам и выбору потенциальных мишеней для действия медикамента, (5) механизмам ответа на заражение паразитом (см. цветную вклейку, рис. V), (6) механизмам ответа на медикаменты разных типов и дозировок для проведения эффективной терапии.

Существует множество банков данных EST. В основном записи содержат поля, указывающие ткань-источник и/или внутриклеточное расположение, стадию развития, условия роста и количественный анализ уровня экспрессии.

WEB-РЕСУРСЫ: БАЗЫ ДАННЫХ ДЛЯ СПЕЦИФИЧЕСКИХ БЕЛКОВЫХ СЕМЕЙСТВ

Протеин киназы

`http://www.sdsc.edu/kinases/`

HIV (ВИЧ) протеазы

`http://www-fbsc.ncifcrf.gov/HIVdb/`

Икосаэдрические вирусы

`http://mmtsb.scripps.edu/viper/main.html`

Иммунология

IGMT: `http://imgt.cines.fr`

KABAT: `http://immuno.bme.nwu.edu/`

MHCPEP: `http://wehih.wehi.edu.au/mhcpep/`

Набор ссылок на базы данных по специфическим белковым семействам

`http://www2.ebi.ac.uk/msd/Links/family.shtml`

В GenBank коллекция dbEST на настоящий момент содержит почти $9 \cdot 10^6$ статей для 348 видов (см. таблицу на с. 157).

Некоторые EST коллекции специализируются на отдельных разновидностях тканей (например, мышцах, зубах) или на конкретных видах. Часто делается попытка связать экспрессию с остальными сведениями об организме. Например, Jackson Lab Gene Expression Information Resource Project for Mouse Development согласовывает данные об экспрессии генов и анатомию развития.

Многие банки данных предоставляют связи между EST разных видов, например гомологию человека и мыши, или взаимосвязи между генами болезни человека и белками дрожжей. Другие коллекции EST специализируются на отдельных типах белков, например цитокинах. Большое внимание уделяется проблеме рака: совокупная информация о мутациях, хромосомных перестройках и изменениях в экспрессируемом паттерне — для того чтобы идентифицировать генетические изменения во время образования и развития опухоли.

Хотя, конечно, существует близкая взаимосвязь между участками транскрипции и участками трансляции, непосредственные измерения белкового содержимого в клетках и тканях (этим занимается протеомика) предоставляют дополнительную ценную информацию. Из-за разности скоростей трансляции различных мРНК непосредственное измерение белка дает более точное описание паттернов экспрессии генов, чем измерение во время транскрипции. Посттрансляционные модификации могут быть обнаружены *только* с помощью исследования белков.

Виды с наибольшим числом записей в dbEST

Виды	Число записей
Человек	3 733 147
Мышь	2 077 301
Крыса	316 344
Плодовая муха	181 552
Соя	180 830
Бык	169 756
Нематоды	135 203
Томаты	126 562
Дикая горчица	113 330
Гладкая шпорцевая лягушка	103 291
Кукуруза	102 551
Данио рерио	100 075
Свинья	91 938

Протеомный анализ включает в себя разделение, идентификацию и количественный анализ белков, представленных в пробе. Он основан на разделении белков с помощью двумерного электрофореза и распознавании отдельных компонентов методом масс-спектрометрии. Протеомные базы данных содержат изображения гелей и их интерпретацию относительно белковых паттернов. Некоторые базы данных показывают изображения и позволяют осуществлять интерактивный выбор точечными метками. При выборе метки открывается окно с соответствующей статьей. Для каждого белка в статье обычно записано (см. Интернет-задание 3.21):

- идентификатор белка
- относительная масса
- функции
- механизм работы
- паттерн эксперссии
- клеточная локализация
- родственные белки
- посттрансляционные модификации
- взаимодействия с другими белками
- связи с другими базами данных

Биоинформатика вносит свой вклад в развитие этих баз данных, а также в разработку алгоритмов сравнения и анализа белковых паттернов, содержащихся в них.

Банки данных метаболических путей

Банк данных KEGG (Kyoto Encyclopedia of Genes and Genomes) собирает отдельные геномы, продукты и функции генов, но его особое достоинство заключается в объединении биохимической и генетической информации. KEGG сосредотачивается на взаимодействиях: молекулярной сборке, метаболических и регуляторных сетях. Его координатором является М. Канехиса.

KEGG организует/систематизирует пять типов данных во всеобъемлющую/полную систему:

1. Каталоги химических структур живых клеток
2. Каталоги генов
3. Карты геномов
4. Карты путей
5. Таблицы ортологов

Каталоги химических соединений и генов (1 и 2) содержат информацию о конкретных молекулах и последовательностях. Геномные карты (3) объединяют информацию о генах в соответствии с тем, как они следуют в хромосомах. В некоторых случаях знание о том, что ген входит в состав оперона, может дать ключ к пониманию его функции.

Карты биохимических путей (4) описывают сети взаимосвязанных молекулярных функций, как метаболических, так и регуляторных. Так, карта метаболического пути в KEGG — это идеализированное представление большого числа возможных реальных метаболических каскадов различных организмов. Каждая карта может служить основой для воссоздания реального метаболического пути конкретного организма. Для этого ферментам, обозначающим звенья в общей схеме, ставятся в соответствие конкретные белки данного организма.

Конкретный фермент организма может быть найден в KEGG в таблице ортологов (5), которая устанавливает связь фермента с его родственниками в других организмах. Это позволяет анализировать родство метаболических путей из разных организмов.

Основа эффективности KEGG заключается в очень плотной сети ссылок между этими типами информации и в наличии дополнительных ссылок на внешние источники, доступ к которым поддерживается системой. Вот два примера вопросов, ответы на которые можно получить с помощью KEGG.

- Предполагается, что простые метаболические пути эволюционируют в более сложные благодаря удвоению генов и их последующему расхождению. Поиск наборов сходных ферментов в каталоге биохимических путей позволяет выявить кластеры паралогов.
- KEGG позволяет проверить, укладывается ли определенный набор ферментов какого-либо организма в одну из известных схем метаболических путей. Пустое звено в схеме указывает на отсутствие в организме соответствующего фермента либо на существование неизвестного альтернативного пути.

Библиографические базы данных

MEDLINE (основанная на базе National Library of Medicine, USA) объединяет медицинскую литературу, в том числе огромное количество работ, посвященных вопросам молекулярной биологии, не обязательно непосредственно клинического характера. MEDLINE входит в состав PubMed — библиографической базы данных, которая предоставляет рефераты научных статей и интегрирована с другими информационными поисковыми системами на базе Национального центра биотехнологической информации (National Center of Biotechnology Information) входящего в Национальную медицинскую библиотеку (the National Library of Medicine) при Национальных институтах здоровья (`http://www.ncbi.nlm.nih.gov/PubMed/`).

Чрезвычайно эффективным свойством PubMed является возможность поиска связанных по тематике статей. Это очень быстрый способ разобраться в литературе на интересующую тему. В сочетании с использованием неспециализированных поисковых Web-систем, не предназначенных конкретно для поиска научных статей, PubMed позволяет легко получать практически полную информацию по большинству тем. *Подсказка.* Если вы хотите начать сбор информации на незнакомую тему, попробуйте добавить ключевое слово *tutorial* при поиске с помощью неспециализированной поисковой системы или слово *review* при поиске в PubMed.

Почти все научные журналы сейчас помещают на Web-страницах оглавления, а во многих случаях и полный текст выпусков. Национальные институты здоровья США (US National Institutes of Health, NIH) основали централизованную Web-библиотеку научных статей под названием PubMed Central (`http://www.pubmedcentral.nih.gov/`). NCBI в сотрудничестве с научными журналами организует систему электронного доступа к полным текстам опубликованных статей.

Обзоры баз данных и серверов по молекулярной биологии

Сложно заниматься изучением какой-либо области в молекулярной биологии и при этом не наткнуться на такого рода обзор. Списки Web-ресурсов по молекулярной биологии встречаются очень часто. По большей части они содержат одну и ту же информацию, хотя могут сильно различаться общему духу ее представления. Основная сложность заключается в том, что если никто не занимается их курированием, то они быстро деградируют в списки «мертвых» ссылок. (Черновой вариант этого раздела книги содержал ссылку на сайт с хорошим обзором ресурсов. По прошествии двух месяцев мы обнаружили, что сайт поменял свой адрес, и исчезло более половины сайтов, перечисленных в обзоре.)

Эта книга не содержит длинного списка с аннотациями важных и рекомендуемых нами сайтов по следующим причинам. (1) Вам не нужен длинный список — вам нужен короткий. (2) Интернет слишком непостоянен, чтобы такой список долго оставался пригодным. *Намного более эффективно использовать неспециализированную поисковую систему для того, чтобы найти то, что вы хотите, тогда, когда вы этого хотите.*

Мой вам совет: потратьте какое-то время на поиск в Интернете. Вам не потребуется слишком долго искать сайт, который бы представлялся достаточно стабильным и соответствовал методам вашей работы. В качестве альтернативы вот ссылка на достаточно всеобъемлющий сайт, который, насколько можно судить, прилагает усилия к тому, чтобы идти в ногу со временем: `http://www.expasy.\ch/alinks.html`. Это подходящее место для того, чтобы начать поиск.

Вход в архивы

Базы нуклеотидных и белковых последовательностей поддерживают возможность очень широкого круга операций по получению и анализу информации.

1. *Получение последовательностей из баз данных.* Последовательности могут быть найдены на основе либо их аннотаций, либо их характерных участков.
2. *Сравнение последовательностей.* Это не просто опция, а целая индустрия! Она впервые была упомянута в гл. 1 и подробно обсуждается в гл. 4. Сравнение последовательностей включает в себя решение такой важной задачи, как поиск родственных последовательностей.
3. *Трансляция последовательностей ДНК в аминокислотные последовательности.*
4. *Простые способы анализа и предсказания структуры.* Например, статистические методы предсказания вторичной структуры белков на основе только их последовательности, в том числе профили гидрофобности, позволяющие в целом находить трансмембранные α-спирали.
5. *Поиск мотивов (паттернов).* Можно провести поиск всех последовательностей, содержащих заданный мотив или комбинацию мотивов. Мотив представляет собой набор вероятностей обнаружить определенный набор остатков в следующих друг за другом позициях. В последовательностях ДНК это могут быть специальные участки — сайты узнавания ферментами, например, такими, как ответственные за сшивание поврежденных генов. В белках короткие и четко локализованные мотивы могут указывать на то, что молекулы выполняют сходную функцию, хотя никакого общего родства между их последовательностями нет. PROSITE — это коллекция записей таких белковых паттернов.
6. *Молекулярная графика.* Необходимо четкое и ясное компьютерное представление очень сложных систем. Возможности этих операций следующие:
   - расположение на трехмерном остове белка остатков, которые, как считают, необходимы для выполнения какой-либо функции. Часто это позволяет выделить активный центр белка в виде кластера боковых цепей в пространстве
   - классификация и сравнение характера фолдинга (способа укладки) белков

- анализ различий между близкородственными структурами, либо между двумя конформациями одной молекулы
- изучение взаимодействия с белком малых молекул для определения функции белка или при разработке лекарств
- интерактивное сопоставление модели белка с «зашумленным» и нечетким изображением молекулы, которое получается при расшифровке структуры белка методом рентгеноструктурной кристаллографии
- разработка и моделирование новых структур.

Доступ к базам данных в молекулярной биологии

Как приобрести навык работы в молекулярной биологии через Интернет?

Было бы сложно научиться ездить на велосипеде, читая книгу с описанием необходимых движений, хотя и проще, чем если бы это была книга про то, как устроен гироскоп. Точно так же место, где нужно учиться использовать Интернетом, — компьютер, через браузер. Это действительно так, но всегда присутствует какой-то начальный период затруднений и неуверенности. Наша цель — предоставить вам только временную поддержку для того, чтобы вы могли с чего-то начать. Дальше — вам разбираться самим!

Этот раздел содержит введение в некоторые крупнейшие базы данных и системы поиска информации в молекулярной биологии. В каждом случае мы будем показывать сравнительно простые способы поиска и приложения. По мере необходимости будем подчеркивать уникальные черты каждой системы.

ENTREZ

Национальный центр информации по биотехнологиям (National Center for Biotechnology Information, NCBI) — составная часть Медицинской национальной библиотеки США — поддерживает базы данных и средства доступа к ним. ENTREZ позволяет работать со следующими базами данных:

- Белковые
- Полипептидные
- Нуклеотидные
- Базы данных структур
- Геномные
- База данных Popset, предоставляющая информацию о популяциях
- База данных OMIM — база данных признаков, наследуемых по Менделю (в основном это генетические болезни. — *Прим. ред.*).

Ссылки между различными базами данных — сильная сторона NCBI. Отправной точкой для поиска последовательностей и структур является Entrez: `http://www.ncbi.nlm.nih.gov/Entrez/`

Давайте найдем для молекулы человеческой эластазы нейтрофилов результаты поиска в различных секциях ENTREZ.

Поиск по базе данных белков ENTREZ

Зайдите на сайт `http://www.ncbi.nlm.nih.gov/Entrez/` Выберите ссылку «Protein: sequence database» («Белки: база данных по последовательностям»), в окошке для поиска наберите «HUMAN ELASTASE» и нажмите «Go». (При работе с Entrez и другими ресурсами надо иметь в виду, что данные очень быстро накапливаются, равно как изменяется дизайн сайта. Поэтому если воспроизводить приведенные здесь примеры, можно получить результаты, отличные от тех, что приведены здесь. Не удивляйтесь — это нормально. — *Прим. ред.*).

Программа выдаст 390 ответов (на с. 163 показаны первые 15). Самый верхний ответ — «предшественник эластазы 1 (ELASTASE 1 PRECURSOR) [*HOMO SAPIENS*]»; другие результаты поиска включают эластазы других видов, ингибиторы человека и пиявки и тирозил-тРНК-синтетазу. (Почему такие белки, как белок пиявки и тРНК синтетаза, должны обнаруживаться при поиске человеческой эластазы (см. Интернет-задание 3.9)? Позже мы рассмотрим, как настроить запрос, чтобы избежать таких ненужных ответов.

Формат ответа следующий. В каждом случае в первой строке указывается имя, синонимичные названия молекулы и ее видовая принадлежность. Не забывайте, что греческие буквы не употребляются. Следующая строка дает представления о типе базы данных: gi = GenInfo Indentifer, (см. с. 40), gb = GeneBank accession number, sp = Swiss-Prot, pir = Protein Idenification Resource, ref = проект NCBI (справочник по последовательностям). В ответе также присутствуют эластазы человека и других видов, а также ингибиторы эластаз.

Открывая первую ссылку, получим картину, сходную с иллюстрацией (с. 164). Первые строчки содержат идентификатор белка в базе данных, название молекулы, дату занесения в базу и т. д. Далее следует более подробное описание белка — организм, из которого он был выделен (в нашем случае человек) с полной таксономией; авторы, которые впервые добавили (описали) этот белок; ссылки на литературу. И в конце представлены научная информация — локализация гена и его продукты (CDS = кодирующая последовательность), и сама последовательность (см. упр. 3.2).

Поиск в банке данных нуклеотидных последовательностей ENTREZ

Сейчас мы снова поищем эластазу человека, human elastase, но на этот раз в банке данных нуклеотидных последовательностей. Давайте попытаемся настроить поиск так, чтобы исключить находки, относящиеся не к эластазе, а к ее ингибиторам.

1. Выберем NUCLEOTIDE на сайте Entrez.
2. Щелкнем мышкой на закладку INDEX, выберем ORGANISM из выпадающего списка ALL FIELDS, введем название организма — Homo sapiens — в поле справа и затем щелкнем по кнопке AND.

Первые 15 результатов поиска «human elastase» в банке данных белковых последовательностей ENTREZ

1. elastase 1 precursor [Homo sapiens]
   gi—4731318—gb—AAD28441.1—AF120493_1[4731318
2. ALPHA-1-ANTITRYPSIN PRECURSOR
   (ALPHA-1 PROTEASE INHIBITOR) (ALPHA-1-ANTIPROTEINASE)
   gi—1703025—sp—P01009—A1AT_HUMAN[1703025]
3. elastase [Mus musculus]
   gi—7657060—ref—NP_056594.1—[7657060]
4. proteinase 3 [Mus musculus]
   gi—6755184—ref—NP_035308.1—[6755184]
5. ANTIMICROBIAL PEPTIDE ENAP-2
   gi—7674025—sp—P56928—ENA2_HORSE[7674025]
6. AMBP PROTEIN PRECURSOR [CONTAINS: ALPHA-1-MICROGLOBULIN (PROTEIN HC) (COMPLEX-FORMING GLYCOPROTEIN HETEROGENEOUS IN CHARGE); INTER-ALPHA-TRYPSIN INHIBITOR LIGHT CHAIN (ITI-LC) (BIKUNIN) (HI-30)]
   gi—122801—sp—P02760—AMBP_HUMAN[122801]
7. ELAFIN PRECURSOR (ELASTASE-SPECIFIC INHIBITOR) (ESI) (SKIN-DERIVED ANTILEUKOPROTEINASE) (SKALP)
   gi—119262—sp—P19957—ELAF_HUMAN[119262]
8. ANTILEUKOPROTEINASE
   gi—113637—sp—P22298—ALK1_PIG[113637]
9. ANTILEUKOPROTEINASE 1 PRECURSOR (ALP) (HUSI-1) (SEMINAL PROTEINASE INHIBITOR) (SECRETORY LEUKOCYTE PROTEASE INHIBITOR) (BLPI) (MUCUS PROTEINASE INHIBITOR) (MPI)
   gi—113636—sp—P03973—ALK1_HUMAN[113636]
10. ALPHA-2-MACROGLOBULIN PRECURSOR (ALPHA-2-M)
    gi—112911—sp—P01023—A2MG_HUMAN[112911]
11. tyrosyl-tRNA synthetase [Homo sapiens]
    gi—4507947—ref—NP_003671.1—[4507947]
12. pancreatic elastase IIB [Homo sapiens]
    gi—7705648—ref—NP_056933.1—[7705648]
13. protease inhibitor 3, skin-derived (SKALP) [Homo sapiens]
    gi—4505787—ref—NP_002629.1—[4505787]
14. pancreatic elastase I (allele HEL1-36) - human (fragment)
    gi—7513237—pir——S70441[7513237]
15. m guamerin - Korean leech

3. Затем выберем SUBSTANCE NAME из того же выпадающего меню ALL FIELDS, введем ELASTASE и опять нажмем AND.
4. Наконец, в выпадающем меню «ALL FIELDS» выбираем «TEXT WORD», вводим слово «INHIBITOR» и щелкаем по кнопке с надписью «NOT».

Первый результат поиска эластазы человека в банке данных белковых последовательностей ENTREZ

```
LOCUS       AF120493_1    258 aa                    PRI       03-AUG-2000
DEFINITION  elastase 1 precursor [Homo sapiens].
ACCESSION   AAD28441
PID         g4731318
VERSION     AAD28441.1  GI:4731318
DBSOURCE    locus AF120493 accession AF120493.1
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 258)
  AUTHORS   Talas,U., Dunlop,J., Khalaf,S., Leigh,I.M. and Kelsell,D.P.
  TITLE     Human elastase 1: evidence for expression in the skin and the
            identification of a frequent frameshift polymorphism
  JOURNAL   J. Invest. Dermatol. 114 (1), 165-170 (2000)
  MEDLINE   20087075
   PUBMED   10620133
REFERENCE   2  (residues 1 to 258)
  AUTHORS   Talas,U., Dunlop,J., Leigh,I.M. and Kelsell,D.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-JAN-1999) Centre for Cutaneous Research, Queen Mary
            and Westfield College, 2 Newark Street, London E1 2AT, UK
COMMENT     Method: conceptual translation supplied by author.
FEATURES             Location/Qualifiers
     source          1..258
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="12"
                     /map="12q13"
                     /cell_type="keratinocyte"
     Protein         1..258
                     /product="elastase 1 precursor"
     CDS             1..258
                     /gene="ELA1"
                     /coded_by="AF120493.1:42..818"
ORIGIN
        1 mlvlyghstq dlpetnarvv ggteagrnsw psqislqyrs ggsryhtcgg tlirqnwvmt
       61 aahcvdyqkt frvvagdhnl sqndgteqyv svqkivvhpy wnsdnvaagy diallrlaqs
      121 vtlnsyvqlg vlpqegaila nnspcyitgw gktktngqla qtlqqaylps vdyaicssss
      181 ywgstvkntm vcaggdgvrs gcqgdsggpl hclvngkysl hgvtsfvsar gcnvsrkptv
      241 ftqvsayisw innviasn
//
```

В результате действий 1–4 составленный запрос будет выглядеть несколько иначе:

HOMO SAPIENS[ORGANISM] AND ELASTASE[TEXT WORD] NOT INHIBITOR[TEXT WORD]

Первый результат поиска эластазы человека в банке данных нуклеотидных последовательностей ENTREZ

```
LOCUS       AF120493      952 bp    mRNA            PRI       03-AUG-2000
DEFINITION  Homo sapiens elastase 1 precursor (ELA1) mRNA, complete cds.
ACCESSION   AF120493
VERSION     AF120493.1  GI:4731317
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 952)
  AUTHORS   Talas,U., Dunlop,J., Khalaf,S., Leigh,I.M. and Kelsell,D.P.
  TITLE     Human elastase 1: evidence for expression in the skin and the
            identification of a frequent frameshift polymorphism
  JOURNAL   J. Invest. Dermatol. 114 (1), 165-170 (2000)
  MEDLINE   20087075
   PUBMED   10620133
REFERENCE   2  (bases 1 to 952)
  AUTHORS   Talas,U., Dunlop,J., Leigh,I.M. and Kelsell,D.P.
  TITLE     Direct Submission
  JOURNAL   Submitted (15-JAN-1999) Centre for Cutaneous Research, Queen Mary
            and Westfield College, 2 Newark Street, London E1 2AT, UK
FEATURES             Location/Qualifiers
     source          1..952
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="12"
                     /map="12q13"
                     /cell_type="keratinocyte"
     gene            1..952
                     /gene="ELA1"
     CDS             42..818
                     /gene="ELA1"
                     /codon_start=1
                     /product="elastase 1 precursor"
                     /protein_id="AAD28441.1"
                     /db_xref="GI:4731318"
                     /translation="MLVLYGHSTQDLPETNARVVGGTEAGRNSWPSQISLQYRSGGSR
                     YHTCGGTLIRQNWVMTAAHCVDYQKTFRVVAGDHNLSQNDGTEQYVSVQKIVVHPYWN
                     SDNVAAGYDIALLRLAQSVTLNSYVQLGVLPQEGAILANNSPCYITGWGKTKTNGQLA
                     QTLQQAYLPSVDYAICSSSSYWGSTVKNTMVCAGGDGVRSGCQGDSGGPLHCLVNGKY
                     SLHGVTSFVSSRGCNVSRKPTVFTQVSAYISWINNVIASN"
BASE COUNT      226 a    261 c    250 g    215 t
ORIGIN
        1 ttggtccaag caagaaggca gtggtctact ccatcggcaa catgctggtc ctttatggac
       61 acagcaccca ggaccttccg gaaaccaatg cccgcgtagt cggagggact gaggccggga
      121 ggaattcctg gccctctcag atttccctcc agtaccggtc tggaggttcc cggtatcaca
      181 cctgtggagg gacccttatc agacagaact gggtgatgac agctgctcac tgcgtggatt
      241 accagaagac tttccgcgtg gtggctggag accataacct gagccagaat gatggcactg
      301 agcagtacgt gagtgtgcag aagatcgtgg tgcatccata ctggaacagc gataacgtgg
      361 ctgccggcta tgacatcgcc ctgctgcgcc tggcccagag cgttaccctc aatagctatg
      421 tccagctggg tgttctgccc caggagggag ccatcctggc taacaacagt ccctgctaca
      481 tcacaggctg gggcaagacc aagaccaatg ggcagctggc ccagaccctg cagcaggctt
      541 acctgccctc tgtggactat gccatctgct ccagctcctc ctactggggc tccactgtga
      601 agaacaccat ggtgtgtgct ggtggagatg gagttcgctc tggatgccag ggtgactctg
      661 ggggcccccc ccattgcttg gtgaatggca agtattctct ccatggagtg accagctttg
      721 tgtccagccg gggctgtaat gtctccagga agcctacagt cttcacccag gtctctgctt
      781 acatctcctg gataaataat gtcatcgcct ccaactgaac attttcctga gtccaacgac
      841 cttcccaaaa tggttcttag atctgcaata ggacttgcga tcaaaaagta aaacacattc
      901 tgaaagacta ttgagccatt gatagaaaag caaataaaac tagatataca tt
//
```

Результат такого поиска — 445 находок, в том числе множество клонов одного и того же гена. Первая запись в списке находок такова: Homo SAPIENS ELASTASE 1 PRECURSOR (ELA1) mRNA, complete cds. Термин «complete cds» обозначает полную кодирующую последовательность (complete coding sequence).

Сравните результат этого поиска с результатом, полученным при поиске по банку данных белковых последовательностей (см. упр. 3.5).

Поиск в банке данных геномов ENTREZ

Поиск для эластазы человека HUMAN ELASTASE дает следующий результат:

1. NC_000967 CAENORHABDITIS ELEGANS CHROMOSOME III[64] LCL-WORM_CHR_III
2. NC_001099 HOMO SAPIENS CHROMOSOME 19[19] REF-NC_001099-HSAP-19
3. NC_001065 HOMO SAPIENS CHROMOSOME 14[14] REF-NC_001065-HSAP-14
4. NC_001044 HOMO SAPIENS CHROMOSOME 11[11] REF-NC_001044-HSAP-11
5. NC_001008 HOMO SAPIENS CHROMOSOME 6[6] REF-NC_001008-HSAP-6

Почему в поиске для эластазы человека в результатах оказываются белки *C. elegans*? Запись NC_000967 — это хромосома III *C. elegans* целиком. В комментариях к одному из детектированных генов содержится:

```
gene="T07A5.1" /note="weak similarity with elastase
(PIR accession number A406659)"
```

т. е. обнаружено слабое сходство с эластазами.

Многие другие гены *C. elegans* аннотированы по сходству с белками человека. *C. elegans* действительно содержит эластазу, однако она *не* похожа на эластазу человека, хотя они и гомологи.

Поиск в банке данных структур ENTREZ

Известна ли пространственная структура эластазы человека? Выберем банк данных STRUCTURE и повторим запрос. Программа возвращает два результата:

1B0F Crystal Structure Of Human Neutrophil Elastase With Mdl 101, 146

1QIX Porcine Pancreatic Elastase Complexed With Human Beta-Casomorphin-7

Обозначения 1B0F и 1QIX — это названия записей в банке данных белковых последовательностей.

Но постойте, мы могли не заметить, что потеряли множество полезных результатов! Есть много структур эластазы, выполненных в комплексе с ингибиторами, которые мы просили программу не учитывать. Удалив фразу NOT INHIBITORS и отправив запрос повторно, мы получим в качестве результата 8 структур:

```
1B0F  Crystal Structure Of Human Neutrophil Elastase With Mdl 101, 146
1QIX  Porcine Pancreatic Elastase Complexed With Human Beta-Casomorphin-7
2REL  Solution Structure Of R-Elafin, A Specific Inhibitor Of Elastase,
      Nmr, 11 Structures
1FLE  Crystal Structure Of Elafin Complexed With Porcine Pancreatic Elastase
1FUJ  Pr3 (Myeloblastin)
1PPG  Human Leukocyte Elastase (Hle) (E.C.3.4.21.37)
      Complex With Meo-Succinyl-Ala-Ala-Pro-Val Chloromethylacetone
1PPF  Human Leukocyte Elastase (Hle) (neutrophil Elastase (Hne)) (E.C.3.4.21.37)
      Complex With The Third Domain Of Turkey Ovomucoid Inhibitor (O mtky3)
1HNE  Human Neutrophil Elastase (HNE) (E.C.3.4.21.37)
      (Also Referred To As Human Leucocyte Elastase (HLE))
      Complex With Methoxysuccinyl-Ala-Ala-Pro-Ala Chloromethyl Ketone
      (MSACK)
```

Однако, если мы сделаем запрос PDB для эластазы (см. дальше), мы найдем (см. упр. 3.2):

```
0EPC   Elastase -(Thr-Pro-Nval-Nmeleu-Tyr-Thr) Co...
0ESC   Elastase with Two Molecules Of Acetyl-Ala-...
0ESZ   Elastase-N-Carbobenzoxy-L-Alanyl-P-Nitroph...
1B0E   Porcine Pancreatic Elastase With Mdl 101 146
1B0F   Human Neutrophil Elastase With Mdl 101 146
1BMA   Benzyl Methyl Aminimide Inhibitor Complexe...
1BRU   Porcine Pancreatic Elastase with The Elast...
1BTU   Porcine Pancreatic Elastase with (3s 4r)-1...
1C1M   Porcine Elastase Under Xe Pressure (8 Bar)
1DKG   The Nucleotide Exchange Factor Grpe Bound ...
1EAI   Complex Of Ascaris Chymotrpsin Elastase In...
1EAS   Elastase with 3- (Methylamino) Sulfonyl Am...
1EAT   Elastase with 2- 5-Methanesulfonylamino-2-...
1EAU   Elastase with 2- 5-Amino-6-Oxo-2-(2-Thieny...
1ELA   Elastase with Trifluoroacetyl-L-Lysyl-L-Pr...
1ELB   Elastase with Trifluoroacetyl-L- Lysyl-L-L...
```

```
1ELC   Elastase with Trifluoroacetyl-L-Phenylalan...
1ELD   Elastase with Trifluoroacetyl-L- Phenylala...
1ELE   Elastase with Trifluoroacetyl-L- Valyl-L-A...
1ELF   Elastase with N-(Tert- Butoxycarbonyl-Alan...
1ELG   Elastase with N-(Tert- Butoxycarbonyl-Alan...
1ELT   Native Pancreatic Elastase From North Atla...
1ESA   Elastase Low Temperature Form (-45 C)
1ESB   Elastase with N-Carbobenzoxy-L-Alanyl-P-Ni...
1EST   Tosyl-Elastase
1EZM   Elastase (Zn Metalloprotease)
1FLE   Elafin with Porcine Pancreatic Elastase
1HLE   Horse Leukocyte Elastase Inhibitor (Hlei)
1HNE   Human Neutrophil Elastase ( Hne) (Also Ref...
1INC   Porcine Pancreatic Elastase with Benzoxazi...
1JIM   Porcine Pancreatic Elastase with The Heter...
1LVY   Porcine Elastase
1NES   Structure: Product Complex Of Acetyl-Ala-P...
1PPF   Human Leukocyte Elastase (Hle) (Neutrophil...
1PPG   Human Leukocyte Elastase (Hle) with Meo-Su...
1QGF   Porcine Pancreatic Elastase with (3r 4s)N-...
1QIX   Porcine Pancreatic Elastase with Human Bet...
1QNJ   The Native Porcine Pancreatic Elastase At ...
1QR3   Porcine Pancreatic Elastase with Fr901277 ...
2EST   Elastase with Trifluoroacetyl -L-Lysyl-L-A...
2REL   Solution R-Elafin A Specific Inhibitor Of ...
3EST   Native Elastase
4EST   Porcine Pancreatic Elastase with Ace-Ala-P...
5EST   Porcine Pancreatic Elastase with Carbobenz...
6EST   Elastase Crystallized 10% Dmf
7EST   Elastase with Trifluoroacetyl -L-Leucyl-L-...
8EST   Porcine Pancreatic Elastase with Guanidini...
9EST   Porcine Pancreatic Elastase with Guanidini...
```

Поиск по библиографической базе данных PubMed

Вероятно, пришло время посмотреть статьи о нашей молекуле. Разумеется, существует огромное количество литературы по эластазе. Поиск HUMAN ELASTASE в базе данных PubMed выдает 6506 статей. Давайте попробуем найти цитаты из статей, описывающих роль эластазы в заболевании, чтобы сократить полученные результаты. Поиск HUMAN ELASTASE DISEASE выдает 1214 статей. Что же насчет специфических мутантов эластазы, связанных с заболеваниями человека? Поиск HUMAN ELASTASE DISEASE MUTATION выдает 28 статей в обратном хронологическом порядке. Ниже приведены первые 10.

1. Hermans MH, Touw IP. Significance of neutrophil elastase mutations versus G-CSF receptor mutations for leukemic progression of congenital neutropenia. Blood. 2001 Apr 1;97(7):2185–6. No abstract available.

2. Li FQ, Horwitz M. Characterization of mutant neutrophil elastase in severe congenital neu- tropenia. J Biol Chem. 2001 Apr 27;276(17):14230–41.
3. Ye S. Polymorphism in matrix metalloproteinase gene promoters: implication in regulation of gene expression and susceptibility of various diseases. Matrix Biol. 2000 Dec;19(7):623–9. Review.
4. Dale DC, Person RE, Bolyard AA, Aprikyan AG, Bos C, Bonilla MA, Boxer LA, Kannourakis G, Zeidler C, Welte K, Benson KF, Horwitz M. Mutations in the gene encoding neutrophil elastase in congenital and cyclic neutropenia. Blood. 2000 Oct 1;96(7):2317–22.
5. McGettrick AJ, Knott V, Willis A, Handford PA. Molecular effects of calcium binding mutations in Marfan syndrome depend on domain context. Hum Mol Genet. 2000 Aug 12;9(13):1987–94.
6. Rashid MH, Rumbaugh K, Free in PMC, Passador L, Davies DG, Hamood AN, Iglewski BH, Kornberg A. Polyphosphate kinase is essential for biofilm development, quorum sensing, and virulence of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A. 2000 Aug 15;97(17):9636–41.
7. Jormsjo S, Ye S, Moritz J, Walter DH, Dimmeler S, Zeiher AM, Henney A, Hamsten A, Eriksson P. Allele-specific regulation of matrix metalloproteinase-12 gene activity is associated with coronary artery luminal dimensions in diabetic patients with manifest coronary artery disease. Circ Res. 2000 May 12;86(9):998-1003.
8. Talas U, Dunlop J, Khalaf S, Leigh IM, Kelsell DP. Human elastase 1: evidence for expression in the skin and the identification of a frequent frameshift polymorphism. J Invest Dermatol. 2000 Jan;114(1):165–70.
9. Horwitz M, Benson KF, Person RE, Aprikyan AG, Dale DC. Mutations in ELA2, encoding neutrophil elastase, define a 21-day biological clock in cyclic haematopoiesis. Nat Genet. 1999 Dec;23(4):433–6.
10. Griffin MD, Torres VE, Grande JP, Kumar R. Vascular expression of polycystin. J Am Soc Nephrol. 1997 Apr;8(4):616–26.

Это ссылки на связь мутаций в эластазе нейтрофилов с нейтропенией — низким уровнем одного типа лейкоцитов, называемых нейтрофилами. В продолжение этого мы можем поискать эластазу в базе данных генетических заболеваний человека.

Интерактивный каталог «Менделевская (по Менделю) наследственность человека» (OMIM)

OMIM — база данных человеческих генов и генетических нарушений. Первоначально она была составлена МакКасиком (V. A. McKusick), Смитом (M. Smith) и их коллегами и опубликована в письменной форме. NCBI усовершенствовал ее, сделав базой данных, доступной в сети Интернет, и вставил ссылки на другие архивы со сходной информацией, включая базы данных последовательностей и медицинскую литературу. Сейчас OMIM хорошо интегрирована с информацией поисковой системы ENTREZ NCBI. Родственная база данных Morbid Map, карта болезней OMIM, имеет дело с генетическими болезнями и их хромосомной локализацией.

Ниже следует несколько выдержек из ответа на слово ELASTASE при поиске в OMIM.

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM
Search for
Limits Preview/Index History Clipboard

***130130** Related Entries, PubMed, Protein, Nucleotide, Structure, Genome, LinkOut

ELASTASE 2; ELA2

Alternative titles; symbols

ELASTASE, NEUTROPHIL; NE
ELASTASE, LEUKOCYTE
MEDULLASIN
PROTEASE, SERINE, BONE MARROW

Gene map locus 19p13.3

TEXT

Aoki (1978) purified a 31,800-Da serine protease from human bone marrow cell mitochondria. Both granulocytes and erythroblasts were found to contain the protease medullasin, but it was not detected in lymphocytes or thrombocytes. It was shown to be located on the inner membrane of mitochondria. Nakamura et al. (1987) reported the complete genomic sequence and deduced the amino acid sequence of the medullasin precursor. It contains 267 amino acids, including a possible leader sequence of 29 amino acids.

■■■

Cyclic hematopoiesis (cyclic neutropenia; 162800) is an autosomal dominant disorder in which blood-cell production from the bone marrow oscillates with 21-day periodicity. Circulating neutrophils vary between almost normal numbers and zero. During intervals of neutropenia, affected individuals are at risk for opportunistic infection. Monocytes, platelets, lymphocytes, and reticulocytes also cycle with the same frequency. Horwitz et al. (1999) used a genomewide screen and positional cloning to map the locus to 19p13.3. They identified 7 different single-basepair substitutions in the ELA2 gene, each on a unique haplotype, in 13 of 13 families, as well as a new mutation in a sporadic case. Neutrophil elastase is a target for protease inhibition by alpha-1-antitrypsin (also called protease inhibitor-1; PI; 107400), and its unopposed release destroys tissue at sites of inflammation. Horwitz et al. (1999) hypothesized that a perturbed interaction between neutrophil elastase and serpins or other substrates may regulate mechanisms governing the clock-like timing of hematopoiesis.

■■■

Набор результатов по эластазе, составленный нами, мог бы поддержать исследование этой системы; например, мы можем нанести мутанты эластазы на структуру молекулы, чтобы установить возможности выяснения причины циклической нейтропении.

Система поиска последовательностей (Sequence Retrieval System, SRS)

Система SRS, первоначально разработанная Этцольдом (T. Etzold), является интегрированной системой, производящей поиск информации во многих различных базах данных последовательностей и обработку найденных последова-

тельностей аналитическими инструментами — сравнение последовательностей и программы, выполняющие выравнивания.

В общей сложности, SRS может производить поиск по 141 базе данных белковых и нуклеотидных последовательностей, метаболических путей, пространственных структур и функций, геномов и болезней и по фенотипической информации. SRS включает много маленьких баз данных таких, как the Prosite и Blocks — базы данных структурных мотивов белков, базы данных транскрипционных факторов и базы данных, специализированные на определенных патогенах.

Категории баз данных, доступных для поиска SRS

Последовательности	Геномы
Базы, связанные с InterPro	Картирование
Базы, связанные с последовательностями	Мутации
Факторы транскрипции TransFac	СНП
Пользовательские базы данных	Локус-специфические мутации
Результаты обработки	Метаболические пути
Трехмерные структуры белков	Паттерны последовательностей

Внутри категории Sequence у SRS есть доступ к следующим базам данных:

EMBL	архивная база данных нуклеотидных последовательностей
EMBLNEW	обновления к последнему полному релизу EMBL
ENSEMBL	аннотированные геномные последовательности
SWISSPROT	курируемая и аннотируемая архивная база данных белковых последовательностей
SPTREMBL	аннотируемая компьютером база данных белковых последовательностей, дополняющая базу данных белковых последовательностей SWISSPROT
REMTREMBL	трансляции кодирующих последовательностей из базы данных нуклеотидных последовательностей EMBL, которые не были предназначены для окончательного введения в SWISS-PROT
TREMBLNEW	трансляции всех новых и обновленных кодирующих последовательностей в EMBL начиная с последнего релиза TrEMBL
SWALL	подробная база данных белковых последовательностей, соединяющая полные аннотации в SWISSPROT с еженедельно обновляющимися трансляциями всех кодирующих последовательностей белков из базы данных нуклеотидных последовательностей EMBL
IMGT	интегрированная база данных, специализированная на иммуноглобулинах, рецепторах Т-клеток и главном комплексе гистосовместимости (ГКГС) всех видов позвоночных
IMGTHLA	белковые последовательности главного комплекса гистосовместимости человека
InterPro	интегрированный ресурс белковых доменов и функциональных сайтов (документационный ресурс белковых семейств, доменов и функциональных сайтов)

Кроме предоставления доступа к множеству баз данных, SRS также предоставляет систему ссылок, связывающих эти базы, и плавный переход между

ними в запускаемых приложениях. Поиск в одной отдельно взятой БД может быть расширен до поиска во всей сети; таким образом, можно легко найти все относящиеся к данному белку записи во всех БД. Поиски похожих последовательностей, выравнивания и т. д. могут быть запущены сразу же, без сохранения результатов поиска в промежуточных файлах на компьютере пользователя.

В сеансе работы с SRS вы начинаете с выбора одной или более БД, где будет производиться поиск. Они сгруппированы по категориям: нуклеотидные, белковые и т. д. Далее вы можете указать критерии запроса. Как и в случае с ENTREZ, можно либо искать их сразу по всем полям записей, либо указать, в какой категории терминов искать. Программа в ответ выдаст список записей, отвечающих запросу. После нахождения можно:

1. посмотреть одну из найденных последовательностей, переходя к полученному файлу,
2. выделить одну (или более) определенную последовательность и провести поиск похожих (или ассоциированных с ней) последовательностей в других БД,
3. запустить приложение, такое как предсказание вторичной структуры, или множественное выравнивание.

Прочие опции на странице представления результатов поиска позволяют вам создавать и сохранять в файле отчеты по выделенным записям. Это может быть просто список найденных последовательностей или же результат более сложного анализа результатов. Например, белковые БД предоставляют график гидрофобности.

Главная страница SRS: `http://srs.ebi.ac.uk/`, также по адресу `http://www.lionbioscience.ac.psiweb.com/publicsrs.html` находится список зеркальных сайтов SRS. Для поиска эластазы человека, откройте SRS и выберите Swiss-Prot. Введите HUMAN ELASTASE в поле Simple query и нажмите кнопку Quick Search. Программа выдаст результат:

RootLibs	acc	des	sl
SWISSPROT:EL1_HUMAN	P11423	ELASTASE 1 (EC 3.4.21.36) (FRAGMENT).	68
SWISSPROT:EL2A_HUMAN	P08217	ELASTASE 2A PRECURSOR (EC 3.4.21.71)	269
SWISSPROT:EL2B_HUMAN	P08218	ELASTASE 2B PRECURSOR (EC 3.4.21.71).	269
SWISSPROT:EL3A_HUMAN	P09093	ELASTASE IIIA PRECURSOR (EC 3.4.21.70) (PROTEASE E)	270
SWISSPROT:EL3B_HUMAN	P08861	ELASTASE IIIB PRECURSOR (EC 3.4.21.70) (PROTEASE E)	270
SWISSPROT:ELNE_HUMAN	P08246	LEUKOCYTE ELASTASE PRECURSOR (EC 3.4.21.37) (NEUTROPHIL ELASTASE)	267
	P09649	(PMN ELASTASE) (BONE MARROW SERINE PROTEASE) (MEDULLASIN).	267
SWISSPROT:ILEU_HUMAN	P30740	LEUKOCYTE ELASTASE INHIBITOR (LEI) (MONOCYTE/NEUTROPHIL ELASTASE INHIBITOR) (M/NEI) (EI).	379
SWISSPROT:ELAF_HUMAN	P19957	ELAFIN PRECURSOR (ELASTASE-SPECIFIC INHIBITOR) (ESI)	117
		(SKIN-DERIVED ANTILEUKOPROTEINASE) (SKALP).	117

Для демонстрации запуска приложений выберем последовательности эластаз млекопитающих и произведем множественное выравнивание с помощью CLUSTAL-W. В сеансе работы с SRS выберите Swiss-Prot и далее кликните на Extended в блоке Query. На полученной странице введите Mammalia в поле Organism, а в поле Description введите elastase ! inhibitor ! fragment. (Восклицательный знак является альтернативным обозначением NOT, т. е. отрицанием. Данный запрос означает выбор записей, в которых описание содержит «elastase» и не содержит слов «inhibitor» и «fragment», поскольку нам нужны полные молекулы эластаз, а фрагменты и ингибиторы не требуются.) Далее кликните на Submit query.

Программа выдаст примерно 20 ответов. Найдите выпадающее меню под надписью Launch, выберите в нем Clustal-W и кликните на Launch. Программа покажет вам созданные ею предварительные входные данные для Clustal-W; это дает вам возможность изменить параметры от взятых по умолчанию, либо прервать процесс, если вы по какой-либо причине не удовлетворены ситуацией. Начните процесс выравнивания повторным нажатием на Launch, и ждите. Результаты показаны на цветной иллюстрации VI.

Ресурс идентификации протеинов (Protein Identification Resource, PIR)

PIR является эффективной комбинацией аккуратно курируемой базы данных, системы доступа к получению информации и рабочего окружения для исследования последовательностей. PIR также создает интегрированную среду анализа последовательностей (Integrated Environment for Sequence Analisys, IESA). Рассматривайте эту среду как надстройку над системой поиска информации. Ее функциональность включает в себя просмотр, поиск по БД и анализ сходства, а также ссылки на другие БД. Пользователь может:

- просматривать записи по аннотациям;
- искать по разным полям аннотаций, таким как Superfamily (суперсемейство), Family (семейство), Title (название), Species (вид), Taxonomy group (таксономия), Keywords (ключевые слова) и Domains (доменная структура);
- анализировать последовательности с помощью BLAST/FASTA-поисков, соответствий паттерну, множественных выравниваний;
- искать глобально или по доменам, а также сортируя по аннотации;
- просматривать статистику по суперсемействам, семействам, видам, таксономии, ключевым словам, доменам и прочим особенностям;
- получать ссылки на другие БД, включая PDB, COG, KEGG, WIT и BRENDA;
- искать последовательности в специализированных группах, таких как геномы человека, мыши, дрожжей.

Адреса для поиска в PIR:

в США:

```
http://www-nbrf.georgetown.edu/pirwww/search/textpsd.html
```

в Европе (зеркало):

```
http://www.mips.gsf.de
```

Поиск эластазы человека в PIR дал 15 записей, которые могут рассматриваться в качестве кандидатов:

Результаты поиска по PIR эластазы человека

ELHUL	leukocyte elastase (EC 3.4.21.37) precursor — human
TIHUSP	antileukoproteinase 1 precursor — human
ITHU	alpha-1-antitrypsin precursor — human
S70439	pancreatic elastase I (allele HEL1-16) probable splice form I — human
S68826	pancreatic elastase (EC 3.4.21.36) isoform 2 precursor — human
S68825	pancreatic elastase (EC 3.4.21.36) isoform 1 precursor — human
A29934	pancreatic elastase (EC 3.4.21.36) IIIA precursor — human
B26823	pancreatic elastase II (EC 3.4.21.71) A precursor — human
C26823	pancreatic elastase II (EC 3.4.21.71) B precursor — human
B29934	pancreatic elastase (EC 3.4.21.36) IIIB precursor — human
A49499	metalloelastase HME (EC 3.4.24.-) — human
S27383	elastase inhibitor — human
JH0614	elafin precursor — human
S70441	pancreatic elastase I (allele HEL1-36) — human (fragment)
A56615	probable pancreatic elastase (EC 3.4.21.36) pseudogene — human

Приведем лучший результат поиска (верхняя строка):

```
ENTRY            ELHUL  #type complete
TITLE            leukocyte elastase (EC 3.4.21.37) precursor [validated] -
                 human
ALTERNATE_NAMES  inflammatory serine proteinase; medullasin; neutrophil
                 elastase
ORGANISM         #formal_name Homo sapiens #common_name man
   #cross-references taxon:9606
DATE             30-Jun-1990 #sequence_revision 30-Jun-1990 #text_change
                 08-Dec-2000
ACCESSIONS       A31976; S04954; S06241; A27064; S00631; A28370; A34570;
                 A05293; A25907; S14736
REFERENCE        A31976
   #authors      Takahashi, H.; Nukiwa, T.; Yoshimura, K.; Quick, C.D.;
                 States, D.J.; Holmes, M.D.; Whang-Peng, J.; Knutsen, T.;
                 Crystal, R.G.
   #journal      J. Biol. Chem. (1988) 263:14739-14747
   #title        Structure of the human neutrophil elastase gene.
   #cross-references MUID:89008342
   #accession    A31976
      ##molecule_type DNA
      ##residues 1-267 ##label TAK
      ##cross-references GB:M20203; GB:J04032; NID:g189147;
                 PIDN:AAA36359.1; PID:g386981
```

```
                              additional references deleted. . .

COMMENT         This is a lysosomal proteinase found in the azurophil
                granules of neutrophils.
COMMENT         This elastase cleaves preferentially bonds after Ala and
                Val. It is believed to be one of the major agents
                responsible for tissue destruction in emphysema and
                rheumatoid arthritis.
GENETICS
   #gene        GDB:ELA2
      ##cross-references GDB:118792; OMIM:130130
   #map_position 19p13.3-19p13.3
   #introns     23/1; 75/2; 122/3; 199/3
CLASSIFICATION  #superfamily trypsin; trypsin homology
KEYWORDS        emphysema; glycoprotein; hydrolase; leukocyte; lysosome;
                rheumatoid arthritis; serine proteinase
FEATURE
   1-27                 #domain signal sequence #status predicted #label
                        SIG\
   28-29                #domain propeptide #status predicted #label PRO+\
   30-247               #product leukocyte elastase #status experimental
                        #label MAT\
   30-242               #domain trypsin homology #label TRY\
   248-267              #domain carboxyl-terminal propeptide #status
                        predicted #label CTP\
   55-71,151-208,
   181-187,198-223      #disulfide_bonds #status experimental\
   70,117,202           #active_site His, Asp, Ser #status predicted\
   88                   #binding_site carbohydrate (Asn) (covalent)
                        #status predicted\
   124,173              #binding_site carbohydrate (Asn) (covalent)
                        #status experimental
SUMMARY         #length 267 #molecular_weight 28518

SEQUENCE
             5         10        15        20        25        30
    1 M T L G R R L A C L F L A C V L P A L L L G G T A L A S E I
   31 V G G R R A R P H A W P F M V S L Q L R G G H F C G A T L I
   61 A P N F V M S A A H C V A N V N V R A V R V V L G A H N L S
   91 R R E P T R Q V F A V Q R I F E N G Y D P V N L L N D I V I
  121 L Q L N G S A T I N A N V Q V A Q L P A Q G R R L G N G V Q
  151 C L A M G W G L L G R N R G I A S V L Q E L N V T V V T S L
  181 C R R S N V C T L V R G R Q A G V C F G D S G S P L V C N G
  211 L I H G I A S F V R G G C A S G L Y P D A F A P V A Q F V N
  241 W I D S I I Q R S E D N P C P H P R D P D P A S R T H

---------------------------------------

PDB structures most related to ELHUL:
     1PPFE (30-247) 100.0%; 1PPGE (30-247) 100.0%; 1HNEE (30-247) 99.5%
     1BOF (30-247) 99.1%

Enzyme Links for ELHUL:
     EC-IUBMB: EC 3.4.21.37
     KEGG: EC 3.4.21.37
     BRENDA: EC 3.4.21.37
     WIT: EC 3.4.21.37
     MetaCyc: EC 3.4.21.37

ALIGNMENTS containing ELHUL:
     FA2856 trypsin - 230.4 19.0
     MO1074 trypsin - 1093.0 16.0
Associated Alignments:
     DA1082 trypsin homology
     SA2887 trypsin superfamily 230.4

Link to iProClass (Superfamily classification and Alignment):
     iProClass Report for ELHUL at PIR.

---------------------------------------
```

Одна из возможностей PIR — поиск специфических пептидных участков. Посмотрев на выравнивание эластаз млекопитающих, на позициях 220–228 (цветная иллюстрация VI) мы наблюдаем консервативный мотив — большинство последовательностей содержат CNGDSGGPLN. Находясь в PIR, можем выбрать «Pattern/Peptide match» и поискать точные совпадения с этим мотивом.

```
1    ELRT2  pancreatic elastase II (EC 3.4.21.71)
     214 - 223       GVTSSCNGDSGGPLNCQASN
2    CPBOA3  procarboxypeptidase A complex compon
     183 - 192       DTRSGCNGDSGGPLNCPAAD
3    S68826  pancreatic elastase (EC 3.4.21.36) i
     212 - 221       GVISACNGDSGGPLNCQLEN
4    S68825  pancreatic elastase (EC 3.4.21.36) i
     212 - 221       GVISACNGDSGGPLNCQLEN
5    A29934  pancreatic elastase (EC 3.4.21.36) I
     213 - 222       YIRSGCNGDSGGPLNCPTED
6    B26823  pancreatic elastase II (EC 3.4.21.71
     212 - 221       GVISSCNGDSGGPLNCQASD
7    C26823  pancreatic elastase II (EC 3.4.21.71
     212 - 221       GVICTCNGDSGGPLNCQASD
8    A26823  pancreatic elastase II (EC 3.4.21.71
     212 - 221       GIISSCNGDSGGPLNCQGAN
9    A25528  pancreatic elastase II (EC 3.4.21.71
     214 - 223       GVTSSCNGDSGGPLNCRASN
10   JQ1473  pancreatic elastase (EC 3.4.21.36) I
     212 - 221       GVISACNGDSGGPLNCQAED
11   B29934  pancreatic elastase (EC 3.4.21.36) I
     213 - 222       DIRSGCNGDSGGPLNCPTED
12   S29239  chymotrypsin (EC 3.4.21.1) 1 precurs
     219 - 228       GGKSTCNGDSGGPLNLNGMT
13   T10495  chymotrypsin (EC 3.4.21.1) BII - pen
     214 - 223       GGKGTCNGDSGGPLNLNGMT
```

Учтите, что названия белков обрезаны; это может привести к вводящим в заблуждение ситуациям, особенно если вы попытаетесь анализировать эти данные с помощью компьютерной программы, где обычно сложно увидеть очевидное. Например, может оказаться, что идентичная последовательность из 10 аминокислотных остатков имеется в карбоксипептидазе — молекуле, не имеющей никакого отношения к эластазе. Однако запись CPBOA3, вторая в списке, действительно является компонентом III комплекса прокарбоксипептидазы A коровы, гомологом эластазы.

Возвращаясь к выравниванию (цветная иллюстрация VI), в некоторых последовательностях видим вариации в паттерне. Более общий поиск по образцу «C[RNQF]GDSG[GS]PL[HNV]» (где [XYZ] означает позицию, способную содержать либо X, либо Y, либо Z) может выдать все эластазы млекопитающих, присутствующие в выравнивании, плюс другие последовательности, общим числом 82. Но даже это будут не все гомологи эластаз в банке данных, что

можно обнаружить после запуска PSI-BLAST, или, оставаясь в PIR, проведя поиск эластаз в базе данных PROT-FAM. Этот паттерн соответствует 20 семействам, все — сериновые протеиназы. Мы явно находимся на пути к получению полного списка гомологов.

ExPASy — экспертная система анализа белков.

ExPASy (Expert Protein Analysis System) — система поиска и обработки данных Швейцарского института биоинформатики. Ее производными (созданными при сотрудничестве с Европейским институтом биоинформатики) являются банки белковых последовательностей SWISS-PROT и TrEMBL. В TrEMBL хранятся «оттранслированные» аминокислотные последовательности, соответствующие нуклеотидным последовательностям, хранящимся в EMBL, но еще не в полном объеме добавленные в SWISS-PROT. (Отметим, что SWISS-PROT — курируемая база данных, и для того чтобы последовательность была добавлена в нее требуется подтверждение эксперта. — *Прим. ред.*)

Главный сайт ExPASy (`http://www.expasy.ch`), со ссылками на SWISS-PROT и TrEMBL, предоставляет доступ к набору поисковых программ, включая ссылку на SRS (Sequence Retrieval System). В частности, с помощью SRS можно проводить поиск непосредственно в SWISS-PROT. Если выбрать опцию *All text* (искать во всем тексте) и поискать в SWISS-PROT слово «elastase» (эластаза), будут обнаружены запись ELNE_HUMAN, непосредственная цель нашего поиска, и еще 108 других записей, включая разнообразные ингибиторы. Возьмем из всего этого множества гомолог эластазы, выделенный из кровяной двуустки (*Schistosoma mansoni*): CERC_SCHMA. В обеих записях содержатся последовательности предшественников белков; в приведенном ниже выравнивании заглавными буквами отмечены участки зрелых ферментов:

```
CERC_SCHMA
-msnrwrfvvvvvtlftycltfervstwlIRSGEPVQHPAEFPFIAFLTTER-TMCTGSL 57
ELNE_HUMAN
mtlgrrlaclflacvlpalllggtalaseIVGGR-RARPHAWPFMVSLQLRGGHFCGATL 59
               :.:*  :.:: ::.  *    . :  * .*:   :*  :**:: *  :   :* .:*

CERC_SCHMA
VSTRAVLTAGHCVCSPLPVIRVSFLTLRNGDQQGIHHQPSGVKVAPGYMPSCMSARQRRP 117
ELNE_HUMAN
IAPNFVMSAAHCVAN--VNVRAVRVVLGAHNLSRREP--TRQVFAVQRIFENGYDP 111
             :::. *::*.***..    :.*  : :  * ::  :::*    .   :  . :  .   *

CERC_SCHMA
IAQTLSGFDIAIVMLAQMVNLQSGIRVISLPQPSDIPPPGTGVFIVGYGRDDNDRDPSRK 177
ELNE_HUMAN
VNLLN--DIVILQLNGSATINANVQVAQLPAQGRRLGNGVQCLAMGWGLLGRNRG-- 164
             :       **.*: *   ..::::.:* .**  .     *.  : :*:*  .:.*.
```

```
CERC_SCHMA
NGGILKKGRATIMECRHATNGNPICVKAGQNFGQLPAPGDSGGPLLPS-LQGPVLGVVSH 236
ELNE_HUMAN
IASVLQELNVTVVTS-LCRRSNVCTLVRGRQAG-VCFGDSGSPLVCNGLIHGIASFVRG 221
              ..:*:: ..*:: .  . ..*   :  *:: *   . ****.**: . *   : ...*

CERC_SCHMA    GVTLPNLPDIIVEYASVARMLDFVRSNI--------- 264
ELNE_HUMAN    GCASGLYPDAFAPVAQFVNWIDSIIQRSEDNPCPHPRDPDPASRTH 267
              * :    ** :.  *.... :* : ..
```

Пространственная структура эластазы нейтрофилов человека получена с помощью рентгеноструктурного анализа, но для эластазы кошачьей двуустки она не известна.

Одна из уникальных возможностей ExPASy — ссылка на SWISS-MODEL: автоматический сервер, для гомологичного моделирования. Заходя в SWISS-MODEL и выбирая *First approach mode* (это самое простое), мы можем просто открыть текст записи CERC_SCHMA, и запустить приложение. Построение пространственной модели гомолога — занятие не из легких, поэтому сервер работает длительное время, без участия пользователя, а результаты высылает по электронной почте.

Мы в дальнейшем еще разберемся со SWISS-MODEL подробнее в гл. 5.

Ресурс Ensembl

Ensembl (`http://www.ensembl.org`) создавался как универсальный источник данных по геному человека. Его записи должны содержать всю доступную информацию о последовательности ДНК человека, аннотации и привязки к оригинальной геномной последовательности, так чтобы все это было доступно для всего сообщества ученых, которые захотят использовать ее в своих целях с индивидуальным подходом. С этой целью, вдобавок к сбору и сортировке данных, много сил ушло на развитие вычислительной инфраструктуры. Было установлено соглашение о номенклатуре: не так-то просто изобрести схему поддержания постоянных идентификаторов, учитывая тот факт, что данные будут подвергаться не только пополнению, но и пересмотру. Наиболее наглядный результат этих стараний — сайт; он изобилует как средствами просмотра общей информации, так и возможностями разобраться в деталях.

Ensembl — совместный проект Европейского института биоинформатики и Центра Сэнгера; в числе его учредителей Э. Бирни (E. Birney), М. Клэмп (M. Clamp), Т. Кокс (T. Cox) и Т. Д. П. Хаббард (T. J. P. Hubbard). Тем не менее, Ensembl создан как открытый проект; поощряющий добавление новой информации пользователями. Наверное, даже самые наивные читатели понимают, что к новой информации предъявляются самые жесткие требования при проверке.

В Ensembl собраны гены, СНП (однонуклеотидные полиморфизмы), повторы и гомологи. Гены могут быть как экспериментально установленными, так и предсказанными по нуклеотидной последовательности. Поскольку экспериментальные данные для аннотации человеческого генома поступают непостоянно и неодинаково успешные, Ensembl предоставляет все сопровождающие доказательства по идентификации каждого гена. Множественные ссылки на базы данных, содержащие смежную информацию, такие как OMIM или базы данных по экспрессии, многократно дополняют и расширяют доступную информацию.

Ensembl структурирован специально для последовательности человеческого генома. Пользователи могут идентифицировать различные его области посредствам разнообразных поисковых программ.

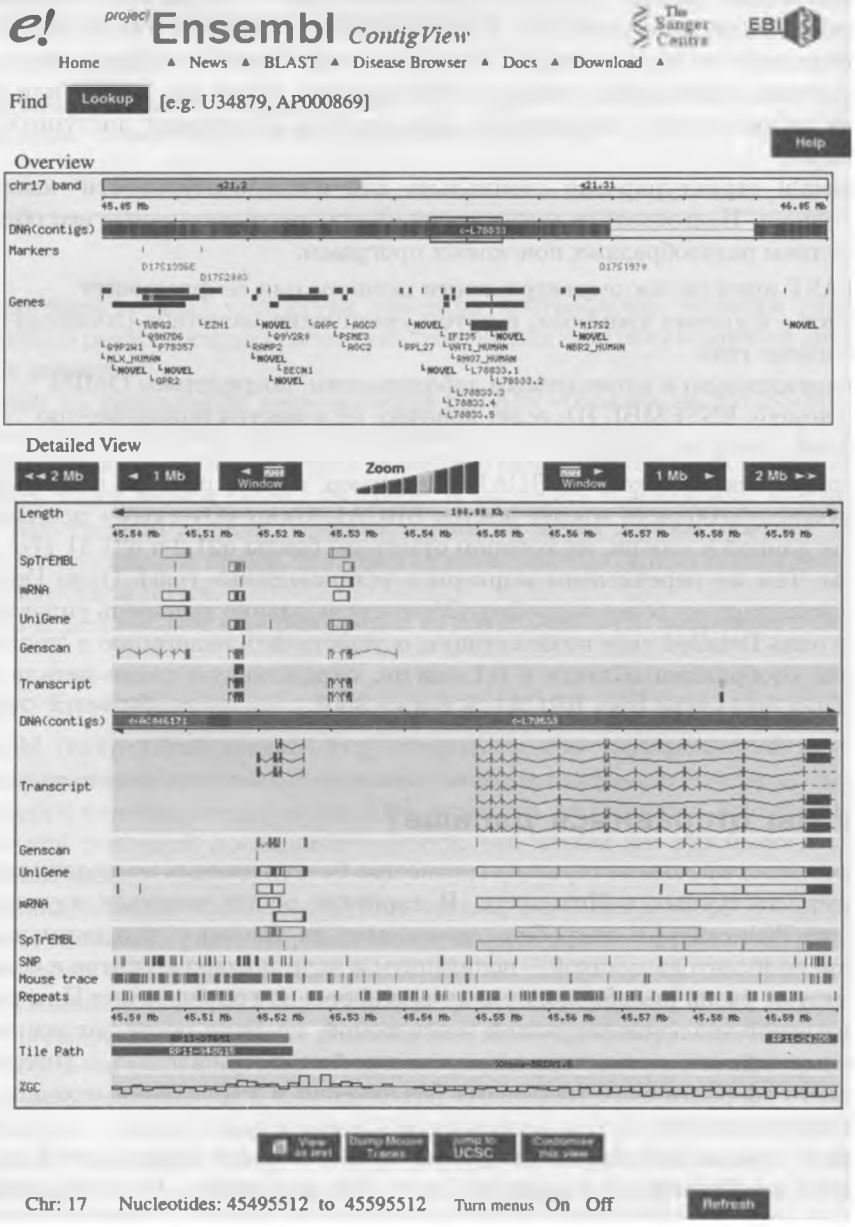
- BLAST ищет по последовательности целиком или ее фрагменту
- Поиск — с уровня хромосом, и потом увеличение масштаба (zooming)
- По имени гена
- Применительно к наследуемым заболеваниям, посредствам OMIM
- По номеру ENSEMBL ID, если, конечно, он известен пользователю
- Общий поиск по тексту.

Через поиск по тексту «BRCA1», например, можно попасть на страницу, показывающую область вокруг локуса BRCA1. Окно «Overview» показывает участок длиной в млн пн, на котором отмечены бенды q21.2 и q21.31 17-й хромосомы. Там же перечислены маркеры и установленные гены. Окно Detailed view показывает все более подробно. Обратите внимание на панель управления вверху окна Detailed view позволяющую осуществлять навигацию и 'zooming'. Под ней отображена область в 0,1 млн пн, отражающую такие детали как, подробная структура гена BRCA1, а также SNP.

Куда мы отправимся дальше?

Мы посетили только некоторые из множества банков данных по молекулярной биологии, доступных в Интернете. В короткие сроки читатели изучат эти и другие Web-сайты и подробно ознакомятся не только с содержимым Интернета, но и с его динамикой — появлением и исчезновением сайтов и ссылок. Существуют различные биологические метафоры для определения Интернета: его называют эволюционирующей экосистемой, по мере роста загрязняемой мертвыми сайтами и ссылками на мертвые сайты. К сожалению, в Интернете нет такого эффективного механизма разложения и утилизации отходов, как в органическом мире.

Банки данных развивают более эффективные связи между собой до тех пор, пока все близкие ссылками не будут явно соединены. Недолго осталось ждать до того момента, когда будет существовать только один банк данных по молекулярной биологии с множеством путей доступа. Ученые смогут формировать собственный доступ к частям информации, создавая «виртуальные базы данных» в соответствии с потребностями.


project Ensembl ContigView
The Sanger Centre
EBI
Home
News
BLAST
Disease Browser
Docs
Download
Find
Lookup
[e.g. U34879, AP000869]
Help
Overview
chr17 band
DNA(contigs)
Markers
Genes
Detailed View
2 Mb
1 Mb
Window
Zoom
Length
SpTrEMBL
mRNA
UniGene
Genscan
Transcript
SNP
Mouse trace
Repeats
Tile Path
ZGC
Chr: 17 Nucleotides: 45495512 to 45595512 Turn menus On Off
Refresh

Литература

Ежегодно январский выпуск журнала «Nucleic Acids Research» содержит подборку статей о базах данных по молекулярной биологии. Их следует держать под рукой для быстрого получения справки.

Bishop, M. J. (1999) *Genetics databases*. (London: Academic). [Краткое руководство по базам данных, доступ и анализ.]

Упражнения, задачи и компьютерные задания

Упражнение 3.1. База данных транспортных средств содержит следующие записи: велосипед, трехколесный велосипед, мотоцикл, автомобиль. Банк хранит только следующую информацию о каждой записи: (1) число колес и (2) движущая сила — человек или мотор. Для каждой возможной пары колес придумайте такую логическую комбинацию элементов запроса, о точном числе колес и о движущей силе, чтобы получить только два выделенных колеса, и ничего больше.

Упражнение 3.2. На с. 165 приведена запись NCBI для предшественника человеческой эластазы 1. Сделайте копию этой страницы, укажите на ней пункты, которые являются: (а) чисто служебными для базы данных; (б) внешним данными, такими как ссылки на литературу; (в) результатами экспериментальных измерений; (г) производной информацией, выведенной из экспериментальных измерений.

Упражнение 3.3. Почему поиск в базе данных структур ENTREZ выдал только 8 структур эластазы, а поиск в PDB выдал гораздо больше?

Задача 3.1. Напишите скрипт PERL, чтобы получить аминокислотную последовательность из записи базы данных белковых последовательности PIR, как показано на с. 142, и преобразовать ее в формат FASTA.

Задача 3.2. Сравните для человеческой эластазы файлы, найденные в категориях «белки» и «нуклеотиды» NCBI. На копиях этих двух страниц отметьте маркером все общие для двух файлов пункты.

Интернет-задание 3.1. Найдите запись SWISS-PROT для ингибитора трипсина поджелудочной железы быка (не для ингибитора секреции трипсина поджелудочной железы) и полную PIR запись для этого белка. Наличием какой информации отличаются эти записи друг от друга?

Интернет-задание 3.2. Найдите список официальных и неофициальных зеркальных сайтов Protein Data Bank. Какой из них ближе всего к вам?

Интернет-задание 3.3. Найдите все структуры миоглобина кашалота в Protein Data Bank и нарисуйте гистограмму дат их размещения в Protein Data Bank.

Интернет-задание 3.4. Найдите структуры белков, которые описал Peter Hudson, один или с коллегами.

Интернет-задание 3.5. Создайте строку поиска для использования инструментальным средством SearchLite из Protein Data Bank, которая (строка) выдаст структуры тиоредоксина *E. coli*, и не выдаст структуры нуклеазы стафилококка.

Интернет-задание 3.6. Для какой части структур, расположенных в Protein Data Bank и определенных при помощи рентгеноструктурного анализа, были размещены также файлы структурных факторов?

Интернет-задание 3.7. В Protein Data Bank запись 8XIA содержит структуру одного мономера изомеразы D-ксилозы из организма *Streptomyces rubiginosus*. Какова вероятная четвертичная структура, выведенная из координат атомов?

Интернет-задание 3.8. Найдите «структурных соседей» для записи 2TRX Protein Data Bank (тиоредоксин *E. coli*), соответствующие SCOP, CATH, FSSP и CE. Существуют ли какие-нибудь структуры, которых все эти классификации определяют как «структурных соседей» 2TRX? Какие структуры определены как «структурные соседи» в некоторых, но не во всех классификациях?

Интернет-задание 3.9. Почему поиск ENTREZ в категории «белок» для эластазы человека выдает тРНК-синтетазу?

Интернет-задание 3.10. На с. 164 приведена аминокислотная последовательность предшественника человеческой эластазы 1. Каковы различия между его последовательностью и последовательностью созревшего белка?

Интернет-задание 3.11. Какова связь между последовательностями эластазы, полученными в результате поиска NCBI и PIR?

Интернет-задание 3.12. Используя SWISS-PROT напрямую или через SRS, получите запись SWISS-PROT для человеческой эластазы. Какую информацию, которой нет в соответствующей записи (а) ENTREZ (белок); (б) PIR, содержит этот файл?

Интернет-задание 3.13. Какие гомологи эластазы нейтрофила человека можно найти при помощи PSI-BLAST?

Интернет-задание 3.14. Найдите как минимум 6 мутаций в человеческой эластазе, связанной с циклической нейтропенией, и отметьте их на выравнивании последовательностей (цветная иллюстрация VI). Сохраняется ли затронутая позиция для каждого мутанта более, чем в половине природных последовательностей?

Интернет-задание 3.15. Какой ген в *C. elegans* кодирует схожий по последовательности с человеческой эластазой?

Интернет-задание 3.16. Какова локализация человеческого гена глюкоза-6-фосфатдегидрогеназы в хромосоме?

Интернет-задание 3.17. Псевдогены эукариот могут быть разделены на те, которые возникают из-за удвоения и расхождения генов, и те, которые были вставлены в геном из мРНК ретровирусом, называемые процессинговыми псевдогенами. Процессированные псевдогены можно отличить по отсутствию интронов. Существуют ли какие-нибудь псевдогены в кластерах гена глобина человека, которые являются процессинговыми псевдогенами?

Интернет-задание 3.18. Предварительный генетический анализ, направленный на изоляцию гена, связанного с кистозным арахноидитом, поместил этот ген между MET онкогеном и RFLP D7S8. Впоследствии было посчитано, что этот участок содержит 1–2 млн по и может содержать 100–200 генов. (а) Сколько пар оснований длиной оказался этот участок на самом деле? (б) Сколько экспрессированных генов, как сейчас считается, содержит этот участок?

Интернет-задание 3.19. Ген синдрома Берардинелли—Сейп был локализован между двумя метками на зоне хромосомы 11q13 — D11S4191 и D11S987. Сколько пар оснований между этими двумя маркерами?

Интернет-задание 3.20. Существует ли база данных, доступная в Интернете, которая собирает специфическую информацию по термодинамике и структуре взаимодействий белков с нуклеиновыми кислотами?

Интернет-задание 3.21. База данных протеома дрожжей содержит запись для cdc6, белка, который регулирует инициацию репликации ДНК. (а) На какой хромосоме локализован ген cdc6-дрожжей? (б) Какой посттрансляционной модификации подвергается это белок, чтобы достичь зрелого активного состояния? (в) Каковы ближайшие известные родственники этого белка в других видах? (г) Для каких белков известно взаимодействие с cdc6-дрожжей? (д) Каково воздействие дистамицина A на активность cdc6-дрожжей? (е) Как влияет актиномицин D на активность cdc6-дрожжей?

4 Выравнивания и филогенетические деревья

Выравнивание последовательностей. Введение

Цели сравнения двух или более последовательностей:

- соизмерить их сходство и установить соответствие между остатками,
- отметить консервативные и вариабельные области,
- высказать соображения об эволюционных взаимосвязях.

Если мы сможем сделать все это, то легко искать родственные последовательности в банках данных. Выравнивание последовательностей в основном применяется при аннотировании геномов, в том числе для определения структуры и функций генов.

Как можно количественно определить уровень сходства последовательностей? Чтобы сравнить нуклеиновые основания или аминокислоты в различных позициях двух или более последовательностей, первым делом мы должны назначить сопоставления. *Выравнивание последовательностей, — это определение соответствия между остатками.* Это основной инструмент биоинформатики.

Любые назначения соответствий, которые сохраняют *порядок* остатков в последовательности — это выравнивание, при этом возможна вставка пропусков.

Даны две текстовые строки:

Первая строка = `a b c d e`	вторая строка = `a c d e f`
Разумное выравнивание выглядит так:	`a b c d e -` `a - c d e f`

Для того чтобы находить *оптимальное выравнивание*, мы должны определить критерий качества выравнивания. Возможные выравнивания для последовательностей `gctgaacg` и `ctataatc`:

Неинформативное выравнивание	`- - - - - - - g c t g a a c g` `c t a t a a t c - - - - - - -`
Выравнивание без пропусков	`g c t g a a c g` `c t a t a a t c`
Выравнивание с пропусками	`g c t g a - a - - c g` `- - c t - a t a a t c`
И еще одно выравнивание	`g c t g - a a - c g` `- c t a t a a t c -`

Большинство читателей сочтут последнее из этих выравниваний лучшим из четырех. Чтобы решить, лучшее ли оно из *всех* возможных, нам необходим способ систематической проверки всех возможных выравниваний. Затем нам надо рассчитать вес выравнивания, отражающий качество каждого возможного выравнивания, и определить выравнивание с оптимальным весом. Отметим, что даже незначительные изменения в схеме оценки могут изменять ранг выравниваний, из-за чего другие становятся лучшими.

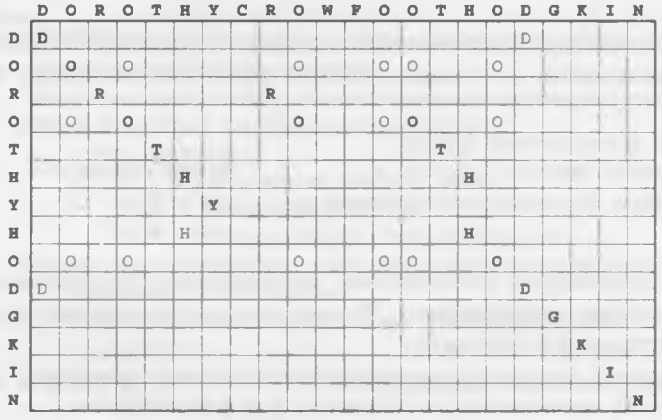
Точечная матрица сходства

Точечная матрица — простейшее изображение, которое дает представление о сходстве между двумя последовательностями. Менее очевидным является ее близкая взаимосвязь с выравниваниями.

Точечная матрица представляет собой таблицу или матрицу. Строки соответствуют остаткам в одной последовательности, колонки — остаткам в другой последовательности. В простейшем варианте позиции в точечной матрице оставляются пустыми, если остатки различны, и заполняются, если они совпадают. Совпадающие фрагменты последовательностей отобразятся в виде диагоналей, идущих из верхнего левого угла в нижний правый.

ПРИМЕР 4.1.

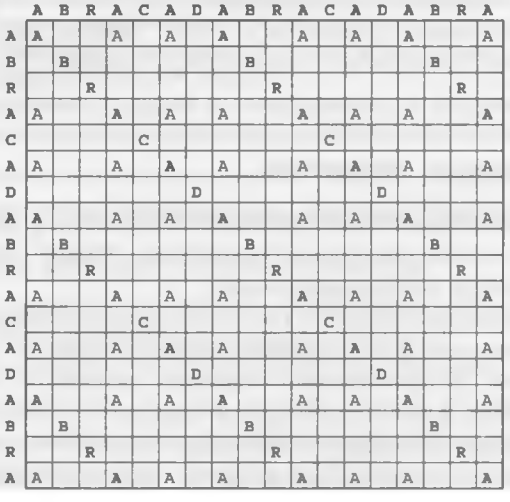
Точечная матрица, показывающая совпадения между коротким (DOROTHYHODGKIN) и полным именами (DOROTHYCROWFOOTHODGKIN) известного кристаллографа.



Буквы, соответствующие *отдельным* совпадениям, не выделены жирным шрифтом. Самые длинные совпадающие участки, выделенные жирным шрифтом, — это первое и последнее имена DOROTHY и HODGKIN. Более короткие совпадающие участки, такие как OTH из dorOTHy и crowfoOTHodkin или RO из doROthy и cROwfoot, являются шумом.

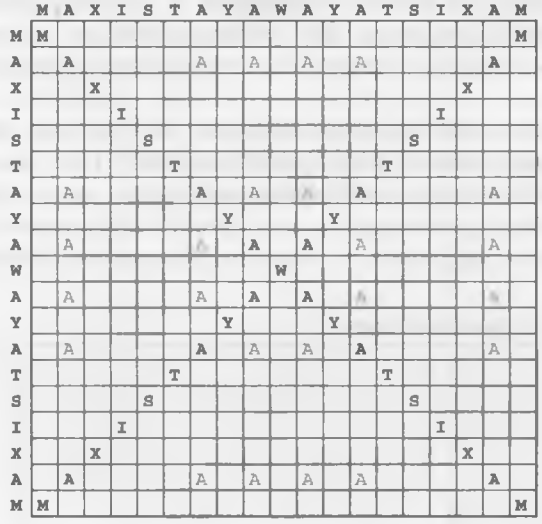
ПРИМЕР 4.2.

Точечная матрица, показывающая совпадения повторяющейся последовательности (ABRACADABRACADABRA) с самой собой.



ПРИМЕР 4.3.

Точечная матрица показывает, насколько палиндромная последовательность MAXISTAYAWAYATSIXAM похожа на саму себя[1]. Палиндром проявляется как побочная диагональ, пересекающаяся с главной диагональю (из верхнего левого угла в нижний правый).



Палиндромы — это не просто игра слов — это участки ДНК, распознаваемые регуляторами транскрипции или ферментами рестрикции, имеющими похожие на палиндром последовательности, которые располагаются поочередно то на одной цепи, то на другой:

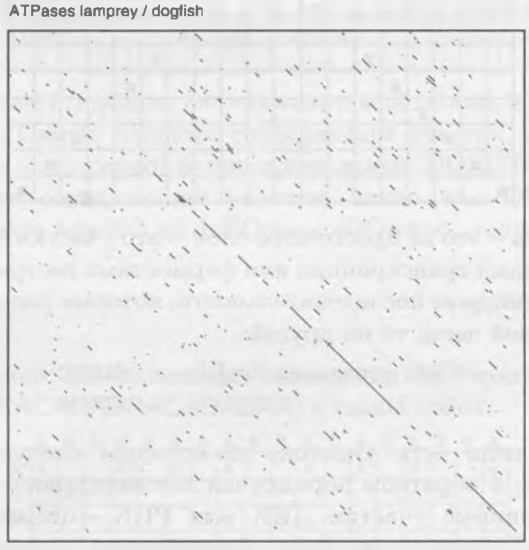
Сайт узнавания EcoRI: GAATTC
CTTAAG

В каждой цепи есть участок, за которым следует комплементарная, записанная в обратном порядке последовательность (см. задание 4.8 и упр. 4.8). Длинные участки ДНК или РНК, содержащие инвертированные повторы такого типа, могут формировать шпилечные структуры. В добавлении к этому, некоторые подвижные элементы, выделенные из растений, содержат настоящие (неточные) палиндромные последовательности — инвертированные повторы некомплементарных последовательностей, расположенных на той же цепи; следующий пример представлен последовательностью генома вируса, вызывающего остановку роста пшеницы (wheat dwarf virus): ttttcgtgagtgcggaggctttt. (Часто палиндромность последовательности определяется тем, что с этим участком ДНК должен взаимодействовать белок в виде димера, одна субъединица которого взаимодействует с одним плечом палиндрома, а другая — с другим плечом на комплементарной цепи. — *Прим. ред.*).

[1] Это является не вполне верным утверждением.

Точечная матрица позволяет быстро проиллюстрировать родство между двумя последовательностями. Яркие признаки сходства четко проявляются. Например, точечная матрица, отображающая родство между генами митохондриальной АТФазы-6 из миноги (*Petromyzon marinus*) и из морской собаки (*Scyliorbinus canicula*), показывает, что сходство между этими последовательностями меньше всего выражено вначале (см. рисунок ниже). Этот ген кодирует субъединицу АТФазного комплекса. У человека мутации в этом гене являются причинами синдрома Ли — неврологических отклонений у детей, вызываемых эффектами замедления окислительного метаболизма в мозговых тканях.

Недостатком точечной матрицы является то, что она плохо определяет сходство дальнеродственных последовательностей. При анализе последовательностей, с одной стороны, необходимо наблюдать за построением точечной матрицы, чтобы быть уверенным в том, что не пропущено ничего важного, но с другой стороны, надо быть готовым к применению более тонких методов анализа.



Часто участки сходства могут быть смещены, появляются на параллельных, но неколлинеарных диагоналях. Это показывает, что в сегментах между участками сходства возникают вставки и делеции. Точечная матрица, отражающая родство между белком PAX-6 из мыши и белком eyless из плодовой мушки *Drosophila melanogaster*, показывает три продолжительных участка сходства. Два из этих участков находятся в начале последовательностей, а третий — в середине. Между двумя из трех участков в последовательности белка из мыши есть более длинный промежуточный участок, чем в последовательности белка из *Drosophila*.

Фильтрование результатов может сократить количество шума в точечной матрице. В сравнении последовательностей АТФаз точки не были показаны до тех пор, пока в центральном участке из 15 упорядоченных остатков не стало,

как минимум, 6 совпадений. Программа для построения точечных матриц на языке PERL (см. с. 190) позволяет пользователю установить значения для *окна* (длина участка, состоящего из упорядоченных остатков) и граничного значения *порога* (число совпадений, находящихся в рамках значений окна).

WEB-РЕСУРСЫ: ТОЧЕЧНЫЕ МАТРИЦЫ

Программа Dotter И. Л. Соннхаммера (E. L. Sonnhammer) считает и отображает точечные матрицы. Эта программа позволяет пользователю осуществлять контроль над расчетами и изменениями результатов путем изменения параметров интерактивно. Программа расположена по адресу `http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html`

Для того чтобы использовать полный набор функций программы Dotter, необходимо установить ее локально.

Web-сайт, предлагающий интерактивное построение точечных матриц расположен по адресу:

`http://www.isrec.isb-sib./ch/java/dotlet/exonintron.html`

Программа на языке PERL для отображения точечных матриц

Программа принимает следующие данные:

1. Заголовок или название задания, отпечатанное сверху над графиком (первая строка ввода данных)
2. Параметры, определяющие фильтрующие параметры *графического окна* и *порогового значения* (вторая строка ввода данных). Точка появится на графике, если данные попадают в центр распределения остатков по длине *окна* таким образом, что число совпадений является большим или равным *пороговому значению*.
3. Две последовательности, где каждая начинается со строки заголовка и заканчивается «*» (звездочкой).

Программа рисует точечные матрицы похожие на те, которые рассмотрены в тексте. Выводные данные представляют собой графический язык называемый PostScript™, который может быть отображен или отпечатан на многих устройствах.

```
#!/usr/bin/perl
#dotplot.pl - reads two sequences and prints dotplot

# read input

$/ = "";
$_ = <DATA>; $_ =~ s/#(.*)\n/\n/g;
$_ =~ /^(.*)\n\s*(\d+)\s+(\d+)\s*\n(.*)\n([A-Z\n]*)\*\s*\n(.*)\n([A-Z\n]*)\*/;
$title = $1; $nwind = $2; $thresh = $3;
$seqt1 = $4; $seq1 = $5; $seqt2 = $6; $seq2 = $7;
$seq1 =~ s/\n//g; $seq2 =~ s/\n//g; $n = length($seq1); $m = length($seq2);

# postscript header

print <<EOF;
%!PS-Adobe-
/s /stroke load def /l /lineto load def /m /moveto load def /r /rlineto load def
/n /newpath load def /c /closepath load def /f /fill load def
1.75 setlinewidth 30 30 translate /Helvetica findfont 20 scalefont setfont
EOF

#print matrix

$dx = 500.0/$n;  $mdx = -$dx; $dy = 500.0/$m;
if ($dy < $dx) {$dx = $dy;} $dy = $dx; $xmx = $n*$dx; $ymx = $m*$dx;
print "0 510 m ($title NWIND = $nwind) show\n";
printf "0 0 m 0 %9.2f l %9.2f %9.2f l %9.2f 0 l c s\n", $ymx,$xmx,$ymx,$xmx;
```

```
for ($k = $nwind - $m + 1; $k < $n - $nwind; $k++) {
    $i = $k; $j = 1; if ($k < 1) {$i = 1; $j = 2 - $k;}
    while ($i <= $n - $nwind && $j <= $m - $nwind) {
        $_ = (substr($seq1,$i -1,$nwind) ^ substr($seq2,$j -1,$nwind));
        $mismatch = ($_ =~ s/[^\x0]//g);
        if ($mismatch < $thresh) {
            $xl = ($i - 1)*$dx; $yb = ($m - $j)*$dy;
            printf "n %9.2f %9.2f m %9.2f 0 r 0 %9.2f r %9.2f 0 r c f\n",
                    $xl,$yb,$dx,$dy,$mdx;
        }
        $i++; $j++;
    }
}
print "showpage\n";

__END__
ATPases lamprey / dogfish                    #TITLE
15 6                                         #WINDOW, THRESHOLD
Petromyzon marinus mitochondrion             #SEQUENCE 1
ATGACACTAGATATCTTTGACCAATTTACCTCCCCAACA
ATATTTGGGCTTCCACTAGCCTGATTAGCTATACTAGCCCCTAGCTTA
ATATTAGTTTCACAAACACCAAAATTTATCAAATCTCGTTATCACACACTA
CTTACACCCATCTTAACATCTATTGCCAAACAACTCTTTCTTCCAATAAAC
CAACAAGGGCATAAATGAGCCTTAATTTGTATAGCCTCTATAATATTTATC
TTAATAATTAATCTTTTAGGATTATTACCATATACTTATACACCAACTACC
CAATTATCAATAAACATAGGATTAGCAGTGCCACTATGACTAGCTACTGTC
CTCATTGGGTTACAAAAAAAACCAACAGAAGCCCTAGCCCACTTATTACCA
GAAGGTACCCCAGCAGCACTCATTCCCATATTAATTATCATTGAAACTATT
AGTCTTTTTATCCGACCTATCGCCCTAGGAGTCCGACTAACCGCTAATTTA
ACAGCTGGTCACTTACTTATACAACTAGTTTCTATAACAACCTTTGTAATA
ATTCCTGTCATTTCAATTTCAATTATTACCTCACTACTTCTTCTATTA
CTAACAATTCTGGAGTTAGCTGTTGCTGTAATCCAGGCATATGTATTTATT
CTACTTTTAACTCTTTATCTGCAAGAAAACGTTT*
Scyliorhinus canicula mitochondrion          #SEQUENCE 2
ATGATTATAAGCTTTTTTGATCAATTCCTAAGTCCCTCCTTTCTAGGA
ATCCCACTAATTGCCCTAGCTATTTCAATTCCATGATTAATATTTCCAACACCAACC
AATCGTTGACTTAATAATCGATTATTAACTCTTCAAGCATGATTTATTAACCGATTTATT
TATCAACTAATACAACCCATAAATTTAGGAGGACATAAATGAGCTATCTTATTTACAGCC
CTAATATTATTTTTTAATTACCATCAATCTTCTAGGTCTCCTTCCATATACTTTTACGCCT
ACAACTCAACTTTCTCTTAATATAGCCTTTGCCCTGCCCTTATGGCTTACAACTGTATTA
ATTGGTATATTTAATCAACCAACCATTGCCCTAGGGCACTTATTACCTGAAGGTACCCCA
ACCCCTTTAGTACCAGTACTAATCATTATCGAAACCATCAGTTTATTTATTCGACCATTA
GCCTTAGGAGTCCGATTAACAGCCAACTTAACAGCTGGACATCTCCTTATACAATTAATC
GCAACTGCGGCCTTTGTCCTTTTAACTATAATACCAACCGTGGCCTTACTAACCTCCCTA
GTCCTGTTCCTATTGACTATTTTAGAAGTGGCTGTAGCTATAATTCAAGCATACGTATTT
GTCCTTCTTTTAAGCTTATATCTACAAGAAAACGTATAA*
```

Точечные матрицы и выравнивание последовательностей

Точечная матрица собирает в одном изображении не только полную информацию о сходстве двух последовательностей, но также предоставляет полный набор (и относительное качество) других возможных выравниваний.

Любой путь по точечной матрице от верхнего левого угла до нижнего правого, двигаясь при каждом шаге только строго направо, вниз и по диагонали, соответствует возможному выравниванию[1]. Если две последовательности близкородственны, то выравнивание может быть считано непосредственно с графика.

На рис. 4.1 изображен пример построения точечной матрицы при выравнивании слов Dorothy Hodgkin.

Если направление движения между последующими ячейками диагональное, то две пары следующих друг за другом сравниваемых остатков оказываются в выравнивании без вставки между ними (сопоставляются). Если направление движения горизонтальное, то в последовательность, служащую указателем рядов, вставляется пропуск. Если же направление движения вертикальное (вниз), то пропуск вставляется в последовательность, индексирующую столбцы.

Следует обратить внимание на то, что ни одно движение не может совершаться вверх или влево, так как это соответствовало бы сравнению нескольких остатков одной последовательности со всего лишь одним остатком другой. Путь, указанный стрелками, соответствует очевидному выравниванию:

```
DOROTHY--------HODGKIN
DOROTHYCROWFOOTHODGKIN
```

Другой способ интерпретации пути по точечной матрице — это порядок редактирования (*edit script*). Он представляет собой указание серий операций, которые трансформируют последовательность, индексирующую столбцы, — горизонтально расположенную последовательность — в последовательность, которая индексирует ряды, — вертикально расположенную последовательность. Каждое движение говорит нам о проведении одной из операций — замены, вставки или делеции. По достижении конца пути получится преобразование одной последовательности в другую. В целом, несколько различных последовательностей редакционных операций могут превратить одну последовательность в другую за одно и то же количество шагов, однако они могут продуцировать при этом различные выравнивания.

Следует подчеркнуть, что, несмотря на то что последовательность редакционных операций получена из оптимального выравнивания, и она *возможно* соответствует реальному эволюционному пути, но невозможно *доказать*, что

[1] Это ограничение на переходы соответствует тому, что в выравнивании должна сохраняться последовательность символов и исключаются перестановки символов. — *Прим. ред.*

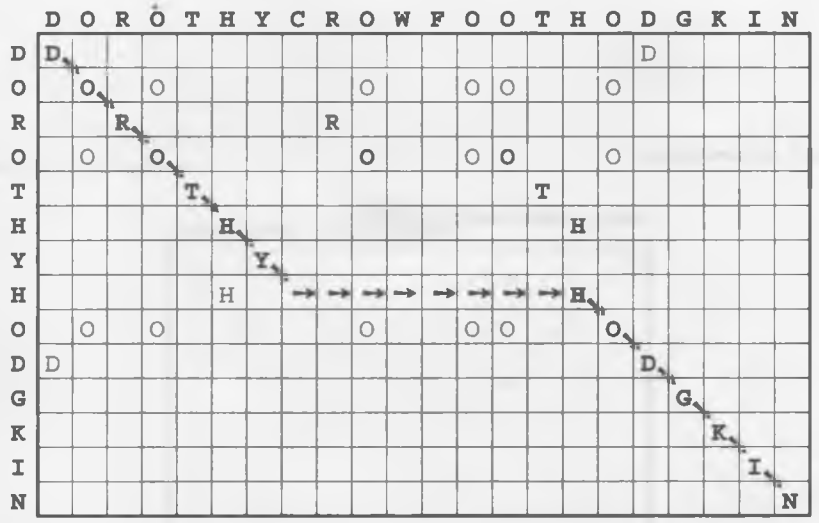


Рис. 4.1. Любой путь по точечной матрице от верхнего левого угла к нижнему правому проходит последовательность ячеек, каждая из которых предсказывает пару позиций: одну из ряда и одну из столбца, которые совпадают с выравниванием; либо означают пробел в одной из последовательностей. Путь не обязательно должен проходить лишь заполненные позиции. Тем не менее, чем больше заполненных позиций, на диагональном отрезке пути, тем больше совпадающих остатков в выравнивании

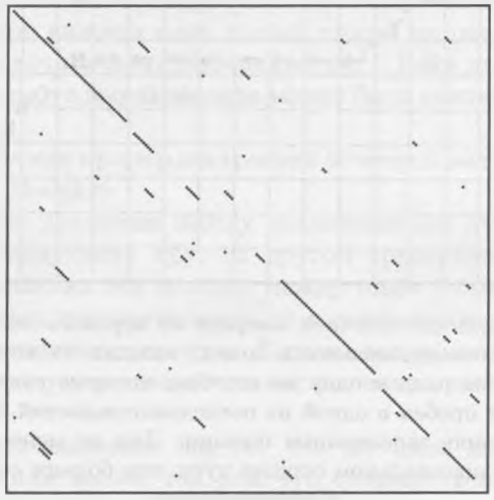
это действительно так. Чем больше редакционное расстояние, тем больше разумных эволюционных путей между двумя последовательностями.

ПРИМЕР 4.4.

Точечные матрицы и выравнивания. Давайте сравним вид графиков между парами белков с возрастанием дальности родства между ними. На рис. 4.2а показана точечная матрица при сравнении сульфгидрильного протеиназного папаина из папайи, с четырьмя гомологами — близким родственником (актинидином из плодов киви) и более отдаленными родственниками (человеческим прокатепсином L, человеческим катепсином B, и стафопаином из *Staphylococcus aureus)*. В примере показаны также выравнивания аминокислотных последовательностей. Так как последовательности постепенно расходятся, становится труднее и труднее угадать правильное выравнивание на точечной матрице. Показанные выравнивания были получены путем сравнения структур.

ПРИМЕР 4.4 *продолжение*

PAPA_CARPA / ACTN_ACTCH



```
ВЫРАВНИВАНИЕ 9pap и 2act (см. рис. 4.2a)
SCORE = 5324  NPOS = 219 NIDENT = 102  %IDENT = 46.58

IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDR-
 | |||||  |||   | || || ||||||  | ||| ||  | |   ||||| || |
LPSYVDWRSAGAVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELIDCGRTQ

RSYGCNGGYPWSALQ-LVAQYGIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYN
   || |||     |      ||     |||      |             |    |   |
NTRGCDGGYITDGFQFIINDGGINTEENYPYTAQDGDCDVALQDQKYVTIDTYENVPYNN

QGALLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGP--NYILI
  ||      ||||| | |||  |  |  ||| ||||  ||||   ||||      |
EWALQTAVTYQPVSVALDAAGDAFKQYASGIFTGPCGTAVDHAIVIVGYGTEGGVDYWIV

KNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
|||| | ||| || || |  |   | ||  |   ||||
KNSWDTTWGEEGYMRILRNVGGA-GTCGIATMPSYPVKY
```

Рис. 4.2а. Выравнивание папаина папайи и актинидина из плодов киви с соответствующей точечной матрицей

ПРИМЕР 4.4 *продолжение*

PAPA_CARPA / CATL_HUMAN

```
ВЫРАВНИВАНИЕ 9pap и 1cjl (см. рис. 4.2б)
SCORE = 3214 NPOS = 220 NIDENT  = 81  %IDENT = 36.82

IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCD-R
     ||| || |||||||| ||| |||||    ||     || |   ||| | ||
V--DWREKGYVTPVKNQGQCGSSWAFSATGALEGQMFRKTGRLISLSEQNLVDCSGPE

RSYGCNGGYPWSALQLVAQY-GIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYN
   |||||    | | |    |      ||||     |    |    |   |
GNEGCNGGLMDYAFQYVQDNGGLDSEESYPYEATEESCKYNPKYS-VANDAGFVDIPKQE

QGALLYSIANQPVSVVLQAAGKDFQLYRGGIFVGP-CGNKVDHAVAAVGYG--PNYIL
          | ||   |    |  |  ||   |       || |  ||||     | |
KALMKAVATVGPISVAIDAGHESFLFYKEGIYFEPDCSSEDMDHGVLVVGYGFESNKYWL

IKNSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN
 |||||  ||  ||        |    ||      ||
VKNSWGEEWGMGGYVKMAKDRRN-H-CGIASAASYPTV-
```

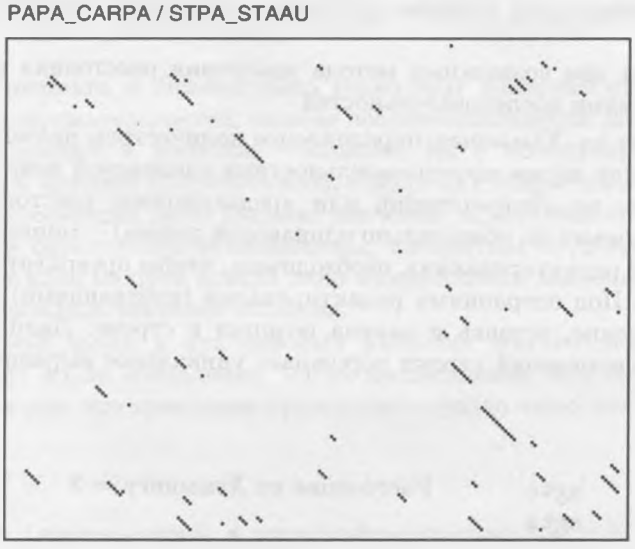
Рис. 4.2б. Выравнивание папаина папайи и человеческого прокатепсина L с соответствующей точечной матрицей. Этот график показывает присутствие нескольких похожих участков, однако достаточно сложно построить полное выравнивание последовательностей, используя точечную матрицу

ПРИМЕР 4.4 *продолжение*

PAPA_CARPA / CATB_HUMAN

```
ВЫРАВНИВАНИЕ 9pap и 1huc (см. рис. 4.2в)
SCORE   =  2073  NPOS = 251  NIDENT  = 66  %IDENT = 26.29

IPEYVD-WRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLD-C-D
      | |          |||||||||| ||  |   | | |      |   ||  |
-DAREQWPQCPTIKEIRDQGSCGSCWAFGAVEAISDRICIHTNVSVEVSAEDLLTCCGS

RRSYGCNGGYP---WSALQLVAQYGI-HYRN-TY---P-YEGVQRYCRSREKG
    |||||||        |    ||       |      |          |    |
MCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEHHVNGSRPPCTGEGDT

PYAAK---TDGVRQVQPYNQGALLYSIANQPVSV-V---LQ--AAGKDFQLYRG
|   |         |   |        |                        || ||
PKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKS

GIFVGPCGNKV-DHAVAAV-GY-GPNYILIKNSWGTGWGENGYIRIKRGTGNSYGVCG
|      |     ||      |   |  | |  ||| | || ||   | ||     |
GVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQ-DHCGIES

LYTSSFYPVKN
       |
EVVAGI-PRTD
```

Рис. 4.2в. Выравнивание последовательности белка папаина из папайи и человеческого белка печени катепсина B с соответствующей точечной матрицей для двух выравниваний. И для самих выравниваний, и для диаграммы, большая степень сходства наблюдается в начале и конце последовательностей, чем в их центральных участках




```
ВЫРАВНИВАНИЕ 9pap и 1cv8 (см. рис. 4.2г)
SCORE = -290  NPOS = 219  NIDENT  = 25 %IDENT = 11.42

IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRRS
                                                        |    |
----------------------EQYVNKLENFKIRE

YGCNGGYPWSALQLVAQYGIHYRNTYPYEGVQRYCRSREKG-PYAAKTDGVRQVQPY--
  | |                 | |  | | |       |          |
TQGNNGWCAGYTMSALLNATYNTNKYHAEAVMRFLHPNLQGQQFQFTGLTPREMIYFGQT

-NQGALLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGPNYILIK
     ||           |    |      |       |    || | ||
QGRSPQLLNRMTTYNEVDNLTKNNKGIAIL-GSRVESRNGMHAGHAMAVVGNAKLNNGQE

NSWGTGWGENGYIRIKRGTGNSYGVCGLYTSSFYPVKN-
        ||              |
VIIIWNPWDNGFMTQDAKNNVIPVSNGDHYQWYSSIYGY
```

Рис. 4.2г. Выравнивание последовательности папаина из папайи и белка стафопаина из *S. aureus* с соответствующей точечной матрицей. При помощи полученной диаграммы выравнивание определить невозможно

Мера сходства последовательностей

Для того чтобы пойти еще дальше в определении выравниваний «на глазок», необходимо определить единицы измерения сходства и различия последовательностей.

Существует два возможных метода измерения расстояния между двумя данными строками последовательностей.

(1) Расстояние *по Хэммингу*, определяемое количеством несовпадающих позиций между двумя последовательностями одинаковой длины.

(2) Расстояние *по Левенштайну* или «редакционное расстояние» (между двумя строками не обязательно одинаковой длины) — минимальное число «операций редактирования», необходимых, чтобы превратить одну строку в другую. Под операциями редактирования (изменениями) подразумеваются удаление, вставка и замена позиции в строке. Данная последовательность изменений строки порождает уникальное выравнивание, но не наоборот.

Например:

```
agtc
cgta
```
Расстояние по Хэммингу = 2

```
ag-tcc
cgctca
```
Расстояние по Левенштайну = 3

Для приложений молекулярной биологии считается, что некоторые определенные изменения (замены) происходят вероятнее других. Например, аминокислотные замены достаточно консервативны: замена одной аминокислоты на другую с похожими свойствами (размер или физико-химические характеристики) происходит с гораздо большей вероятностью, чем замена на другую аминокислоту со значительно отличающимися параметрами. Или, например, делеция расположенных рядом нуклеотидных оснований или аминокислот — более вероятное событие, чем делеция того же количества позиций, независимо расположенных в последовательности. По этой причине при расчете расстояния между последовательностями имеет смысл назначить различные «цены» за каждый из видов замены. Таким образом, компьютерная программа сможет вычислить не просто кратчайший путь из сделанных замен, а оптимальные выравнивания. Складывая цены отдельных изменений позиции программа, может присвоить общую цену для каждого из путей. Для определения цены замены в данной позиции программа учитывает цену произошедшей мутации, изменяющейся в зависимости от мутировавших остатков. Для вертикальных и горизонтальных передвижений позиции программа добавляет подходящий штраф за образование пропуска или вставки.

Схемы оценки

Система, по которой рассчитываются «цены» изменений данной позиции, должна учитывать замены и вставки или делеции (вставка как изменение одной последовательности — то же самое, что и делеция для второй)[1]. Цена делеции нескольких символов будет иметь цену, зависящую от их длины пропуска.

Методы Хэмминга и Левенштайна позволяют измерить степень *несходства* двух последовательностей: схожие последовательности дают маленькие значения расстояния, а несхожие — большие. Но в молекулярной биологии выравниваниям принято назначать веса, исходя из степени *сходства*, т. е. схожие последовательности дают высокие значения, а различающиеся — низкие. Поскольку эти определения эквивалентны, алгоритмы оптимального выравнивания могут идти по пути поиска либо минимальных значений несхожести, либо — максимальных значений сходства.

Для подсчета замен в нуклеиновых кислотах обычно используют или простую схему: +1 за совпадение, −1 за несовпадение, или более сложную, основанную на том, что транзиции происходят гораздо чаще, чем трансверсии.

ПРИМЕР 4.5.

Транзиции (*пурин*↔*пурин* и *пиримидин*↔*пиримидин;* т. е. a↔g и t↔c) встречаются чаще *трансверсий*(*пурин*↔*пиримидин;* т. е. (a, g)↔(t, c)). Это наблюдение можно описать матрицей замен.

Например, одна из возможных матриц:

	a	t	g	c
a	20	10	5	5
t	10	20	5	5
g	5	5	20	10
c	5	5	10	20

Для аминокислотных последовательностей было предложено несколько схем. Аминокислоты можно группировать по их физико-химическим свойствам, и добавлять +1 к цене выравнивания, если аминокислота в позиции первой последовательности принадлежит той же группе, что и аминокислота из второй последовательности, и −1, если аминокислоты принадлежат к различным группам.

Другой способ — строить схему подсчета на основе данных о множестве белков. Такой метод впервые предложила М. О. Дэйхофф (M. O. Dayhoff). Она собирала статистику по частотам аминокислотных замен в известных белках, и результаты ее работ использовались для подсчета цен выравниваний

[1] Поскольку при сравнении двух последовательностей нельзя различить вставку от делеции, в англоязычной литературе часто используют универсальный термин indel (от insertion + deletion). Мы в этом случае будем использовать термин делеция, понимая, что в соответствующем месте может быть произошла вставка. — *Прим. ред.*

в течение многих лет. В последствии они были заменены новыми матрицами, полученными в результате обработки огромного количества расшифрованных последовательностей.

Получение матриц замен

Чем сильней расходятся последовательности, тем больше в них накапливается мутаций. Для того чтобы измерить относительную возможность какой-нибудь конкретной замены, например серин → треонин, необходимо подсчитать количество таких замен в парах гомологичных, уже выровненных последовательностей. Можно использовать величины относительных частот таких замен для построения матрицы замен. Относительно частая замена должна цениться дороже относительно редкой. Но что делать, если в данных позициях происходило несколько замен? Это бы внесло ошибку в общую статистику. Этой проблемы можно избежать, ограничивая набор анализируемых последовательностей таким образом, чтобы все они были достаточно близки и можно было бы предположить, что ни одна позиция не менялась больше одного раза.

Мера расхождения последовательностей оценивается PAM в единицах. PAM (Percent Accepted Mutation) — процент зафиксированных мутаций. Таким образом, две последовательности имеют расстояние 1 PAM, если они совпадают на 99%. Для пар последовательностей с уровнем расхождения менее 1 PAM вероятнее всего, что в процессе эволюции ни одна из позиций не менялась больше чем один раз. Статистика, собранная на основе информации о парах последовательностей, расположенных примерно на таком эволюционном расстоянии, и ее корректировка для разных аминокислотных составов последовательностей порождают *матрицу замен PAM1*. Можно воспользоваться этим принципом, чтобы получить матрицу, подходящую для последовательностей, которые расходятся гораздо сильнее. Уровень в 250 PAM, соответствующий примерно 20%-й идентичности последовательностей, — минимальный уровень сходства, для которого можно надеяться получить правильное выравнивание, основываясь только на анализе самой последовательности (без привлечения дополнительных данных, таких, как структура белка)[1)]. Поэтому только эти уровни сходства подходят для практических работ (см. с. 202). (Некоторые авторы разработали матрицы замен, подходящие для разных уровней сходства последовательностей).

Появление реверсий (как непосредственных, так и получающихся из-за одной или многих замен) объясняет уменьшение различий в скоростях наступления мутаций при расхождении последовательностей.

PAM	0	30	80	110	200	250
% сходства	100	75	50	60	25	20

[1)] Расстояние 250 PAM означает, что при эволюции последовательности длиной 100 аминокислотных остатков произошло 250 мутационных событий. Однако при этом мутации происходили в *случайных* позициях и поэтому по случайным причинам в некоторых позициях мутаций не произошло, а в некоторых произошло 3 и более событий. — *Прим. ред.*

На с. 202 показана матрица PAM250. Цены выравниваний выражены в логарифмических величинах:

$$\text{Цена мутации } i \leftrightarrow j = \lg \frac{\text{наблюдаемая скорость мутаций } i \leftrightarrow j}{\text{скорость мутаций, исходя из частот встречаемости аминокилот}}$$

Во избежание дробных чисел все величины умножены на 10 и округлены. Величины в ячейках матрицы отражают вероятность мутации. Значение +2 (например, за замену C↔S) предполагает то, что в сравниваемых последовательностях данная мутация происходит с вероятностью в 1.6 раз выше, чем в случайной выборке. Расчеты производятся следующим образом: 2 было получено после умножения на 10, поэитому на самом деле это 0.2, т. е. lg относительной вероятности мутации. Таким образом, вероятность мутации равна $10^{0.2} = 1.6$.

Вероятность двух независимых мутаций — это произведение их вероятностей. При использовании логарифмов веса выравниваний можно складывать.

Матрицы BLOSUM

С. Хеникофф и Дж. Г. Хеникофф разработали семейство матриц BLOSUM для расчета весов для замен в аминокислотных последовательностях при выравнивании. Они ставили своей целью использовать вместо матрицы Дэйхофф (Dayhoff) другую, более эффективную для идентификации удаленных связей, с использованием гораздо большего объема информации, который стал доступным со времени работ Дэйхофф.

Матрица BLOSUM основана на базе данных выровненных последовательностей белков BLOCKS; отсюда и ее название — BLOcks SUbstitution Matrix. Из семейства родственных белков, выравниваемых без пропусков, Хэникофф и Хэникофф вычислили отношение числа рассмотренных пар аминокислотных остатков на любой позиции, к числу пар, ожидаемых для всех аминокислотных последовательностей. Как и в матрице Дэйхофф, результат представлен в виде логарифмов частот замен. Чтобы избежать излишнего влияния родственных последовательностей, Хэникофф заменили группы белков, которые имеют идентичность последовательности выше, чем пороговая, либо одним представителем, либо их средневзвешенным значением. Порог в 62% создает часто используемую матрицу замещений BLOSUM62 (см. с. 202). Она предлагается в качестве варианта всеми программами, и большинством из них эта матрица используется по умолчанию. Матрицы BLOSUM заменили матрицу Дэйхофф в широком спектре приложений.

Взвешивание вставок/делеций

Для формирования полной схемы определения весов выравнивания нам нужен, кроме матрицы замен, способ оценки пропусков. Насколько важны вставки и делеции, по сравнению с заменами?

Матрицы замен, использованные для расчета подобия аминокислотной последовательности. Значения расположены в алфавитном порядке в соответствии с трехбуквенным наименованием аминокислот. Показаны только нижние треугольники матриц, поскольку вероятности замещений считаются симметричными. (Не потому, что мы уверены, что вероятность любой замены равна вероятности обратного перехода, но потому, что мы не можем определить разницу между двумя этими событиями.)

Матрица Дэйхофф PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala (A)	2																			
Arg (R)	−2	6																		
Asn (N)	0	0	2																	
Asp (D)	0	−1	2	4																
Cys (C)	−2	−4	−4	−5	12															
Gln (Q)	0	1	1	2	−5	4														
Glu (E)	0	−1	1	3	−5	2	4													
Gly (G)	1	−3	0	1	−3	−1	0	5												
His (H)	−1	2	2	1	−3	3	1	−2	6											
Ile (I)	−1	−2	−2	−2	−2	−2	−2	−3	−2	−5										
Leu (L)	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6									
Lys (K)	−1	3	1	0	−5	1	0	−2	0	−2	−3	5								
Met (M)	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6							
Phe (F)	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9						
Pro (P)	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6					
Ser (S)	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2				
Thr (T)	1	−2	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3			
Trp (W)	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17		
Tyr (Y)	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	
Val (V)	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4

Матрица BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala (A)	4																			
Arg (R)	−1	5																		
Asn (N)	−2	0	6																	
Asp (D)	−2	−2	1	6																
Cys (C)	0	−3	−3	−3	9															
Gln (Q)	−1	1	0	0	−3	5														
Glu (E)	−1	0	0	2	−4	2	5													
Gly (G)	0	−2	0	−1	−3	−2	−2	6												
His (H)	−2	0	1	−1	−3	0	0	−2	8											
Ile (I)	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu (L)	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys (K)	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met (M)	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe (F)	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro (P)	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser (S)	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr (T)	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp (W)	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr (Y)	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val (V)	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Следует различать внесение одиночного пропуска

```
aaagaaa
aaa-aaa
```

и протяженного пропуска[1]):

```
aaaggggaaa
aaa----aaa
```

Для выравнивания последовательностей ДНК программа CLUSTAL-W рекомендует использовать матрицы идентичности для замен (+1 для совпадения, 0 для несовпадения) и штрафы: 10 для открытия делеции и 0.1 для продолжения делеции. Для выравнивания аминокислотных последовательностей рекомендуется использовать матрицу BLOSUM62 для замен, и штрафные значения: 11 для открытия делеции и 1 для продолжения делеции.

Расчет выравнивания для двух последовательностей

Теперь, когда у нас есть схема определения веса, мы можем применить ее для поиска оптимальных выравниваний — мы ищем выравнивание, которое достигает максимального количества баллов. Знаменитый алгоритм для определения глобальных оптимальных выравниваний двух последовательностей основан на математическом методе, называемом динамическим программированием (детали его описаны далее). Этот алгоритм был чрезвычайно важен в молекулярной биологии. Из нескольких существенных особенностей метода выделим следующие две:

- положительная особенность состоит в том, что метод гарантирует *глобальный* оптимум: *наилучший* результат выравнивания при заданном наборе параметров — матрице замещений и штрафных значениях для пропусков — без каких-либо приближений.
- отрицательная особенность состоит в том, что многие выравнивания могут привести к оптимальному, но одному и тому же числу баллов. И совершенно не обязательно, что хотя бы одно из них имеет отношение к биологически корректному выравниванию. Например, при сравнении последовательностей α- и β-цепей гемоглобина цыпленка В. Фитч и Т. Смит нашли 17 выравниваний, каждое давало одинаковое оптимальное число баллов, из которых корректно одно (на основании сравнения структур, «суда последней инстанции»). Существует 1317 выравниваний, которые дают число баллов в пределах 5% от оптимума.

Есть еще отрицательная особенность, связанная с компьютерным обеспечением: время, требуемое для выравнивания двух последовательностей длиной n и m пропорционально $m \times n$, потому что это — размер редактируемой матрицы,

[1]) Более четко — штраф за делецию устанавливается по формуле $s = \alpha + \beta \cdot l$, где α — штраф за открытие делеции, β — штраф за продолжение делеции, l — размер делеции. Такая схема в литературе называется аффинным штрафом за делецию; в данном примере $\alpha = 10$, $\beta = 0.1$. — *Прим. ред.*

которая должна быть заполнена[1]. Следовательно, метод динамического программирования не подходит при поиске соответствия для пробной последовательности в полной базе данных последовательностей, и еще меньше подходит для выравниваний «все-против-всех». Проблема поиска в базе данных — это на самом деле проблема поиска соответствия интересующей нас последовательности с очень длинной последовательностью, длина которой равна всей базе данных.

Вариации и обобщения

Вариации описанного метода применяются к трем связанным с выравниванием вопросам: поиск выравнивания одной полной последовательности с другой полной последовательностью, выравнивание фрагмента (заранее неизвестного — лишь бы вес выравнивания был побольше. — *Прим. ред.*) одной последовательности с полной другой последовательностью, фрагмента одной последовательности с фрагментом другой последовательности (см. с. 41). Алгоритм глобального выравнивания был впервые применен для выравнивания биологической последовательности С. Б. Нидлманом и К. Д. Вуншем. Т. Смит и М. Ватерман модифицировали его для идентификации локальных совпадений.

Приближенные методы для быстрого поиска в базах данных

По общепринятой практике гены из нового генома сравниваются со всеми последовательностями из базы данных. Приближенные методы могут работать хорошо и быстро для поиска похожих последовательностей, но при поиске очень отдаленных связей они работают хуже, чем точные методы. На практике они дают удовлетворительный результат во многих случаях, когда последовательность запроса достаточно похожа на одну или несколько последовательностей в базе данных, и поэтому имеет смысл пробовать их первыми.

Типичный приближенный подход выбирает некое малое целое значение k и находит все подслова длиной k остатков (k-tuple) в последовательности поиска, которые появляется также в какой-либо последовательности в базе данных. Последовательность-кандидат — это последовательность в базе данных, содержащая большое число подходящих подслов длины k, которые встретились в последовательности поиска и в последовательности-кандидате. Затем для отобранного набора последовательностей-кандидатов производятся вычисления приближенного оптимального выравнивания с временным и пространственным ограничением, которое предусматривали проходы по матрице в пределах диагоналей, содержащих многочисленные совпадающие наборы k. Существует несколько вариаций на эту тему.

---

[1] Отметим, что построение точечной матрицы сходства требует того же времени, а использование наивного алгоритма подавления шума, который реализован в представленной здесь программе, требует времени $m \times n \times w$, где w — ширина окна. — *Прим. ред.*

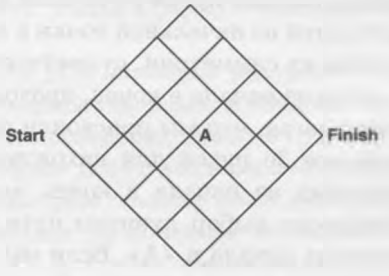
Алгоритм динамического программирования для построения оптимального парного выравнивания последовательностей[1)]

Схема, содержащая все возможные выравнивания, может быть построена в виде матрицы наподобие той, которая используется для изображения точечной матрицы сходства. Остатки одной последовательности индексируют строки, а остатки другой последовательности — столбцы. Любой путь по матрице, начинающийся в левом верхнем углу и заканчивающийся в правом нижнем, соответствует одному выравниванию (см. рис. 4.1). Задача — найти путь наименьшего веса, и трудность состоит в том, что таких путей нужно рассмотреть очень много.

В качестве примера представьте, что вы хотите приехать из Мальмо (на юге Швеции) в Тромсо (на севере Норвегии). Ваш путь будет состоять из нескольких участков, проходящих через ряд промежуточных городов. Существует множество различных комбинаций участков, которые в результате будут давать полный путь, от начала до конца.

Компьютерный подход к нахождению оптимального пути начинается с присвоения цены каждому отдельному возможному участку пути. Эта цена учитывает не простые финансовые затраты, а общую оценку относительных преимуществ разных фрагментов пути. Само расстояние, несомненно, учитывается в цене стоимости, но и другие факторы, такие как качество дороги и возможность ознакомиться с достопримечательностями, тоже вносят вклад в эту цену. Для любого выбранного пути общая стоимость путешествия является суммой цен отдельных фрагментов. Очевидно, что неэффективно повторять какой-то этап пути или посещать какой-то город дважды, следовательно, мы должны согласиться с тем, что любая промежуточная остановка будет севернее предыдущей. Такой формализм придуман скорее с целью минимизации стоимости, чем увеличения качества пути; для наших целей оба подхода равносильны. Алгоритм может рассмотреть возможные комбинации для определения оптимального конечного пути.

На этом рисунке представлена схема, иллюстрирующая задачу:



[1)] Этот материал можно пропустить при первом чтении. Сомневающиеся читатели могут ознакомиться с замечаниями А. М. Леска (1988), «TATA for now...», *Trends Biochem. Sci.*, 13, 410).

Рис. 4.3. Возможные пути из Мальмо в Тромсо. Как вы можете определить оптимальный путь? (© Коллинз. Перепечатано с разрешения ХарперКоллинз Паблишерз).

Для того чтобы уловить ключевую идею динамического программирования, во-первых, подумайте: сколько путей от начала до конца проходит через точку «A»? Существует 6 путей из начальной точки в точку «A» (запишите их все). Таким образом, исходя из симметрии, существует 6 путей из точки «A» в конец, а общее число путей из начала в конец, проходящих через точку «A», равно 36 (почему?). Предполагая, что мы присвоили цены отдельным этапам, нужно ли нам проверять все 36 путей для нахождения пути минимальной стоимости, который проходит из начала в конец через точку «A»? Нет — здесь есть важное наблюдение: выбор лучшего пути из точки «A» в конец не зависит от выбора пути из начала в «A». Если мы определяем лучший из 6 путей, ведущих от начальной точки в точку «A», а также лучший из путей от «A» к концу, лучший путь от начала к концу, проходящий через «A», будет определяться как лучший путь от начала к «A» и следующий за ним лучший путь от «A» к конечной точке. В этом случае нужно рассматривать уже не более 12 путей, проходящих через точку «A».

Очень сильное упрощение возможно путем систематического разделения проблемы на все более мелкие части. Метод динамического программирования по нахождению оптимального пути в матрице основан именно на этой идее.

Утверждение относительно проблемы построения оптимального выравнивания и ее решения методом динамического программирования состоит в следующем: имея две строки символов, возможно, неодинаковой длины: $A = a_1 a_2 \cdots a_n$ и $B = b_1 b_2 \cdots b_m$, где каждый элемент a_i и b_j — буква некого алфавита A, рассмотрим последовательности, как операции редактирования, конвертирующие последовательности A и B в общую последовательность. Отдельные операции редактирования включают в себя:

Замена b_j на a_i отображается записью (a_i, b_j).

Делеция a_i из последовательности A отображается как (a_i, ϕ).

Делеция b_j из последовательности B отображается как (ϕ, b_j).

Если мы расширим алфавит, включив в него нулевой символ ϕ: $\mathcal{A}^+ = \mathcal{A} \cup \{\phi\}$, последовательность операций редактирования будет набором упорядоченных пар (x, y), где $x, y \in \mathcal{A}^+$.

Функция стоимости d определяется операциями редактирования: $d(a_i, b_j)$ = цена мутации в выравнивании, где позиция i последовательности A соответствует позиции j последовательности B, и мутация производит замену $a_i \leftrightarrow b_j$. $d(a_i, \phi)$ или $d(\phi, b_j)$ = стоимость делеции или вставки. Определим расстояние минимального веса между последовательностями A и B как

$$D(A, B) = \min_{A \to B} \Sigma d(x, y)$$

где $x, y \in \mathcal{A}^+$ и минимум берется из всей последовательности операций редактирования, которые конвертируют A и B в общую последовательность.

Проблема состоит в том, чтобы найти $D(A, B)$ и одно или более выравниваний, которые ей соответствуют.

Алгоритм, решающий эту проблему за время $\mathcal{O}(mn)$ создает матрицу $\mathcal{D}(i, j)$, $i = 0, \ldots n$; $j = 0, \ldots m$, такую, что $\mathcal{D}(i, j)$ — минимальное расстояние между строками, которые состоят из первых i символов строки A и первых j символов строки B. Тогда $\mathcal{D}(n, m)$ — искомое минимальное расстояние $D(A, B)$.

Алгоритм считает $\mathcal{D}(i, j)$ с помощью рекурсии. Значение $\mathcal{D}(i, j)$ соответствует преобразованию первоначальных последовательностей $A_i = a_1 a_2 \cdots a_i$ и $B_j = b_1 b_2 \cdots b_j$ в общую последовательность путем L операций редактирования $S_k, k = 1, \ldots, L$, которые могут быть рассмотрены в порядке возрастания позиций в строках. Рассмотрим *отмену* последней операции редактирования. Получающаяся в результате укороченная последовательность операций редактирования S_k, $k = 1, \ldots, L - 1$, представляет собой последовательность операций конвертирования подстроки A_i и подстроки B_j в общий результат. Более того, это должна быть *оптимальная* последовательность операций редактирования для этих подстрок: если какая-нибудь другая последовательность операций S'_k имеет меньшую цену редактирования для этих подстрок,

то и цена редактирования для полной задачи будет меньшей. Таким образом, должен быть рекурсивный метод расчета $\mathcal{D}(i,j)$.

Установим соответствие между индивидуальными операциями редактирования и шагами между смежными ячейками матрицы (см. рис. 4.1):

$(i-1,j-1) \to (i,j)$	соответствует замене $a_i \to b_j$.
$(i-1,j) \to (i,j)$	соответствует делеции a_i из A.
$(i,j-1) \to (i,j)$	соответствует вставке b_j в A в позицию i.

Последовательности операций редактирования соответствуют ступенчатому пути по матрице

$$(i_0,j_0) = (0,0) \to (i_1,j_1) \to \cdots (n,m)$$

где $0 \leqslant i_{k+1} - i_k \leqslant 1$, (для $0 \leqslant k \leqslant n-1$), $0 \leqslant j_{k+1} - j_k \leqslant 1$ (для $0 \leqslant k \leqslant m-1$.) Учитывая возможные последовательности операций редактирования и соответствующие шаги по матрице, предшественник оптимальной цепочки операций редактирования, ведущих от $(0,0)$ к (i,j), где $i,j > 0$, должен быть оптимальной последовательностью операций редактирования, ведущих в одну из ячеек $(i-1,j),(i-1,j-1)$, или $(i,j-1)$; и, соответственно, $\mathcal{D}(i,j)$ должна зависеть только от значений $\mathcal{D}(i-1,j),\mathcal{D}(i-1,j-1)$ and $\mathcal{D}(i,j-1)$ (в согласовании, конечно, с параметризацией, установленной функцией стоимости d).

Таким образом, алгоритм следующий:

Вычислить матрицу $\mathcal{D}$ размерностью $(m+1) \times (n+1)$, применяя

(1) начальные условия на верхний ряд и левую колонку:

$$\mathcal{D}(i,0) = \sum_{k=0}^{i} d(a_k,\phi)$$

$$\mathcal{D}(0,j) = \sum_{k=0}^{j} d(\phi,b_k)$$

Эти значения влияют на штраф за открытие делеции при несовпадающих остатках в начале любой последовательности и тогда

(2) рекуррентные соотношения:

$$\mathcal{D}(i,j) = \min\{\mathcal{D}(i-1,j)+d(a_i,\phi), \mathcal{D}(i-1,j-1)+d(a_i,b_j), \mathcal{D}(i,j-1)+d(\phi,b_j)\}$$

для $i = 1,\ldots,n; j = 1,\ldots,m$. Это означает, что мы учитываем все три возможных шага для $\mathcal{D}(i,j)$:

Операция	Общая цена
Вставка пропуска в последовательность A	$\mathcal{D}(i-1,j) + d(a_i,\phi)$
Замена $a_i \leftrightarrow b_j$	$\mathcal{D}(i-1,j-1) + d(a_i,b_j)$
Вставка пропуска в последовательность B	$\mathcal{D}(i,j-1) + d(\phi,b_j)$

Следует выбрать минимальное значение. Для каждого значения ячейки определено не только значение $\mathcal{D}(i,j)$, но и стрелка назад к (одной и более) ячейке (-ам) $(i-1,j),(i-1,j-1)$ или $(i,j-1)$ определена операцией минимизации. Заметьте, что более одного предшественника могут давать одно и то же значение.

Когда расчеты закончены, $\mathcal{D}(n,m)$ — это оптимальное расстояние $D(A,B)$. Выравнивание, соответствующее последовательности операций редактирования, отмеченной стрелками, может быть восстановлено прохождением обратного пути по матрице от (n,m) к $(0,0)$. Это выравнивание, соответствующее минимальному расстоянию $D(A,B) = \mathcal{D}(n,m)$, может быть не единственным.

ПРИМЕР 4.6.

Выровняем[1)] строки $A =$ `ggaatgg` и $B =$ `atg` в соответствии с простой системой параметров: совпадение $= 0$, несовпадение $= 20$, вставка или делеция $= 25$.

Здесь приведено положение алгоритма после инициализации верхнего ряда и самой левой колонки (выделены курсивом), и элемент во втором ряду и второй колонке равен 20 (выделен жирным шрифтом).

	ϕ	a	t	g
ϕ	*0*	*25*	*50*	*75*
g	*25*	**20**		
g	*50*			
a	*75*			
a	*100*			
t	*125*			
g	*150*			
g	*175*			

Значение **20** было выбрано как минимум из $25+25$ (вертикальный шаг, или вставка делеции в строку `atg`), $0+20$ (замена `a`↔`g`) и $25+25$ (горизонтальный шаг, или вставка делеции в строку `ggaatgg`). В связи с тем, что замена (шаг по диагонали) предполагала минимальное значение, ячейка, содержащая 0 в верхнем левом углу матрицы, является предшественником ячейки, в которую мы только что ввели 20. (Если два или даже три возможных шага порождают одно и то же значение, то у результирующей ячейки будет много предшественников.)

[1)] К сожалению, при описании алгоритма автор минимизировал редакционное расстояние, а в приведенном примере автор максимизирует сходство. — *Прим. ред.*

Матрица после завершения расчетов имеет следующий вид:

	ϕ		a		t		g
ϕ	0	←	25	←	50	←	75
	↑	↖		↖		↖	
g	25		20	←	45		50
	↑	↖	↑	↖		↖	
g	50		45		40		45
	↑	↖		↖	↑	↖	
a	75		50		65		60
	↑	↖	↑	↖		↖	↑
a	100		75		70		85
	↑		↑	↖		↖	
t	125		100		75		90
	↑		↑		↑	↖	
g	150		125		100		75
	↑		↑		↑	↖	↑
g	175		150		125		100

Она включает в себя информацию о пути снизу вверх в виде стрелок от каждой ячейки к его предшественнику(-ам). Для некоторых приложений нам может понадобиться только значение $D(A, B)$, но не выравнивание; в таком случае нет необходимости сохранять стрелки. Выделенные жирным шрифтом стрелки намечают путь оптимального выравнивания, цепь предшественников от нижнего правого до верхнего левого. В некоторых случаях одна ячейка может иметь двух предшественников. Они отвечают различным выравниваниям с одним и тем же счетом.

В двух ячейках путь разветвляется. Таким образом, получаются четыре оптимальных выравнивания с одинаковым счетом.

```
ggaatgg    ggaatgg    ggaatgg    ggaatgg
---atg-    ---at-g    --a-tg-    --a-t-g
```

Система, оценивающая вес вставки, приписывает меньший штраф за продолжение делеции, чем за ее открытие. Поэтому первые две последовательности имеют больший счет, чем остальные. Хотя более сложные системы оценки штрафов за вставки/делеции требуют более сложных рекуррентных соотношений для заполнения матрицы[1)].

Этот алгоритм определяет оптимальное глобальное выравнивание двух последовательностей. Он не подходит для поиска локальных участков высокой консервативности в двух последовательностях или исследования выравнивания длинной последовательности с коротким фрагментом, потому что алгоритм устанавливает штрафы за ставки перед и после сходных участков. Метод Т. Смита и М. Ватермана решает эту проблему. Их модификация основного алгоритма динамического программирования находит

[1)] Строго говоря, описанный здесь алгоритм применим только к линейным штрафам за делеции, когда вес инициации делеции равен нулю. Для аффинных штрафов надо применять более сложный вариант динамического программирования. Отбор выравниваний, учитывающих штраф за инициацию *post factum*, некорректен. — *Прим. ред.*

оптимальные локальные выравнивания, выбирая подстроки из обоих последовательностей, которые более всего похожи друг на друга. Эти изменения затрагивают[1)]

1. *Инициализацию матрицы.* Установка значений верхнего ряда и левой колонки. В методе Смита—Ватермана верхний ряд и левая колонка имеют нулевые значения. В результате любая последовательность может «скользить» вдоль другой перед началом выравнивания без прибавления штрафа за делеции остатков, оставленных позади.
2. *Заполнение матрицы.* В примере с глобальным выравниванием на каждом шагу выбор делался между совпадением, вставкой или делецией, даже если ни одна из этих возможностей не подходит и даже если цепь из неподходящих событий понижает счет по пути, содержащему хорошо выравниваемый участок. Метод Смита—Ватермана предлагает еще один выход: конец участка выравнивается.
3. *Счет и путь выравнивания.* Вес глобального выравнивания это число в ячейке матрицы в нижнем правом углу. В методе Смита—Ватермана это оптимальное посчитанное значение, где бы оно не появилось в матрице. Для глобального выравнивания путь для определения фактического выравнивания начинается в нижней правой ячейке. В методе Смита—Ватермана оно начинается в ячейке, содержащей оптимальное значение, и продолжается назад только до тех пор, пока продолжается участок локального сходства.

Метод Смита—Ватермана предлагает единственный оптимальный путь для нашего примера:

```
ggaatgg
  atg
```

Заметьте, что перед и после выровненного участка нет делеций.

(Пример взят из: Tyler, E. C., Horton, M. R. and Krause, P. R. (1991) "A review of algorithms for molecular sequence comparison," Comp. Biomed. Res. 24, 72-96.)

Значимость выравниваний

Допустим, выравнивание показывает интересное сходство двух последовательностей. Несет ли это сходство смысл или же оно имеет случайное происхождение (мы уже поднимали этот вопрос в гл. 1)? Для некоторых простых явлений, таких как подбрасывание монеты или бросание костей, возможно подсчитать

[1)] Для поиска локального сходства можно использовать только постановку задачи о *максимизации сходства*, но нельзя использовать постановку задачи о *минимизации редакционного расстояния*. На самом деле, единственное отличие алгоритма Смита—Ватермана — это запрет на отрицательные веса. Всякий раз, когда наилучший вес в ячейке матрицы оказывается отрицательным, в ячейку записывается 0 и ставится метка, что отсюда дальше пути нет. — *Прим. ред.*

ожидаемое распределение результатов, а также вероятность любого отдельного результата. В случае последовательностей определение совокупности, из которой выбрано выравнивание, не является тривиальным. Например, генерация контрольных случайных строк нуклеотидов или аминокислот не учитывает влияние неслучайных явлений.

Практический подход к данной проблеме заключается в следующем: если значения выравнивания не выше ожидаемого значения, полученного путем *случайной перетасовки* последовательности, то, вероятно, такое сходство выравниваемых последовательностей имеет случайное происхождение, нежели закономерное. Мы можем многократно перетасовать одну из последовательностей, выравнивая каждый раз со второй, фиксированной, последовательностью, собирая распределение получающихся значений (типичное распределение весов показано на рис. 4.4). При поиске по базе данных для измерения статистики лучше всего использовать группу результатов, полученных из всей базы данных[1)].

Ясно, что если рандомизированные последовательности дают такой же вес, как и исходные, то, скорее всего, выравнивание не несет биологического смысла. Мы можем измерить среднее значение и стандартное отклонение весов выравниваний рандомизированных последовательностей и проверить, является ли вес исходных последовательностей необычно высоким. Величина Z-score отражает отклонение первоначального результата от совокупности:

$$Z\text{-score} = \frac{\text{вес} - \text{среднее}}{\text{стандартное отклонение}}$$

Значение Z-score, равное 0, означает, что обнаруженное сходство не лучше среднего количества случайных перестановок последовательности и могло бы, с высокой вероятностью, получиться случайно. Другие значения, используемые в качестве мер значимости: P — вероятность того, что найденное сходство может быть получено случайным образом, и для поиска по базам данных E — ожидаемое количество сходств, аналогичных или лучших, чем исходное, которые могли бы быть получены случайным образом в базе данных заданного размера (см. с. 213).

Существует множество правил интерпретации процента идентичных остатков в оптимальном выравнивании. Если два белка содержат больше 45% идентичных остатков в их оптимальном выравнивании, то эти белки имеют очень похожие структуры и, скорее всего, общую или, по крайней мере,

[1)] Статистическая значимость весов выравниваний — очень важная характеристика. При этом надо четко понимать, что статистическая значимость всегда оценивается относительно *случайной модели*, а эти модели могут быть разными и иметь разную степень адекватности реальности. Тест со случайной перестановкой последовательности соответствует Бернуллиевской модели случайной последовательности как, впрочем, и другие, приведенные здесь оценки. Эта модель учитывает только частоты появления аминокислотных остатков, но не учитывает, например, предпочтения следования аминокислотных остатков, что соответствует Марковской модели. Эта модель также не отражает в достаточной мере статистические свойства реальных аминокислотных последовательностей. Поэтому, все приведенные здесь оценки носят только ориентировочный характер, и отнюдь не означают реальную статистическую значимость. — *Прим. ред.*

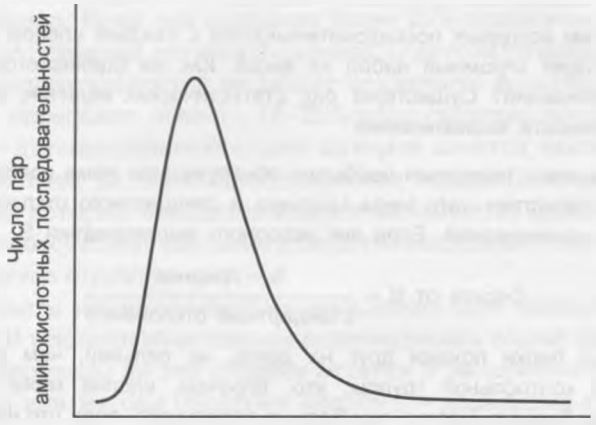


Рис. 4.4. Распределение оптимальных локальных выравниваний между парами случайных аминокислотных последовательностей одинаковой длины имеет экстремум. Для любого веса x вероятность найти вес $\geqslant x$ можно рассчитать: $P(\text{score} \geqslant x) = 1 - \exp(-Ke^{-\lambda x})$, где K и λ — параметры, связанные с расположением максимума и шириной распределения. Обратите внимание на длинный хвост в правой части графика. Он означает, что несколько стандартных отклонений значений выше среднего значения имеют большую вероятность случайного возникновения (т. е. они менее значимы), чем значения, подчиняющиеся нормальному распределению

Как не погореть, играя с совпадениями

Парные выравнивания и системы поиска в базах данных часто выявляют незначительные, но весьма забавные сходства последовательностей. Но как нам решить, имеем ли мы дело с настоящими гомологиями? Статистика не может ответить на этот биологический вопрос прямо, но может указать вероятность, с которой наша последовательность окажется столь же схожей с наблюдаемой, сколь и с последовательностью, не имеющей к ней никакого отношения, чисто случайную. Для этого нам нужно сравнить наши результаты с выравниваниями тех же последовательностей с большим количеством произвольных белков. Последовательности из этого контрольного набора должны в общих чертах быть похожими на наши исходные последовательности, но там должно быть мало родственных им последовательностей. Только если наблюдаемое совпадение нехарактерно для выравниваний из контрольного набора, мы можем считать его значимым.

И откуда нам лучше взять этот набор последовательностей, с которым мы должны сравнивать наше выравнивание? В случае с парным выравниванием мы можем взять одну из двух последовательностей, создать множество ее перемешанных копий, используя генератор случайных чисел, и выровнять их всех по очереди со второй последовательностью. В случае с поиском по базам данных — непосредственно сама база данных играет роль такого набора.

Выравниваниям исходных последовательностей с каждым членом контрольного набора соответствует огромный набор их весов. Как же оценивается вес нашего исходного выравнивания? Существует ряд статистических величин, используемых для оценки значимости выравнивания.

- *Z*-score показывает, насколько необычно обнаруженное нами совпадение, т. е. в терминах статистики — это мера среднего и стандартного отклонения весов полученных выравниваний. Если вес исходного выравнивания S,

$$Z\text{-score от S} = \frac{\text{S} - \text{среднее}}{\text{стандартное отклонение}}$$

  $Z = 0$ — наши белки похожи друг на друга, не сильней, чем (в среднем) на белки из контрольной группы, что, впрочем, вполне может произойти случайно. Чем больше *Z*-score, тем больше вероятность того, что наблюдаемое выравнивание появилось неслучайно. Опыт показывает, что *Z*-score $\geqslant 5$ уже говорит о значимости исходного выравнивания.
- Многие программы выдают величины *P* (*P*-value) — вероятности того, что выравнивание не лучше, чем случайное. Взаимоотношение *Z*-score и *P* зависит от распределения весов контрольных выравниваний, которое не соответствует нормальному распределению.

  Вот грубые ориентиры для интерпретации значений *P*-value:

$P \leqslant 10^{-100}$	точное совпадение
P между 10^{-100}–10^{-50}	последовательности почти идентичны, например, аллели или полиморфизмы
P между 10^{-50}–10^{-10}	близкородственные последовательности; гомология очевидна
P между 10^{-5}–10^{-1}	обычно дальнеродственные последовательности
$P > 10^{-1}$	по-видимому, соответствие незначимо

- Что касается систем поиска по базам данных, некоторые программы (в том числе и PSI-BLAST) указывают *E*-value. *E*-value выравнивания — это ожидаемое количество последовательностей, которые бы имели *Z*-score такой же (или лучше), как если бы мы в качестве запроса дали программе случайную последовательность. *E*-value находят, умножая значение *P* на размер базы данных, в которой производят поиск. (На самом деле это не совсем так. *P* — вероятность увидеть хотя бы одно выравнивание такого качества при сравнении с банком. Поэтому ничего умножать не надо, а надо понимать, что при малых *P* эти величины примерно равны.. — *Прим. ред.*).

  Вот как приблизительно следует интерпретировать полученные *E*-value:

$E \leqslant 0.02$	вероятно, последовательности являются гомологами
E между 0.02 и 1	гомология не очевидна
$E > 1$	следует ожидать, что это случайное совпадение

Надо отметить, что статистика — полезна и необходима, но не может заменить здравый смысл и тщательный и аккуратный анализ результатов, особенно многообещающих!

сходную функцию. Если они содержат более 25% идентичных остатков, они, вероятно, имеют сходный паттерн фолдинга. С другой стороны, низкая степень сходства последовательностей не может исключить возможность гомологии. Р. Ф. Дулитл определил область 18–25%-ного сходства последовательностей как «область двусмысленности», для которой хочется высказать предположение о гомологии, но такое предположение может быть опасным. Парные выравнивания, которые находятся ниже этой области, мало, что могут означать. При этом отсутствие значимого сходства последовательностей не мешают наличию сходства структур.

Хотя «область неоднозначности» ненадежна для выводов, мы и тут не беспомощны. В решении об истинном родстве важна также «текстура» выравнивания: изолированы ли эти сходные остатки и распределены по всей последовательности или же они образуют айсберги — локальные участки высокого сходства (другой термин Дулитла), которые могут соответствовать общему активному сайту? Нам также может понадобиться и другая информация — об общих лигандах или функциях. Конечно, если пространственные структуры известны, то мы можем проверить их сходство непосредственно.

Вот несколько примеров:

- Миоглобин кашалота и леггемоглобин люпина имеют 15% идентичных остатков в оптимальном выравнивании. Это даже ниже определенной Дулитлом области неоднозначности. Но мы также знаем, что обе молекулы имеют сходные трехмерные структуры, содержат гемовые простетические группы и связывают кислород. Они действительно являются удаленными гомологами.
- Последовательности N- и C-концевых частей роданеза имеют 11% идентичных остатков в оптимальном выравнивании. Если бы они возникли в независимых белках, нельзя было бы судить об их родстве, исходя лишь из последовательностей. Однако такая ситуация в одном белке дает основание полагать, что они произошли путем дупликации и дивергенции генов. Очевидное сходство их структур подтверждает их родство.
- Чтобы не допустить опрометчивых действий, рассмотрим две протеазы: химотрипсин и субтилизин. Их последовательности похожи на 12%. Эти ферменты выполняют сходную функцию и в активном центре несут три характерных для них остатка. Тем не менее они имеют разную пространственную укладку и не родственны. Их общая функция и механизм — пример конвергентной эволюции. Отсюда предупреждение: не стоит отстаивать родственную связь между белками с непохожими последовательностями, основываясь только на схожести функций и механизмов!

Множественное выравнивание последовательностей

«Сама по себе аминокислотная последовательность нам ни о чем не говорит, вот две гомологичные последовательности уже что-то нашептывают, ну, а когда их много, они орут вовсю». В природе белковая последовательность сама по себе содержит всю необходимую информацию, определяющую пространствен-

ную укладку белка. Так как же выравнивание нескольких последовательностей сделает эту информацию более вразумительной? Выравнивание выявляет и демонстрирует нам принцип расположения консервативных аминокислотных остатков, что позволяет более достоверно установить далекие родственные отношения. Программы, предсказывающие пространственные структуры белков, дают более правдоподобные результаты, когда основываются на множественных выравниваниях, а не на одной последовательности.

Просмотр и правка вручную множества множественных выравниваний — одна из наиболее полезных форм деятельности, которыми может заняться молекулярных биолог вне лаборатории. Даже не думайте, что поймете в них чего-нибудь без выделения цветом разных типов аминокислотных остатков. Вот один из возможных разумных способов раскраски:

Цвет	Тип остатка	Аминокислоты
Желтый	Маленькие неполярные остатки	Gly, Ala, Ser, Thr
Зеленый	Гидрофобные	Cys, Val, Ile, Leu, Pro, Phe, Tyr, Met, Trp
Фиолетовый	Полярные	Asn, Gln, His
Красный	Отрицательно заряженные	Asp, Glu
Синий	Положительно заряженные	Lys, Arg

Чтобы множественное выравнивание было информативным, оно должно содержать разные по эволюционному расстоянию последовательности. Если все последовательности чересчур близкие, то информация, которую они несут, избыточно дублируется, и из этого выравнивания можно извлечь немногие выводы. А если все последовательности далеки друг от друга, трудно будет построить аккуратное выравнивание (кроме тех белков, для которых известны структуры), и в таком случае достоверность результатов и сделанных на их основе выводов оказывается под вопросом. В идеале оно должно содержать широкий набор белков разного уровня сходства, включающий далеко отстоящие экземпляры среди множества близких гомологов.

Связь множественных выравниваний последовательностей и структур

Тиоредоксины — ферменты, которые можно обнаружить во всех клетках. Они участвуют в широкой области биологических процессов, включая клеточное деление, свертывание крови, прорастание семени, деградацию инсулина, восстановление окислительных повреждений и ферментативную регуляцию. Обычный механизм этих процессов представляет собой восстановление дисульфидных белковых связей.

Цветная иллюстрация VII показывает множественное выравнивание последовательностей 16 тиоредоксинов. Структура тиоредоксинов *E. coli* состоит из центральной β-поверхности, составленной из пяти листов. Данная поверхность выстлана с каждой стороны α-спиралями. Эти спирали и листы идентифицируются символами α и β.

Другие тиоредоксины, предположительно, повторяют большую часть, однако не всю, вторичной структуры ферментов *E. coli*. На иллюстрации VII также показано суммарное выравнивание, на котором буквы различных размеров обозначают различное соотношение аминокислот. (Т. Шнейдер и М. Стивенс смоделировали последовательности Лого; данный пример был произведен с использованием сервера С. Е. Бренера, `http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi`).

Структурные и функциональные особенности тиоредоксинов, которые можно (мы надеемся) идентифицировать из множественных выравниваний последовательностей:

- *Наиболее высококонсервативные участки, возможно, отвечающие за активный сайт связывания.* Дисульфидный мостик между 32 и 35 остатками в тиоредоксине *E. coli* является частью мотива WCGPC[K или R], консервативного по всему семейству. Другие консервативные участки в последовательностях, включая пары PT в позициях 76–77 и GA в позициях 92–93, вовлечены в процесс связывания с субстратом.
- *Участки, богатые вставками и делециями, возможно, отвечающие за образование поверхностных петель. Позиция, содержащая консервативный Gly или Pro, возможно, отвечает за поворот.* Повороты, ассоциированные со вставками и делециями, возникают в позициях 9, 20, 60 и 95. Консервативный глицин в позиции 92 тиоредоксина *E. coli* в самом деле является частью мотива поворота. Это необычная конформация основной цепи, именно та, которая легко достижима только для глицина (см. гл. 5). Консервативный пролин в позиции 76 тиоредоксина *E. coli* также ассоциируется с поворотом. Это очередная необычная конформация основной цепи, в данном случае доступная исключительно пролину.
- *Консервативный паттерн гидрофобности с периодом 2 (как и для других остатков) — с более вариабельными переходными остатками и, включая гидрофильные остатки, предполагает β-тяж на поверхности.* Данный паттерн наблюдается в β-тяже между 50-м и 60-м остатками.
- *Консервативный паттерн гидрофобности с периодом ~ 4 предполагает наличие спирали.* Данный паттерн наблюдается на участке спирали между 40-м и 49-м остатками.

Тиоредоксины — члены суперсемейства, включающего в себя множество более удаленно-родственных гомологов: глутаредоксин (донор водорода для рибонуклеотидной редукции в синтезе ДНК), белковая дисульфидная изомераза (которая катализирует обмен несочетающихся дисульфидных мостиков в белковых фолдингах), фосдуцин (регулятор сигнальных путей G-белков) и глутатион S-трансфераз (белки химической защиты). Представленные в таблице множественные выравнивания последовательностей тиоредоксинов сами

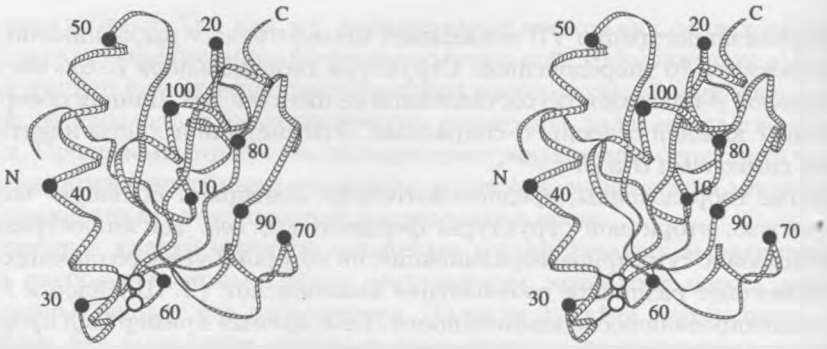


Рис. 4.5. Структура тиоредоксина *E. coli* [2TRX](см. также цветную иллюстрацию VII). Номера остатков соответствуют аналогам в таблице множественного выравнивания последовательности. Концевые N- и C-остатки также отмечены. Шары указывают позиции Cα-атомов каждого десятого остатка. Дисульфидный мостик между Cys32 и Cys35 показан между 30-м и 60-м остатками

по себе являются паттернами, которые могут быть использованы для поиска более удаленных родственников.

Программы для поиска множественного выравнивания последовательностей по базам данных

Поиск гомологов известных белков по базам данных — центральная тема биоинформатики. Действительно, это не терпит отлагательств; мы продемонстрировали это в гл. 1 на примере PSI-BLAST. Здесь мы опять прибегнем к поиску по базам данных с целью попытаться понять, каким образом мы можем лучше всего использовать доступную информацию для построения эффективных процедур. Требуется обеспечить высокую чувствительность — выделение даже самых удаленных родственников — и высокую селективность — минимизировать число последовательностей, о которых известно, что они ненастоящие гомологи. Здесь мы обсуждаем, как использовать множество последовательностей для решения задачи поиска гомологов. В гл. 5 мы вдобавок обсудим, как применить структурную информацию.

Мы узнаем знакомое лицо, реагируя больше на его общие черты, нежели на индивидуальные особенности. Подобно этому, множественные выравнивания последовательностей содержат скрытые паттерны, которые характеризуют семейства белков.

За прошедшие десять лет был достигнут большой прогресс в области разработок методов применения множественных выравниваний последовательностей известных белков для идентификации родственных последовательностей при поиске по базам данных. Результаты весьма важны для современных приложений биоинформатики, включая аннотацию геномов. Вот эти наиболее важные методы: профили, PSI-BLAST и скрытые марковские модели (Hidden Markov Model, HMM).

Профили

Профили выделяют наборы признаков, наблюдаемых в однородной выборке многих последовательностей (имеется в виду функционально или эволюционно однородной выборке. — *Прим. ред.*). У них имеется несколько способов применения:

- Они позволяют более аккуратно строить выравнивания дальнеродственных последовательностей.
- Наборы остатков высокой степени консервативности позволяют предположить их принадлежность к активному сайту и определять функцию.
- Консервативные паттерны облегчают идентификацию других похожих последовательностей.
- Консервативные паттерны могут использоваться для классификации подсемейств в множестве гомологов.
- Наборы остатков, в которых консервативность проявлена в низкой степени и в которых встречаются вставки и делеции, с высокой долей вероятности проявляются в петлях на поверхности. Эта информация применялась при разработке вакцин, поскольку указанные области с высокой долей вероятности провоцируют образование антител, которые будут хорошо взаимодействовать с нативными структурами.

  Большинство методов предсказания структур более эффективны, если они основаны на множественном выравнивании последовательностей. Например, моделирование по гомологии критически зависит от качества выравнивания.

Для использования профилей в процессе идентификации гомологов, необходимо сопоставлять интересующие нас последовательности с последовательностями из базы данных, приведенными в таблице выравниваний, придавая консервативным позициям больший вес по сравнению с изменяемыми позициями. Если регион абсолютно консервативен, как, например, WGCPC, процедура должна требовать обязательного присутствия этого мотива.. Но слишком высокая степень жесткости данной операции может повлечь за собой пропуск интересных дальних родственников, поэтому должны быть допущены некоторые отклонения.

Необходимо ввести количественную меру консервативности. Для этого в каждой позиции в таблице выровненных последовательностей дано распределение аминокислот. Например, для позиций 25–30 тиоредоксина приводятся следующие аминокислоты:

Число остатков	Количество каждой аминокислоты																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
25	1									2								13		
26			16																	
27					16															
28																7	1		5	3
29	16																			
30			1	4									2			1	7	1		

Пусть нас интересует какая-то последовательность (последовательность *запроса*), представляющая собой потенциальный гомолог тиоредоксина. Мы хотим оценить, насколько эта последовательность соответствует известным последовательностям в абсолютно консервированных позициях, например, 26, 27 и 29. Соответствие в этих позициях принесет очень высокое значение веса, а несоответствие в этих позициях — низкое значение. Для позиций со средней степенью консервации мы хотим получить незначительный положительный вклад в вес выравнивания, если в запрошенной последовательности в данной позиции есть S или W, и меньшее значение, если имеются T или Y. Общая идея — подсчитывать значение каждого остатка из последовательности запроса, основываясь на распределении аминокислот в этой позиции согласно таблице множественного выравнивания последовательностей.

Соблазнительно было бы попытаться применить эти значения сразу же, но это было бы слишком просто. Например, если регион в запрошенной последовательности, которая соответствует позициям 25–30 в тиоредоксине, содержит VDFSAE, то данный фрагмент получит значения веса $13 + 16 + 16 + 6 + 16 + 4 = 71$. Это практически наивысшее из возможных значений. Альтернативная запрошенная последовательность ACGVAP получит значения веса $1 + 0 + 0 + 5 + 16 + 2 = 24$, что намного меньше[1)]. Конечно, для каждой запрошенной последовательности нам необходимо протестировать все возможные выравнивания и брать наибольшее общее значение. Последовательности с наибольшим значением веса лучше всего подходят наборам, представленным в таблице.

Этот простой подход работал бы, если бы наша таблица содержала большую и репрезентативную выборку образцов последовательностей тиоредоксинов. Но только в этом случае простая выборка дала бы правильную картину *потенциального* распределения остатков в данной позиции. Если бы наша выборка образцов была маленькой, набор наблюдаемых частот не смог бы отразить весь набор характеристик данного семейства. Или, если бы в выборке содержалось большее количество похожих последовательностей, то они бы были слишком часто представлены в выборке. Мы можем это видеть на цветной иллюстрации VII, где тиоредоксины позвоночных формируют весьма близкородственный набор. Если бы мы включили еще 20 тиоредоксинов позвоночных в выравнивание, профиль эффективно распознавал бы только тиоредоксины позвоночных.

Матрицы замен аминокислотных остатков позволяют сделать выборку более расплывчатой и, следовательно, более общей.

Наблюдаемое распределение аминокислот в любой позиции остатка представляет собой множество из двадцати чисел $(a_1, a_2, a_3, \ldots, a_{20})$, где a_i — количество аминокислот типа i наблюдаемого в данной позиции (для позиции 25 тиоредоксина $a_1 = 1$, поскольку аланин наблюдается один раз, и $a_{18} = 13$,, что соответствует валину). Затем, следуя простой схеме, получаем, что значение

[1)] На самом деле при вычислении веса в сопоставлениях используются не количество встреченных остатков того или иного типа в колонке, а логарифмы частот встречаемости. – *Прим. ред.*

Ala в позиции 25 равно всего лишь единице, значение Val равно 11. В целом значение аминокислоты типа i равно a_i. В данной схеме подсчет встречаемости аминокислот в каждой колонке порождает массив a, который используется при построении выравнивания последовательности и профиля.

Более приемлемая схема расчета оценивала бы любую аминокислоту согласно ее шансам быть замененной одной из наблюдаемых аминокислот. Если $D(i,j)$ — это аминокислотная матрица замены — PAM250 или, например, BLOSUM62, тогда аминокислота i может получить значение $a_1D(i,1) + a_2D(i,2) \cdots a_{20}D(i,20)$. Эта схема распределяет значения среди наблюдаемых аминокислот, взвешенных согласно вероятностям замен. Аминокислота в последовательности запроса могла бы получить большее значение, либо если она часто появляется в запросе в данной позиции, либо если для нее существует высокая вероятность появления с помощью мутации из остатков тех типов, которые обычны для данной позиции[1)].

Данный подход более эффективен для распознавания отдаленных родственников при использовании ограниченного набора известных последовательностей. В этом случае вектор значений для аминокислот является произведением матрицы замен и вектора частот встречаемости остатков. Еще более правильным в данном случае было бы использование комбинации наблюдаемой выборки и частот встречаемости аминокислот в качестве распределения аминокислот.

Результатом является набор вероятных значений для каждой аминокислоты (или ее пропуска) в каждой позиции выравнивания, который называется позиционно-специфической матрицей замен (PSSM — Position Specific Substitution Matrix, или PWM — Position Weight Matrix — эти два понятия идентичны. — *Прим. ред.*). Альтернативный метод получения позиционно-специфической матрицы замен, основанных на трехмерных структурах, описан в гл. 5.

Вычисления, которые необходимы для поиска оптимального выравнивания запрошенной последовательности и профиля среди всех возможных выравниваний являются обобщением метода динамического программирования выравнивания двух последовательностей.

Недостаток простого профиля заключается в том, что множественное выравнивание последовательностей должно быть произведено заранее и учитывается в фиксированном виде. Программы PSI-BLAST и скрытые марковские модели (HMM) набирают силу с ростом набора статистических данных.

PSI-BLAST

PSI-BLAST — это программа, которая подбирает данные для последовательностей, аналогичных запрошенной. Она является обобщенной версией программы BLAST. Эта программа (и ее варианты) независимо сравнивает каждую запись в базе данных с запрошенной последовательностью. Программа PSI-BLAST начинает работу с такого же поэлементного сравнения. Затем она

[1)] На самом деле следует использовать не матрицу PAM или BLOSUM, а матрицу *вероятностей замен*. — *Прим. ред.*

BLAST programs come in several flavours

Program	Type of query sequence	Search in database of
BLASTP	Amino acid sequence	Protein sequences
BLASTX	Translated nucleotide sequence	Protein sequences
TBLASTN	Amino acid sequence	Translated nucleotide sequences
TBLASTX	Translated nucleotide sequence	Translated nucleotide sequences
PSI-BLAST	Amino acid sequence	Protein sequence database

Эти программы сравнивают аминокислотные последовательности друг с другом, используя по умолчанию матрицу BLOSUM62. Поиск нуклеотидных последовательностей как в качестве запроса, так и в качестве базы данных для поиска производится после их трансляции в аминокислотную последовательность в шести возможных рамках считывания. Другое семейство программ, BLASTN, сравнивает нуклеотидные последовательности запроса и базы данных непосредственно.

строит локальное множественное выравнивание последовательностей, полученных при первичном запросе, и затем обращается к базе данных, используя уже это множественное выравнивание. Затем процесс повторяется (по полученному набору кандидатов снова строится множественное выравнивание), и результат уточняется в ходе нескольких итераций, пока не исчерпается заданное количество циклов или пока процедура не сойдется, т. е. пока результаты двух последовательных запросов не совпадут.

Причиной, по которой потребовалось создание программы BLAST, являлось то, что полномасштабные методы динамического программирования недостаточно быстры для задачи полного поиска в большой базе данных. Часто база данных содержит последовательности, очень похожие на запрошенную последовательность. Менее точные, но более быстрые программы вполне способны идентифицировать близкие совпадения, что в большинстве случаев и требуется. Например, если существует необходимость найти гомологи мышиного белка в геноме человека, то степень сходства скорее всего будет высокой, и более быстрые методы подойдут для решения этой задачи. Но в случае необходимости нахождения гомологов человеческого белка в *C. elegans* или дрожжах, различия будут более тонкими, и, следовательно, требуется программа с более высокой степенью точности. (Возможно, это удивит вас, но количество времени, потраченного программой на получение результата, все еще имеет значение, поскольку, хоть компьютерное время сейчас и дешевле, чем раньше, но размер баз данных по всему миру быстро увеличивается, что приводит к возрастанию затрат на компьютерное обеспечение исследований).

Метод, на котором основана программа BLAST, восходит к точечным матрицам сходства, где выявлялись хорошо совпадающие локальные участки. Для каждой записи в базе данных проверяются короткие сопредельные участки,

Поток данных в PSI-BLAST

1. Сравнить независимо каждую последовательность с последовательностью запроса.
2. Отобрать значимые хиты; построить множественное выравнивание для значимых блоков совпадения запрошенной последовательности и последовательностей из базы данных.
3. Построить профиль по множественному выравниванию.
4. Сравнить каждую последовательности из базы данных с профилем.
5. Отобрать статистически значимые хиты.
6. Вернуться к шагу 2. Процесс завершается, если при очередной итерации результат не изменился. Итеративный характер процедуры отражен в названии программы.

совпадающие с короткими сопредельными участками запрошенной последовательности, для чего используется матрица аминокислотных замен, но без пропусков. Участки фиксированной длины быстро определяются при использовании хеш-таблиц.

Как только программа BLAST определяет подходящий участок, он пытается его расширить. В некоторых версиях программы допускается наличие пропусков. На выходе программа предлагает набор локальных сегментных совпадений. В примере, приведенном в гл. 1: даже при применении очень простого алгоритма находятся все совпадающие участки пяти смежных остатков, которые затем комбинируются и расширяются.

```
  My.care.is.loss.of.care,.by.old.care.done,
    |||||||||     |||||||||||||    ||||||  ||
Your.care.is.gain.of.care,.by.new.care.won
```

Программа PSI-BLAST, которая использует итеративный поиск последовательности, гораздо эффективнее BLAST при изучении менее близкородственных связей. PSI-BLAST точно определяет в три раза больше гомологов, чем BLAST, в регионах, в которых совпадения составляют меньше 30%. Следовательно, этот метод весьма хорошо применим в случае анализа целых геномов. PSI-BLAST способна идентифицировать белковые домены известной структуры для 39% генов *M.genitalium*, 24% генов дрожжей и 21% генов *C.elegans*.

Единственным более эффективным методом, основанным на анализе последовательностей, является метод скрытых марковских моделей, который описан в следующем разделе. Для получения результатов существенно более высокого качества необходимо явно использовать структурную информацию. Этот процесс описан в следующей главе.

Скрытые марковские модели (HMM)

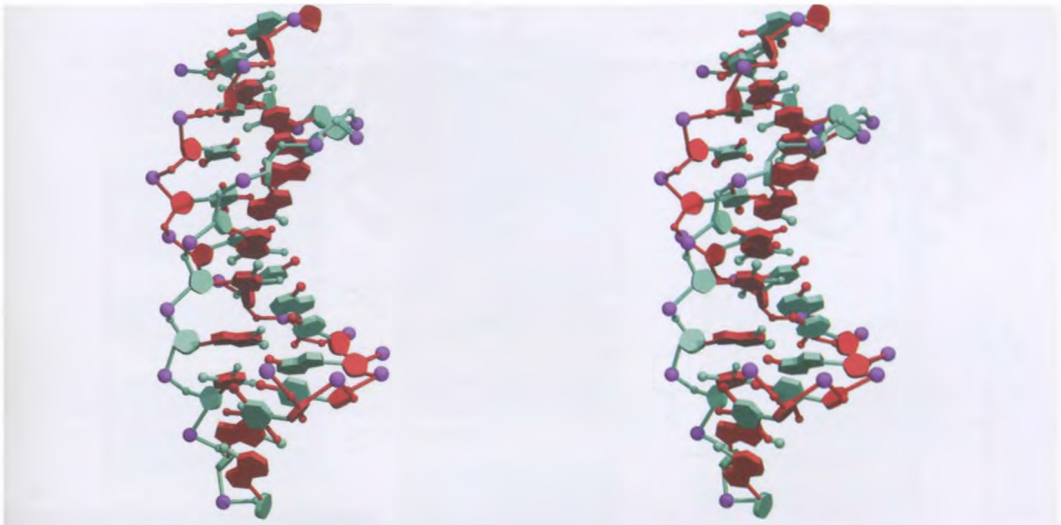
Скрытые марковские модели (Hidden Markov Models, HMM) служат для описания тонких различий, существующих между семействами гомологичных последовательностей. Метод эффективен при сравнении дальних родственников и при предсказании путей сворачивания белков. Только этот метод, полностью базирующийся на анализе последовательностей (т. е. не использующий структурную информацию), может соперничать с программой PSI-BLAST при идентификации дальних гомологов. Эти программы эффективны также при распознавании сворачивания, что оценивается в соревнованиях CASP.

HMM работает с множественными выравниваниями последовательностей. Тем не менее HMM часто используется для генерации последовательностей. Обычная таблица множественного выравнивания последовательностей также может использоваться для генерации этих последовательностей путем выборки аминокислот в последовательных позициях; каждая аминокислота выбирается из вероятностного позиционно-специфического распределения, полученного из профиля. Но модели HMM являются более общими методами, чем профили.

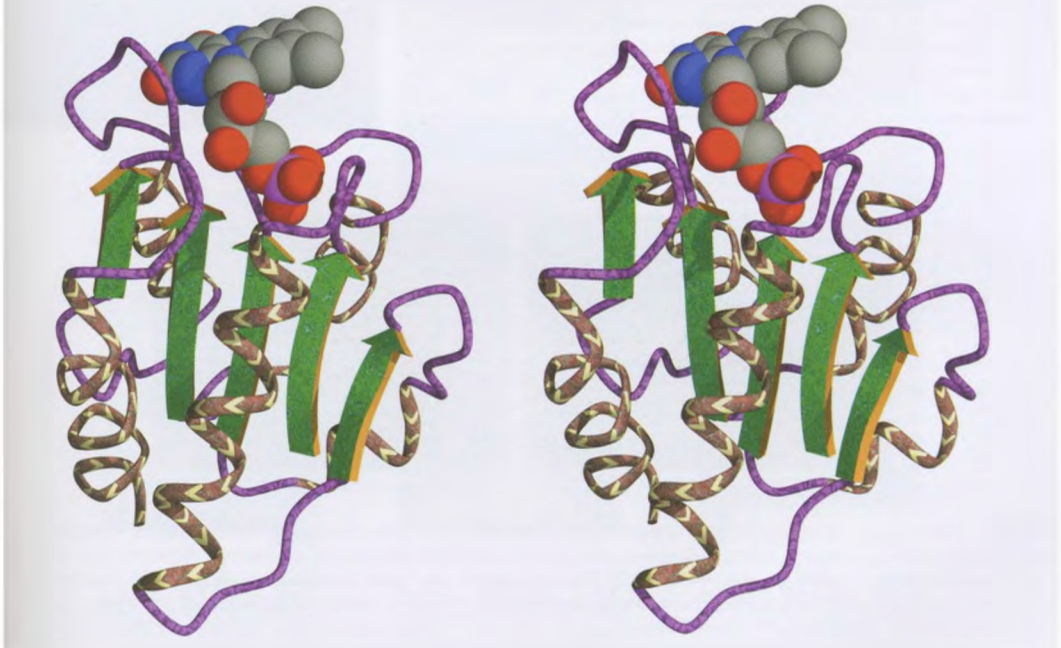
1. Они включают вероятность возникновения делеций или вставок в генерируемой последовательности с позиционно-зависимыми штрафами для них.
2. Применение профилей требует предварительного построения множественного выравнивания последовательностей. При таком подходе статистика запросов определяется после построения выравнивания. HMM же выполняет выравнивание и оценку вероятностей одновременно.

Внутренняя структура HMM показывает механизм генерации последовательностей (см. рис. 4.6). Начало — там, где Start, а затем по стрелкам доходим до End. Каждая стрелка приводит к определенному состоянию системы. На каждой из этих стадий вы предпринимаете определенное действие (например, выделяете остаток) и затем выбираете стрелку, которая приводит вас к следующему состоянию системы. Действие и выбор следующей стадии управляются набором вероятностей. С каждой стадией, на которой выделяется остаток, связаны: одно распределение вероятностей — на каждые 20 аминокислот, второе распределение вероятностей — на выбор следующей стадии процесса. Оба этих распределения вероятностей подбираются так, чтобы дать адекватное описание определенного семейства последовательностей. Таким образом, математически один и тот же процесс может быть использован для различных семейств последовательностей.

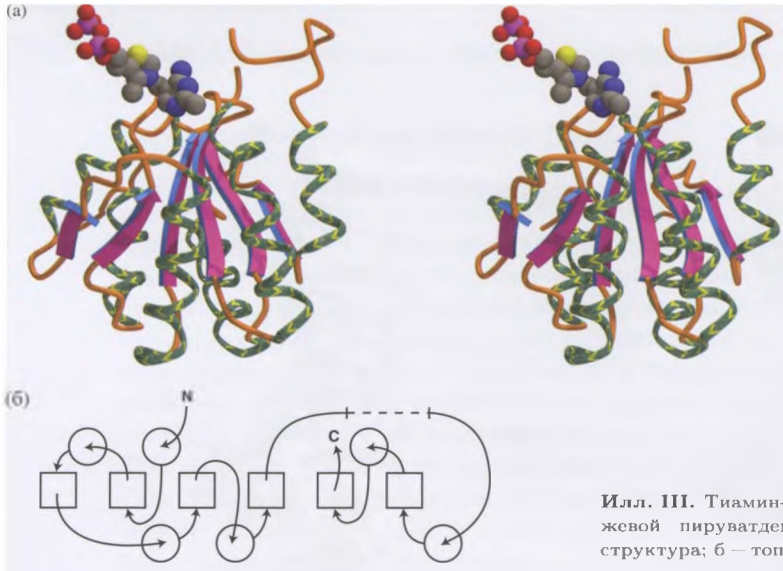
Динамика данной системы такова, что только настоящее положение влияет на выбор своего преемника — система не «помнит» своей истории. Это характеристика процессов изученных в XIX в. русским математиком А. А. Марковым. Рассмотрим череду положений от последовательности исходных аминокислот до конечной последовательности. Некоторые пути через систему могут порождать одинаковые последовательности. Видима только конечная последовательность, переходные же состояния находятся внутри системы, т. е. спрятаны. Паттерн наследует семейство последовательностей с помощью распределений



Илл. I. Двойная спираль РНК. (См. с. 7.)

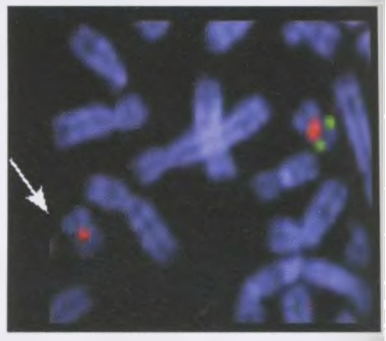


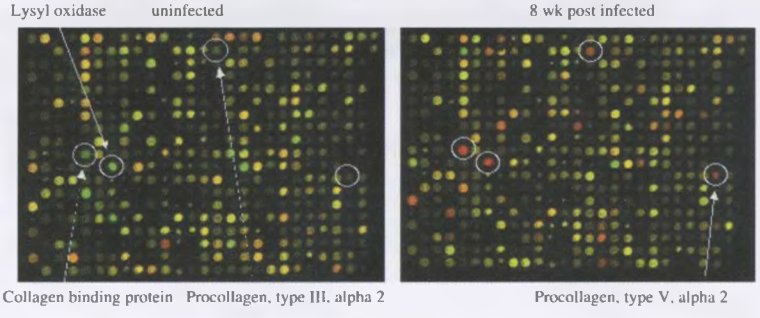
Илл. II. Флаводоксин из *Clostridium beijerinckii*, связанный с кофактором FMN [5NLL]. Большие стрелки обозначают тяжи β-листов. Положение этой структуры в иерархической классификации базы данных SCOP описано на с. 269.



Илл. III. Тиамин-связывающий домен из дрожжевой пируватдекарбоксилазы а — трехмерная структура; б — топологическая схема (см. с. 269)

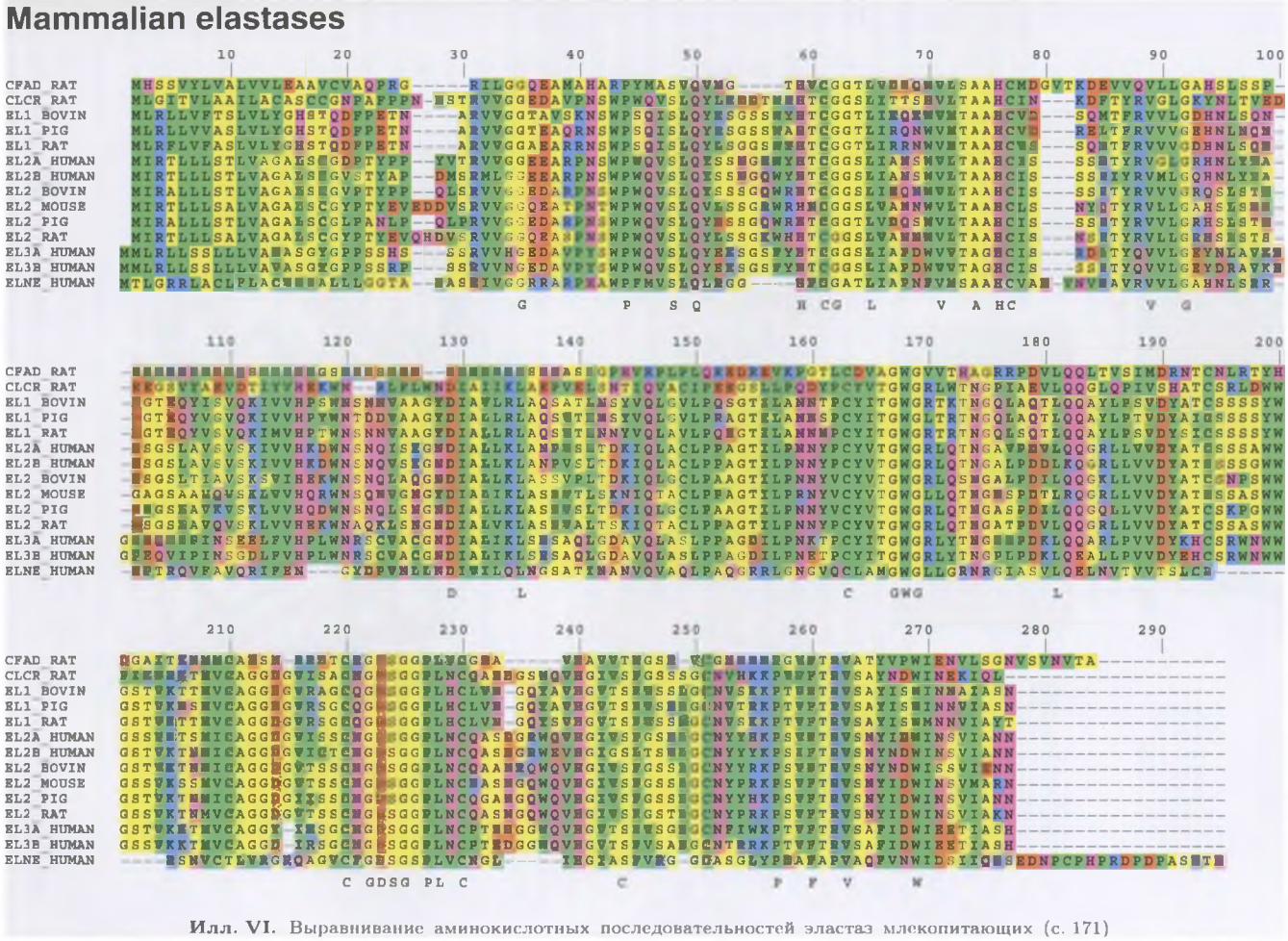
Илл. IV. FISH (Флуоресцентная *in situ* гибридизация) может детектировать наличие локус-специфических зондов и визуализировать их позицию на хромосомах Красным показаны зонды к центромерным районам хромосомы 20, что позволяет идентифицировать две гомологичные копии, которые появляются в метафазе. Зеленые зонды к маркеру B20S108 в районе 20q11.2–13.1, который присутствует в одной копии хромосомы 20 и отсутствует в другой копии (отмечено стрелкой). Этот образец получен от пациента, страдающего полицитемией (анормальным увеличением клеток крови, в основном эритроцитов, появляющихся в костном мозге). Предполагается, что делетированный район в длинном плече хромосомы 20 содержит ген супрессора опухолей, потеря которого дает вклад в развитие лейкемии (см. с. 89). (Courtesy of Dr E Nacheva, Department of Haematology, University of Cambridge.)



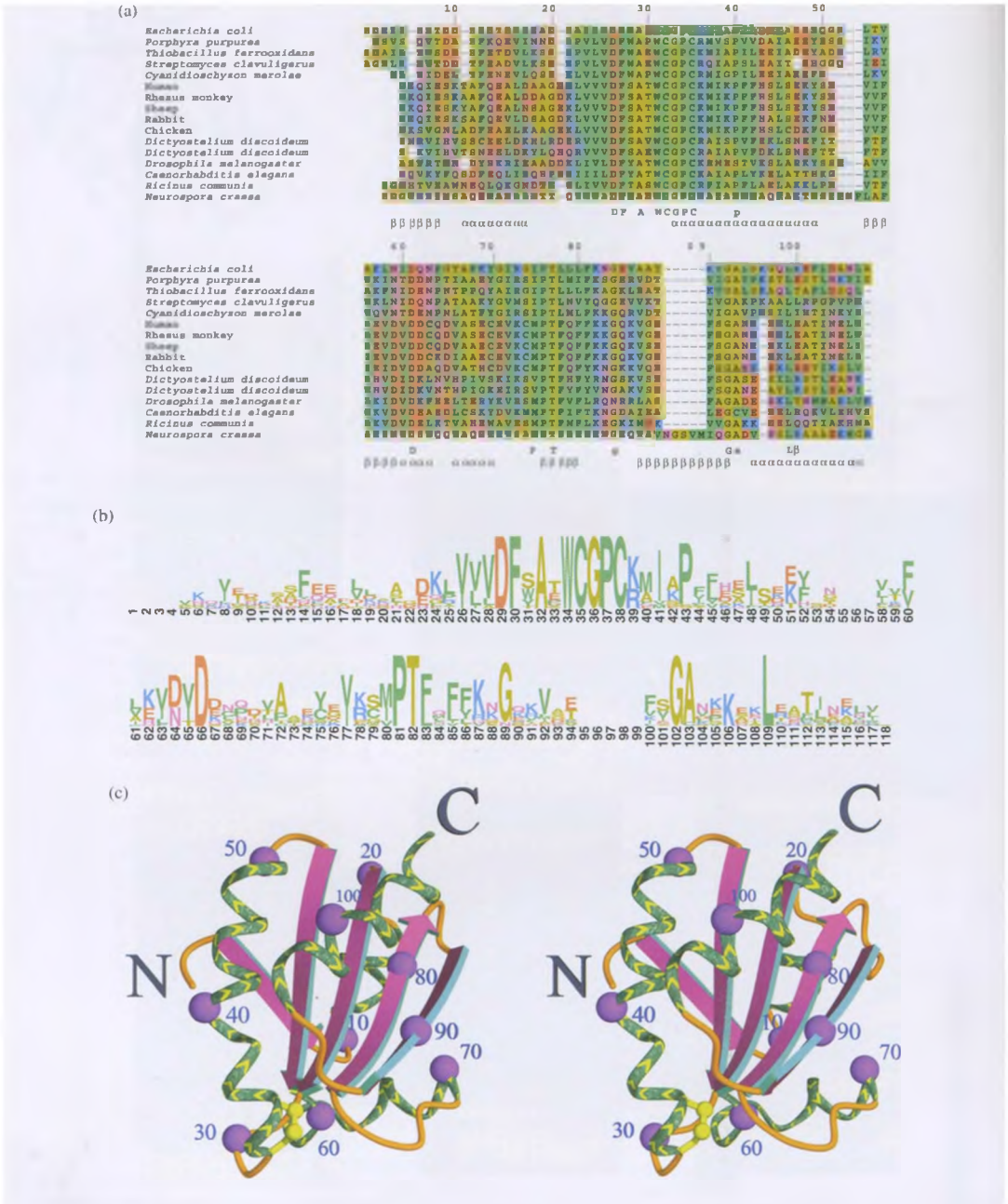


Илл. V. Цитосомы, или кровяные сосальщики — широко распространенные паразиты в субтропиках. Представлено использование кДНК микрочипов для измерения эффекта от инфекции *Schistosoma mansoni* в транскрипционной пробе из печени мыши. Каждое пятно на чипе показывает активность одного гена. Пятна на двух чипах позволяют сравнивать активность генов в неинфицированном контроле (*слева*), животном (*справа*), инфицированном 8 недель назад. Зеленые пятна показывают уровень не индуцированной экспрессии, красные — индуцированный уровень.

Целью такого эксперимента является идентификация дифференциального паттерна экспрессии генов. В представленном случае некоторые гены, которые повысили свою экспрессию в ответ на заражение (отмечены на рисунке) участвуют в синтезе и включении коллагена. Это ассоциировано с баллансированным механизмом защиты хозяина, когда яйца паразита заключаются в фиброзные гранулы. Изучение паттерна экспрессии генов, связанных с развитием этого заболевания может пролить свет на механизмы патогенеза (с. 84). (Courtesy of Dr K Hoffmann, Department of Pathology, University of Cambridge.)



Илл. VI. Выравнивание аминокислотных последовательностей эластаз млекопитающих (с. 171)



Илл. VII. *a* — выравнивание аминокислотных последовательностей тиоредоксина из *E. coli* и его гомологов. Некоторые последовательности оборваны на концах. Нумерация остатков соответствует последовательности из *E. coli* (*верхняя строка*). Спирали (α) и тяжи (β) размечены в соответствии со структурой тиоредоксина из *E. coli* по банку PDB (запись 2TRX) *b* — последовательность лого, полученная из выравнивания. *c* — структура тиоредоксина из *E. coli* [2TRX] содержит β-лист из пяти тяжей, на концах которых расположены α-спирали. Нумерация остатков соответствует выравниванию. Отмечены также N и C концы. Шарики отмечают положение каждого десятого остатка. Активный бисульфидный мостик между Cys32 и Cys35 показан желтым. (см. с. 215–216)

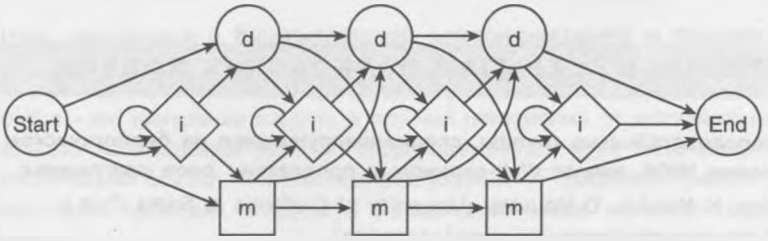


Рис. 4.6. Структура скрытой марковской модели (HMM). Каждой позиции во множественном выравнивании HMM соответствуют состоянию сопоставления (m) и делеции (d). Состояния вставки (i) появляются между позициями остатков, в начале и в конце

- Состояния сопоставления положения «испускают» остаток. В данном случае слово «сопоставление» означает только то, что *некоторые* аминокислоты есть и в последовательности, лежащей в основе HMM, и в последовательности, полученной на выходе, а не то, что это должны быть обязательно *одинаковые* аминокислоты. Вероятность «испускания» каждой из 20 аминокислот в каждой совпадающей позиции — это свойство модели. Как с профилями, вероятности зависят от позиции.
- Состояния делеций пропускают колонки во множественном выравнивании. Достижение удаленного положения после совпадающего или вставленного соответствует открытию делеции, и вероятность такого перехода отражает позиционно-специфический штраф за открытие делеции. Достижение состояния делеции после предыдущего удаленного положения соответствует увеличению делеции.
- Состояния вставки появляются между двумя последовательными позициями в выравнивании. Если система входит в состояние вставки, то новый остаток не соответствует позиции в таблице выравнивания, появившейся в «испускаемой» последовательности. Состояние вставки может следовать само за собой, вставляя больше, чем один остаток. Последовательность остатков, полученных из состояний сопоставления и вставки, порождает конечную последовательность.

После выполнения действия, свойственного какому-либо состоянию (m, d или i), выбор следующего состояния определяется другим распределением вероятностей. В каждой возможной последовательности положений каждая колонка построенного выравнивания должна быть либо сопоставлением, либо делецией — не существует способа пройти сеть без посещения или m-положения, или d-положения в каждой позиции

вероятностей, ассоциированных с состояниями системы (модели). Отсюда и название — скрытая марковская модель (Hidden Markov Model).

Программное обеспечение применения HMM для анализа биологических последовательностей может делать такие вещи, как:

1. *Обучение.* Имея ряд невыровненных гомологичных последовательностей, можно выровнять их и подогнать вероятности переходов и порождения остатков, чтобы определить HMM, описывающую заданный набор последовательностей.
2. *Поиск дальних гомологов.* Имея HMM и исследуемую последовательность, можно посчитать вероятность того, что HMM могла бы сгенерировать эту последовательность. Если HMM, рассчитанная для известного семейства

WEB-РЕСУРСЫ: СКРЫТАЯ МАРКОВСКАЯ МОДЕЛЬ

Две исследовательские группы, специализирующиеся на биологическом применении HMM, имеют Web-серверы и предлагают свои программы:

R. Hughey, K. Karplus, D. Haussler (University of California at Santa Cruz.): `http://cse.ucsc.edu/research/compbio/sam.html`

`http://cse.ucsc.edu/research/compbio/HMM-apps/HMM-applications.html`

S. R. Eddy (Washington University, St. Louis, MO, USA) `http://hmmer.wustl.edu/`

Результаты анализа известных последовательностей и структур также доступны по следующим адресам:

Pfam — это база данных множественных выравниваний и HMM для многих белковых доменов, разрабатывается A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, E. L. Sonnhammer:
`http://www.sanger.ac.uk/Software/Pfam`.

J. Gough, K. Karplus, R. Hughey, C. Chothia создали HMM для всех PDB суперсемейств:
`http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY/`

последовательностей, могла бы это сделать с достаточно большой вероятностью, то видимо заданная последовательность принадлежит к этому семейству.

3. *Выравнивание дополнительных последовательностей.* Вероятность какой-либо последовательности из возможных (для данной HMM) может быть посчитана из индивидуальной вероятности перехода «положение-за-положением». Нахождение наиболее вероятной последовательности положений, которые использовала бы HMM для создания одной или нескольких тестируемых последовательностей, покажет их оптимальное выравнивание в семействе.

Филогения

Мы видели разные примеры эволюции в белках и геноме. Это означает распространение на молекулярный уровень идей, которые занимали умы биологов, начиная с Дарвина или даже еще раньше. Основной принцип: *источник подобия — общее происхождение.* Несмотря на то что существует множество исключений, возникших в результате конвергентной эволюции, важность этого принципа не может быть переоценена как для современных рациональных наблюдений, так и для прорубания окна в историю жизни.

Понятия, связанные с биологической классификацией и филогенезом

Гомология означает, в частности, происхождение от общего предка.

Подобие — это измерение сходств и отличий независимо от источника сходства. Подобие — это наблюдаемые данные, собранные *сейчас* и не включающие каких-либо исторических гипотез. И напротив, рассуждения о гомологии требуют гипотез об исторических событиях, которые в большинстве своем не могут поддаваться наблюдению.

Кластеризация — это сведение вместе сходных предметов, различая классы объектов, более сходных с любыми другими, чем те объекты, которые не входят в эти классы. Большинство людей согласятся с уровнями подобия, но кластеризация более субъективна. При классификации объектов некоторые люди предпочитают более обширные классы, допускающие больше вариаций, другие предпочитают меньшие, более строгие классы. Они называются *разветвителями* или *группировщиками*.

Иерархическая кластеризация — это многоступенчатое группирование кластеров из кластеров.

Филогения — описание биологических взаимосвязей, обычно отображается в виде филогенического дерева. Филогенез *предполагает* гомологию между объектами и *зависит* от классификации. Филогенез устанавливает топологию взаимосвязей, базируясь на классификации, соответствующей сходству одного или нескольких свойств, или на модели эволюционных процессов. Во многих случаях филогенетические взаимосвязи основываются на том, что различные свойства согласуются и влекут за собой другие. Если какие-нибудь свойства вызывают появление несообразных филогенетических взаимосвязей, то они все сомнительны. И наоборот, отмечено, что некоторые похожие данные могут быть соотнесены с различными возможными топологиями или «деревьями».

Целями филогенетических исследований является выявление взаимосвязей между видами, популяциями, индивидами или генами[1)]. Здесь под «взаимосвязями» подразумевается родство или генеалогия, т. е. схема (модель) распределения потомков от общего предка. Результаты обычно представлены в виде генеалогического древа. Таксономия страусообразных — больших нелетающих птиц — типичный тому пример (рис. 4.7, *а*). Полагают, что предок всех страусообразных был летающей птицей, возможно родственной современным тинаму.

Дерево, показывающее всех потомков от одного предка, называется *укорененным*. (Корень дерева обычно находится сверху или сбоку, ботаники должны свыкнуться с этим.) Кроме того, мы можем установить взаимосвязи, но не расположить их в соответствии с историческим ходом событий. Родственные взаимосвязи между вьюрками Галапагосских островов, изученные Дарвином,

[1)] Основной термин — это таксон — систематическая группа. *Наблюдаемая* систематическая группа, например, та, для которой мы хотим получить модель предка, называется «рабочей таксономической единицей» («operational taxonomic unit», OTU).

и плюс родственные виды с соседних Кокосовых островов, показаны на *неукорененном* дереве (рис. 4.7, *б*). Добавление данных о видах Южно-Американского материка к островным вьюркам могло бы привести нас к *укорененному дереву*.

Само дерево может показывать только взаимосвязи или топологию, в этом случае длина ветвей не несет никакой информации. Более честолюбивая цель состоит в том, чтобы показать количественное соотношение для расстояний между систематическими группами, например соотнести длину со временем после отделения от предка. Как мы можем получить информацию о взаимосвязях между различными организмами, имея для различных видов животных набор данных, характеризующих эти группы организмов, например последовательность ДНК или белка, или структуру белка, или форму зубов? Редко случается так, что можно непосредственно установить предка и видовые взаимоотношения. Эволюционные деревья, построенные с помощью генетических данных, часто основываются на подразумеваемых моделях подобия, которые наблюдаются между современными видами. В основном мы принимаем версию о том, что чем больше похожи свойства, тем ближе в родственном отношении виды, хотя это просто чудовищное приближение. Тем не менее, из взаимоотношений между свойствами мы хотим получить модель предка: *топологию* филогенетических взаимоотношений (говоря менее формально, родословную).

До какой степени топология взаимосвязей зависит от выбора сравниваемых свойств? В частности, существует ли систематическое противоречие между результатами молекулярного и палеонтологического анализа?

Молекулярный подход к филогенезу развивался, опираясь на традиционную систематику, основанную на морфологических свойствах, эмбриологии и стратиграфии (информации о геологической ситуации для ископаемых). У классических методов есть ряд достоинств. Классическая систематика основана на изучении вымерших организмов по окаменелостям. Таким образом удается регистрировать появление и исчезновение видов геологическими методами. Молекулярные биологи же очень редко обращаются к вымершим видам. Четкую ДНК имеют лишь некоторые субфоссильные остатки видов, вымерших недавно (век или два назад), включая образцы из квагги (родственник зебре), тасманского тигра (сумчатого тасманского волка) и некоторых новозеландских птиц (в том числе моа). Мы уже видели примеры последовательности из мамонта. Некоторые последовательности ДНК неандертальского человека были получены из особи, умершей ок. 30 000 лет назад. Но *Парк Юрского периода* это все-таки фантастика!

Решающий момент в признании молекулярных методов произошел в 1976 г., когда В. М. Сарих и А. К. Уилсон, основываясь на иммунологических данных, установили, что люди отделились от шимпанзе 5 млн лет назад. До этого времени палеонтологи датировали это разветвление 15 млн лет и очень неохотно соглашались с молекулярным подходом. Новое толкование данных по ископаемым останкам привело к установлению более позднего расхождения этих двух видов и уничтожило барьер между классическим подходом и молекулярными методами.

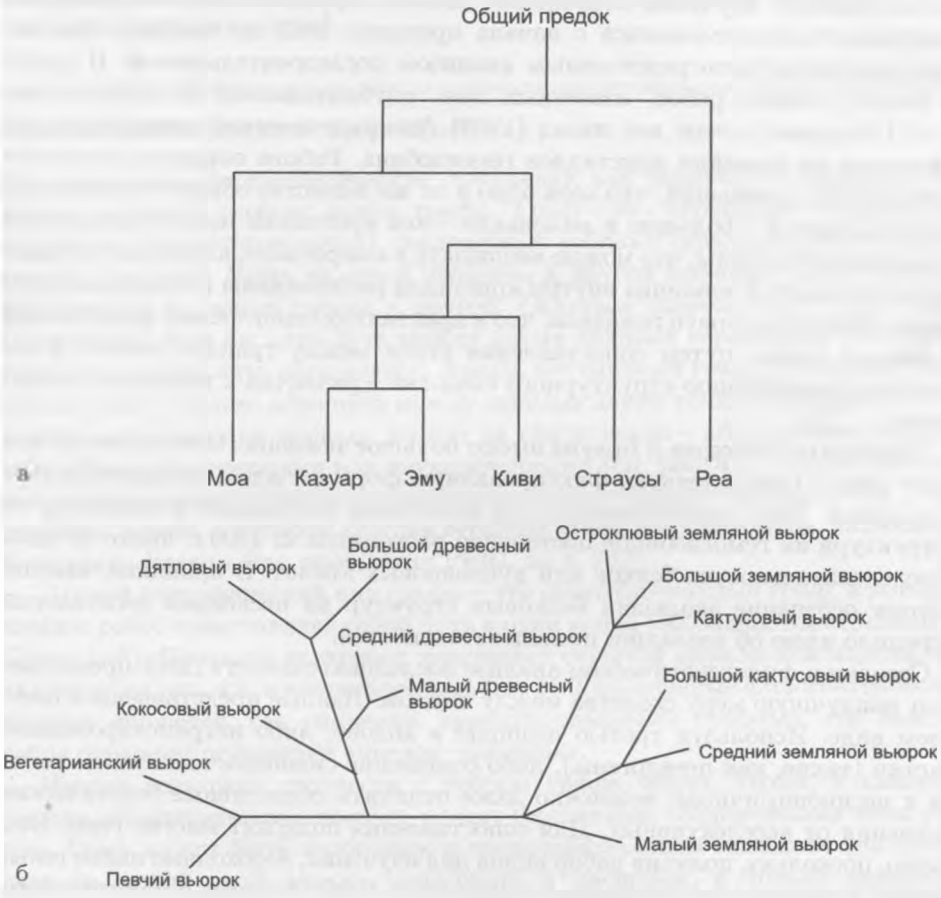


Рис. 4.7. *а* — Филогенетическое дерево страусообразных (больших нелетающих птиц) основано на анализе последовательностей митохондриальных ДНК. Общий предок — это *корень* дерева. Возникает неожиданный вывод из этого дерева: моа и киви не близкие родственники и, следовательно, Новая Зеландия была колонизована двумя страусообразными или их предком. *б* — *Неукорененное* дерево взаимоотношений между вьюрками Галапагосских и Кокосовых островов. Дарвин изучал галапагосских вьюрков в 1835 г., отмечая различия в форме их клювов и соответствие между формой клюва и рационом. Вьюрки, которые едят фрукты, имеют клюв, как у попугаев, а вьюрки, которые едят насекомых, имеют длинный узкий клюв. Эти наблюдения послужили отправной точкой для развития идей Дарвина. Уже в 1839 г. он написал в *«Путешествии на "Бигле"»*: «Посмотрев на эту градацию и разделение структуры в одной маленькой, близкородственной группе птиц, действительно можно было представить себе, что из исходного, небольшого числа птиц этого архипелага был взят один вид и изменен для разных целей».

Несомненно, что многие молекулярные свойства использовались для филогенетического изучения неожиданно давно. Серологическая перекрестная реактивность использовалась с начала прошлого века до тех пор, пока не была вытеснена непосредственным анализом последовательностей. В одной из самых ранних работ, известных мне, опубликованной Е. Т. Рейхертом и А. П. Брауном почти век назад (1909) филогенетический анализ рыб основывался на строении кристаллов гемоглобина. Работа опиралась на закон Стено (1669), гласивший, что хотя одно и то же вещество образует кристаллы разного размера — большие и маленькие — эти кристаллы имеют одинаковые межплоскостные углы, что можно наблюдать в микроскопе, а поэтому атомные или молекулярные единицы внутри кристалла расположены (упакованы) одинаково. Рейхерт и Браун показали, что в кристаллах гемоглобина, выделенных из разных видов, путем сопоставления углов между гранями можно устанавливать корреляцию структурного сходства — различия с видовыми различиями.

Результаты Рейхерта и Брауна имеют большое значение. Они показали, что белки имеют определенную, фиксированную форму, — идея, не признаваемая в то время. Они предположили, что если виды постепенно расходятся, то и структура их гемоглобинов постепенно расходится. В 1909 г. никто не знал о последовательностях белков или нуклеиновых кислот. В принципе, именно поэтому осознание эволюции белковых структур на несколько десятилетий опередило идею об эволюции последовательностей.

Сегодня в филогенетическом анализе последовательности ДНК предоставляют наилучшую меру сходства между видами. Данные представлены в цифровом виде. Используя третью позицию в кодоне, либо нетранслированные участки (такие, как псевдогены), либо отношение синонимичных замен кодонов к несинонимичным, возможно даже отличить селективные генетические изменения от неселективных. Для сопоставления подходят многие гены. Это удачно, поскольку, получив набор видов для изучения, необходимо найти гены, которые разошлись на подходящее расстояние. Гены, которые остаются почти неизменными среди интересующих нас видов, не дают никакого различия в степени сходства. Гены, которые разошлись слишком сильно, не могут быть выровнены. Аналогичная ситуация возникает при радиоактивной датировке, требующей выбор изотопа с периодом полураспада, сравнимым с исследуемым промежутком времени.

К счастью, гены сильно различаются по степени изменчивости. Митохондриальный геном млекопитающих (циклическая двуцепочная молекула ДНК длиной примерно 16 000 пн) предоставляет набор быстро изменяющихся последовательностей, полезный для изучения эволюции близкородственных видов. Напротив, последовательности рибосомальных РНК были использованы К. Возом (C. Woese), чтобы идентифицировать три больших таксономических империи: Архебактерий, Бактерий и Эукариот.

Напротив, разные степени изменений у последовательностей разных генов могут привести к различным и даже противоречивым результатам в филогенетических исследованиях. Это особенно верно, если то, что нам нужно, — не просто топологическая схема родства, а длина ветвей. В дополнение, го-

ризонтальный перенос генов и конвергентная эволюция представляют собой конкурирующие явления, которые смешиваются с выводом относительно филогенетических отношений.

Филогенетические деревья

Мы описываем филогенетические отношения как деревья. В информатике дерево является особым видом графа. Граф — это структура, содержащая вершины (абстрактные точки), соединенные ребрами (представлены линиями между точками). *Путь* из одной вершины к другой проходит вдоль группы, как поездка из одного города в другой. *Связным графом* называется граф, содержащий хотя бы один путь между двумя любыми вершинами. Исходя из этого, мы можем дать определение *дереву*: это связный граф, в котором существует *один и только один* путь между любыми двумя точками. Одна вершина может быть выбрана *корнем*; но это не обязательно — абстрактные деревья могут быть укорененными или неукорененными (см. рис. 4.7). Неукорененные деревья показывают топологию отношений, но не модель наследования. Укорененное дерево, в котором каждая вершина имеет двух потомков, называется *бинарным деревом* (см. программу PERL на с. 233).

Другой специфический вид графа — это ориентированный граф, в котором каждое ребро представляет собой путь в один конец (обозначается стрелкой. — *Прим. ред.*). Примеры включают диаграмму скрытой марковской модели, показанную на рис. 4.6, и нейронную сеть (гл. 5). Укорененные филогенетические деревья являются, без сомнений, ориентированными графами, где каждое ребро отражает отношение «предок—потомок».

Иногда возможно соотнести с ребром графа число, чтобы, в каком-то смысле, обозначить «расстояние» между вершинами, соединенными этим ребром. Граф может быть изображен в пропорциональном масштабе (на самом деле не любой граф можно нарисовать в масштабе, а только дерево!. — *Прим. ред.*). Длина пути через граф — это сумма длин ребер.

Словарь терминов, связанных с графами

Граф — абстрактная структура, содержащая вершины (точки) и ребра (соединяющие точки линиями).

Путь — группа последовательно соединенных ребер.

Связный граф — граф, содержащий хотя бы один путь между двумя вершинами.

Дерево — связный граф, в котором существует *один и только один* путь между любыми двумя точками.

Длина ребра — число, соотнесенное с каждым ребром и обозначающее, в каком-то смысле, расстояние между вершинами, соединенными этим ребром.

Длина пути — сумма длин всех ребер, которые составляют путь.

В филогенетических деревьях длины ребер обозначают либо какую-то меру различия между двумя видами, либо длину времени, прошедшего с их разделения. Предположение о том, что различия между живущими видами отражает время их дивергенции (расхождения), верно только в том случае, если степени дивергенции одинаковы для всех ветвей дерева. Известно много исключений; например, среди млекопитающих грызуны демонстрируют сравнительно высокие скорости эволюции для многих белков (см. Интернет-задание 4.8).

Вообще, существуют два подхода к построению филогенетического дерева. Один подход не имеет никакого отношения к исторической модели родства между видами. Начинают с измерения расстояний между видами и строят дерево с помощью процедуры иерархический кластеризации. Такой подход называется *фенетическим*. Альтернативный подход, *эволюционный*, состоит в рассмотрении возможных путей эволюции, в предположении о возможном предке каждой вершины и в выборе оптимального дерева в соответствии с какой-либо моделью эволюционных изменений: фенетика основана на сходстве; кладистика основана на генеалогии.

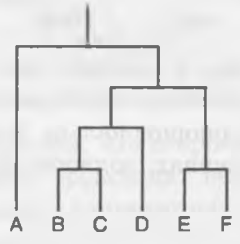
Методы кластеризации

Фенетический и кластризационный подходы к заданию филогенетических отношений явно не исторические. Действительно, иерархическая кластеризация прекрасно справляется с построением дерева даже при отсутствии эволюционных связей. В универмагах товары собраны в секциях в соответствии с типом товара, например, одежда или мебель, и распределены по подсекциям, еще сильнее связанным между собой, такие как мужская и женская обувь. Мужская и женская обувь имеет общего предка, но из этого не следует, что общий предок есть у обуви и мебели.

Простая процедура кластеризации осуществляется следующим образом: дана выборка видов, где для каждой пары установлена мера сходства или различия. Она может зависеть от физических черт тела, таких как разница в среднем росте взрослого организма у представителей двух видов. Либо можно использовать число несходных оснований в выравниваниях митохондриальных ДНК. Для построения дерева из выборки различий сначала выбирают два наиболее близкородственных вида и добавляют вершину, изображающую их общего предка. Затем замещают два выбранных вида группой, содержащей обоих, и заменяют расстояние от этой пары до остальных на среднее от расстояния от двух выбранных видов до остальных. Теперь мы имеем набор парных различий не между самостоятельными видами, а между группами видов. (Каждый оставшийся самостоятельный вид воспринимается как набор, содержащий только один элемент.) Потом процесс повторяют, как показано в следующем примере.

```
#!/usr/bin/perl
#drawtree.prl - draws binary trees (root at top)
#usage:  echo '(A((BC)D)(EF))' | drawtree.prl > output.ps

print <<EOF;
%!PS-Adobe-\n%%BoundingBox: atend
/n /newpath load def /m /moveto load def /l /lineto load def
/rm /rmoveto load def /rl /rlineto load def /s /stroke load def
1.0 setlinewidth 50 100 translate 2 2 scale
/Helvetica findfont 10 scalefont setfont
EOF

$tree = <>; chop($tree); $_ = reverse($tree); s/[()]//g;

$x = 0; $y = 0;
while ($nd = chop()) {
    print "$x $y m ($nd) stringwidth pop -0.5 mul 0 rm ($nd) show\n";
    $xx{$nd} = $x; $x+=20; $yy{$nd} = 10;
}

while ($tree =~ s/\(?([A-Z])([A-Z])\/)?\/1/) {
    print "n $xx{$1} $yy{$1} m\n";
    ($yy{$1} > $yy{$2}) || {$yy{$1} = $yy{$2}}; $yy{$1} += 20;
    print "$xx{$1} $yy{$1} l $xx{$2} $yy{$1} l $xx{$2} $yy{$2} l s\n";
    $xx{$1} = 0.5*($xx{$1} + $xx{$2});
}
print "n $xx{$tree} $yy{$tree} m 0 20 rl s showpage\n";

$rx = 2*$x + 30; $yt = 2*$yy{$tree} + 146;
print "%%BoundingBox: 40 95 $rx $yt\n";
```

Программа PERL используется для построения бинарных деревьев. На входе: (A((BC)D)(EF) дает следующий результат, представленный как PostScript файл, который может быть распечатан на большинстве принтеров и выведен на большинство терминалов.



ПРИМЕР 4.7.

Рассмотрим 4 вида, имеющие гомологичные последовательности ATCC, ATGC, TTCG, TCGG. Приняв количество замен в качестве меры расстояния между каждой парой видов, применим простую процедуру кластеризации для построения дерева организмов.

Матрица расстояний

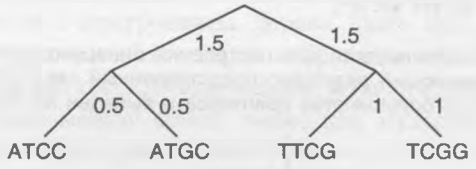
	ATCC	ATGC	TTCG	TCGG
ATCC	0	1	2	4
ATGC		0	3	3
TTCG			0	2
TCGG				0

Поскольку эта матрица симметричная, мы заполнили только верхний правый треугольник. Наименьшее расстояние равно 1 (выделено жирным) между ATCC и ATGC. Поэтому первый кластер будет {ATCC, ATGC }. Дерево имеет фрагмент:

Сокращенная матрица расстояний:

	{ATCC, ATGC}	TTCG	TCGG
{ATCC, ATGC}	0	$\frac{1}{2}(2+3) = 2.5$	$\frac{1}{2}(4+3) = 3.5$
TTCG		0	**2**
TCGG			0

Следующий кластер — {TTCG, TCGG}, расстояние 2. Наконец, объединяя кластеры {ATCC, ATGC }и {TTCG, TCGG}, получаем дерево:



Длины веток установлены в соответствии с правилом: длина ветки между узлом X и Y равна половине расстояния между X и Y.

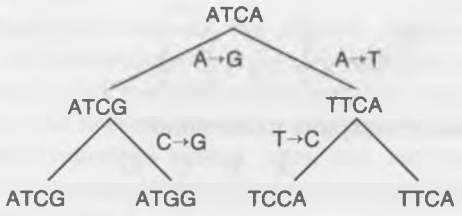
Поскольку длина ветвей пропорциональна времени дивергенции, то таксоны, представленные в вершинах, должны определяться из внешних данных.

Этот процесс построения дерева называется UPGMA (Unweighted Pair Goup Method with Arithmetic mean — метод невзвешенной группировки с арифметическим средним). Модификация метода UPGMA, сделанная N. Siatou и M. Nei, называется методом ближайшего соседа (Neighbour Joining), который разработан для того, чтобы скорректировать неравномерность эволюции на разных ветвях дерева.

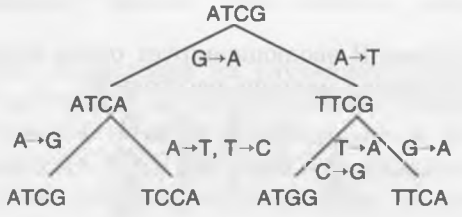
Кладистические методы

Кладистические методы имеют дело исключительно с паттернами наследования, полученными из анализа возможных деревьев таксонов. Они нацелены на выбор правильного дерева и используют детальные модели эволюционных процессов. Наиболее популярными кладистическими методами является метод молекулярной филогении — методы максимальной экономии (Parsimony) и метод наибольшего правдоподобия. Они ориентированы на данные о последовательностях и начинают с множественного выравнивания. Ни метод максимальной экономии, ни метод максимального правдоподобия не ориентируются на анатомические особенности организмов, такие как средний вес взрослой особи.

Метод наибольшей экономии предложен Фитчем (W. Fitch). Он определяет оптимальное дерево так, чтобы минимизировать количество эволюционных событий (мутаций). Например, пусть даны виды, в которых есть гомологичные последовательности ATCC, ATGC, TTCG, TCGG. Дерево



постулирует 4 мутации. Альтернативное дерево



постулирует 7 мутаций. Отметим, что второе дерево постулирует, что замена G → A в четвертой позиции происходит независимо дважды. Предыдущее дерево является оптимальным в соответствии с методом максимальной экономии, поскольку никакое другое дерево не дает меньшего количества мутаций. Во многих случаях несколько деревьев могут давать одинаковое количество мутаций, меньшее, чем другие деревья. В этих случаях метод максимальной экономии не дает единственного ответа.

Метод наибольшего правдоподобия устанавливает количественную характеристику для вероятности мутационных событий, вместо того чтобы просто их подсчитывать. Аналогично методу максимальной экономии метод наибольшего правдоподобия восстанавливает предков в каждом узле дерева, основываясь на вероятностях мутационных событий. Для каждой топологии дерева варьируется скорость эволюции и оптимизируются параметры с тем, чтобы максимизировать правдоподобие порождения наблюдаемых последовательно-

стей. Оптимальное дерево — это такое дерево, которое обеспечивает максимум правдоподобия порождения наблюдаемых данных.

Оба метода (максимальной экономии и наибольшего правдоподобия) лучше ранее описанных кластерных методов. Это показано на независимых наблюдениях, например с помощью классической палеонтологии, когда эти методы давали правильные ответы. Правильность этих методов показана также и на симулированных (модельных) данных.

Проблема переменной скорости эволюции

Предположим, что четыре вида A, B, C, D имеют филогенетическое дерево:

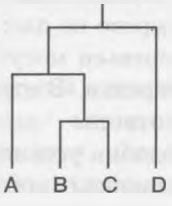
Это дерево соответствует матрице расстояний:

	A	B	C	D
A	0	3	3	3
B		0	2	2
C			0	1
D				0

Предположим, что вид B эволюционирует очень быстро, хотя дерево не меняется. Тогда наблюдаемая матрица расстояний:

	A	B	C	D
A	0	3	3	20
B		0	2	20
C			0	20
D				0

Этой матрице соответствует дерево



Все обсужденные здесь методы не застрахованы от ошибок подобного рода, если скорость эволюции сильно варьирует на разных ветвях дерева. Чтобы проверить есть ли такие вариации, следует принять в рассмотрение внешнюю группу (outgroup) — виды, которые заведомо более удалены от всех видов, для которых строится дерево. Например, если скорость эволюции у приматов

постоянная, то мы ожидаем увидеть примерно одинаковые расстояния между приматами и, скажем, коровой. Если это не так, то неверно предположение о постоянстве скорости эволюции приматов.

Вычислительный анализ

Кладистические методы (максимальной экономии и наибольшего правдоподобия) более точны, чем простые методы кластеризации, такие как UPGMA, но требуют намного больше вычислительных ресурсов для решения разумных задач. Полное количество возможных деревьев, которое должно быть просмотрено кладистическими методами, очень быстро растет с увеличением количества видов. В результате во многих интересных случаях эти методы могут дать только приближенное решение, даже при существенных предположениях.

Поскольку вычисление филогении зачастую приближенное, то важно проверить их. Методы включают:

1. Сравнение филогений полученных из разных характеристик таксонов. Являются ли они согласованными? Если деревья, полученные из разных характеристик имеют согласованные поддеревья, то они, по-видимому правильные, в то время как несогласованные поддеревья неправильные.
2. Анализ подмножеств таксонов может дать тот же ответ по отношению к подмножеству — поддерево, построенное на подмножестве должно соответствовать полному дереву.
3. Формальные статистические тесты, включающие в себя пересчет на подмножестве исходных данных. Эти методы называются *jackknife* и *bootstrap*[1)].

   Jackknife проводит вычисления на случайном подмножестве данных. Для построения филогений по множественному выравниванию отбираются случайные наборы позиций выравнивания и для них проводятся вычисления независимо. Если при этом восстанавливаются одинаковые поддеревья, то дерево признается правильным, а в противном случае — неправильным.

   Bootstrap работает аналогично, но только случайно отобранные позиции могут появляться в выборке несколько раз так, чтобы размер выборки совпадал с исходной выборкой.
4. Если дерево имеет очень длинные ветви, т. е. серьезные основания предполагать, что мы имеем неравномерность эволюции, необходимо использовать внешнюю группу.

[1)] Оценки *jackknife* и *bootstrap* часто относят не ко всему дереву, а к ребрам дерева. Каждое ребро дерева определяет два подмножества листьев — по разные стороны от ребра. Сделав достаточное количество испытаний *bootstrap*, можно оценить процент случаев (восстановленных деревьев по частичным данным), когда эти подмножества совпадают. Так оценивают значимость каждого ребра. Стандартные программы восстановления деревьев указывают на ветвях не только их длины, но и значения *bootstrap*. — *Прим. ред.*

WEB-РЕСУРСЫ: ФИЛОГЕНЕТИЧЕСКИЕ ДЕРЕВЬЯ

Сообщество таксономистов предприняло большие усилия для создания зрелого программного обеспечения. Пакет программ PHYLIP (PHYLogeny Inference Package), созданный Дж. Фельсенстейном представляет собой интегрированную коллекцию многих методов. Программы работают на разных типах компьютеров, свободно распространяются и могут быть легко получены.

Сводка инструментов для филогенетического анализа, включая полезный список Web-ресурсов представлена на сайте `http://evolution.genetics.washington.edu/phylip/software.html` и в работе Whelan S. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. 2001 May; 17(5): 262–72.

Некоторые пакеты построения множественных выравниваний (такие как CLUSTAL-W) имеют возможности построения филогенетического дерева из выравниваний, которые они создают.

Литература

Altschul, S. F. and Koonin, E. V. (1998) 'Iterated profile searches with PSI–BLAST – a tool for discovery in protein databases', *Trends in Biochemical Sciences* 23, 444–7. [Описание наиболее важного инструмента для поиска сходства последовательностей по базам данных.]

Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994) 'Issues in searching molecular sequence databases', *Nature Genetics* 6, 119–29. [Общее описание проблемы разделения информации и интерпретации результатов.]

Eddy, S. (1996) 'Hidden Markov models', *Current Opinion in Structural Biology* 6, 361–5. [Введение в важную математическую область, создающую мощные методы определения удаленных последовательностей и распознавания белковых укладок.]

Efron, B. and Gong, G. (1983) 'A leisurely look at the boostrap, the jackknife, and cross-validation', *The American Statistician* 37, 36–48. [Классическая работа по статистическим методам для калибровки методов распознавания образов.]

Li, W-H. (1997) *Molecular Evolution* (Sunderland, MA, USA: Sinauer.) [Подробное обсуждение эволюции и филогении.]

Penny, D., Hendy, M. D., Zimmer, E. A., and Hamby, R. K. (1990) 'Trees from sequences: Panacea or Pandora's box?', *Australian Systematic Botany* 3, 21–38. [Предостерегающие замечания про определение филогенетических деревьев.]

Упражнения, задачи и компьютерные задания

Упражнение 4.1. Чему равно расстояние Хэмминга между словами DESCLENSION и RECREATION?

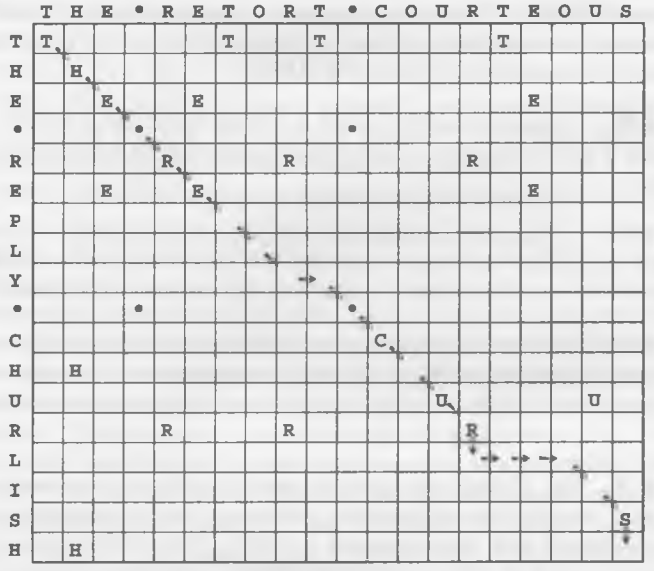
Упражнение 4.2. Чему равно расстояние Левенштейна между словами BIOINFORMATICS и CONFORMATION?

Упражнение 4.3. 'I wasted time and now doth time waste me'. (а) Постройте ожидаемую точечную матрицу для сравнения этой фразы с самой собой. (б) Постройте точную точечную матрицу, отмечая только совпадающие буквы, и сравните результаты.

Упражнение 4.4. Какое значение окна и порога (программа со с. 190) вы предложите для того, чтобы исключить одиночные вхождения в точечной матрице DOROTHYHODKIN, но чтобы остались остальные вхождения.

Упражнение 4.5. Какая замена (W ↔ F или H ↔ R) более вероятна для матриц (а) PAM150 и (б) BLOSUM62.

Упражнение 4.6. Какому выравниванию соответствует показанный путь по матрице сходства?



Упражнение 4.7. Допустим, что при планировании вашего путешествия из Мальмо в Тромсо (с. 206) у вас есть специальное требование посетить Уппсала. Можете ли вы оптимизировать маршрут так, чтобы ваши затраты были минимальными?

Упражнение 4.8. Как бы вы использовали точечную матрицу, чтобы получить палиндромную последовательность ДНК, частично присутствующую в обеих цепях, как в специфических сайтах рестрикции эндонуклеаз?

Упражнение 4.9. Модифицируйте программу PERL на с. 190, которая рисует точечные матрицы, принимая последовательности в формате FASTA.

Упражнение 4.10. Какому значению P соответствовало бы Z-score, равный 1 в нормальном распределении?

Упражнение 4.11. Для каждого из выравниваний на рис. 4.2а определите, относится ли оно к серой зоне, более подобна, чем серая зона или менее подобна, чем серая зона.

Упражнение 4.12. На рис. 4.2а изображено выравнивание последовательностей папаина из папайи и фруктового актинидина киви, а также приведена соответствующая точечная матрица. Выравнивание указывает на два места, на которых один из остатков удален из последовательности папайи, и одно место, на котором

удален остаток из последовательности актинидина. На фотокопии рис. 4.2а на точечной матрице укажите позиции этих вставок и делеций.

Упражнение 4.13. Представьте, что рандомизация последовательности не самый лучший способ создания контрольной группы последовательностей для анализа статистической достоверности парного выравнивания последовательностей, потому что появление в природной последовательности дипептидов или трипептидов имеет неодинаковые вероятности. Какой улучшенный способ для создания контрольной группы последовательностей вы бы предложили?

Упражнение 4.14. Сравнение последовательностей ДНК в гомологичных хромосомах разных людей показывает, что в среднем различны 1 из 700 пар оснований некодирующей ДНК. 95% генома человека не кодирующие. Оцените число полиморфизмов в человеческом геноме, чтобы получить некоторое представление о количестве потенциальных маркеров ДНК.

Упражнение 4.15. Покажите вычисления, которые приводят в клетку со значением 65 в примере 4.6. Что является следствием наблюдений того, что из нее выходят две стрелки?

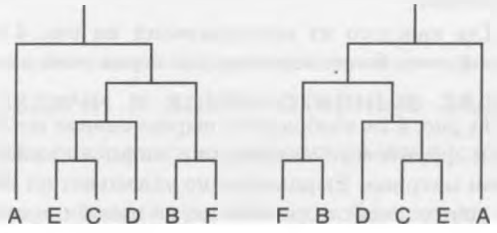
Упражнение 4.16. α-Спираль, сформированная основаниями 32–49 в тиоредоксине *E. coli*, была прервана. На фотокопии рис. 4.5 укажите, где находится этот перерыв. У какого из этих оснований, вероятно, произошло искажение?

Упражнение 4.17. У каких из этих оснований в тиоредоксине *E. coli* происходит поворот в конформации цепи, который не соответствует участкам, в (или возле) которых произошли делеции в множественном выравнивании последовательности?

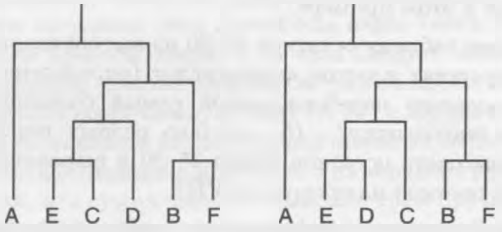
Упражнение 4.18. (а) Используя простой метод перебора, определите, какой гексапептид дает наибольшее возможное значение для совпадений позиций 25–30 в таблице значений для тиоредоксина (с. 219). (б) Используя схему значений, распределенных между всеми аминокислотами, согласно матрице BLOSSUM62, сравните значение этого гексапептида со значением гексапептида VDFSAE.

Упражнение 4.19. (а) Сделайте перебор участков из остатков № 90–95 тиоредоксина, аналогично таблице на с. 219. Какому распределению соответствует последовательность, выровненная по этим основаниям, образующая простой профиль, используя переборы как веса? (б) ISSAVK (в) FVGAKE.

Упражнение 4.20. (а) Являются ли топологии следующей пары деревьев идентичными?



(б) Идентичны ли топологии следующей пары деревьев?



Упражнение 4.21. Нарисуйте все возможные укорененные деревья, связывающие три таксона. Сколько таких деревьев можно создать?

Упражнение 4.22. Для финального графа в примере 4.7 определите,- как попала туда ветвь, длиной 1.5, исходящая из узла, присоединяющемуся к кластерам {ATCC, ATGC} и {TTCG, TCGG}?

Упражнение 4.23. Используя исходную матрицу расстояний и дерево, полученное в примере 4.7, для каждой пары видов сравните исходные расстояния между ними, с суммой длин путей, соединяющих их в дереве.

Упражнение 4.24. Приведите пример полностью связного графа, не являющегося деревом.

Упражнение 4.25. Последовательности митохондриальной ДНК, взятых у европейского, африканского и азиатского скота, показывают, что европейские и африканские породы более тесно связаны друг с другом, чем индийские. Чтобы исключить возможность появления этого результата, как ошибочного результата дифференциальной скорости эволюции двух поколений, предложите разумный выбор внешней группы.

Упражнение 4.26. Преобразуйте дерево, полученное в примере 4.7, в шкалу с расстояниями между делениями, пропорциональными назначенным длинам ветвей.

Упражнение 4.27. В методе динамического программирования для выравнивания двух последовательностей, длины n, мы заметили, что время работы алгоритма пропорционально n^2. Как зависит от n требование к необходимой памяти при обычной реализации алгоритма:

а) если мы хотим определить оптимальное выравнивание, то должна ли храниться информация об обратном пути?

б) если мы хотим получить только счет и не выравнивание, то нужно ли хранить информацию об обратном пути? (Заметьте, что едва отличимые способы выполнения алгоритма — существенно уменьшают требования свободного места, по сравнению с обычной реализацией.)

Задача 4.1. Изобразите точечную матрицу сходства для последовательностей из генома вируса карликовости пшеницы `ttttcgtgagtgcgcggaggctttt` против самого себя. Является ли эта последовательность настоящим палиндромом?

Задача 4.2. (а) Как бы вы изменили алгоритм из разд. 5, чтобы найти оптимальные вхождения относительно короткого паттерна $A = a_1 a_2 \cdots a_n$ в длинную последовательность $B = b_1 b_2 \cdots b_m$ где $n \ll m$. (В участках последовательности B, которые предшествуют и совпадают с последовательностью A, нет делеций.) Это соответствует мотивам выравнивания, как описано в гл. 1. (б) Переделайте расчеты примера 4.6 для выравнивания строк `ggaatgg` и $B =$ `atg`, как задачу выравнивающегося мотива, используя ту же систему счета: совпадения — 0, несовпадения — 20, штраф за открытие внутренней делеции — 25, про-

должение — 22. (в) Как результаты, которые вы получили, отличаются от тех, которые получили в этом примере?

Задача 4.3. Создайте таблицу остатков 25–30 из выравнивания тиоредоксина, как на с. 219, но в терминах классов аминокислот (определено на с. 216). (а) Используя метод простого перебора, какой самый большой счет может получить какой-либо гексопептид? (б) Сколько разных гексапептидов дает этот счет? (в) Каковы счета остатков номер 25–30 в выравнивании каждой последовательности на цветной иллюстрации VII?

Задача 4.4. Как бы вы могли модифицировать метод построения профилей, чтобы сохранить способность отбирать тиоредоксины не млекопитающих животных, если в таблицу было добавлено большое количество дополнительных родственных последовательностей млекопитающих. Рассмотрите (а) методы, которые пытаются убрать избыточность игнорируя конкретные последовательности; (б) методы, которые сохраняют все последовательности, но включают матрицу весов, чтобы сбалансировать представленность близкородственных последовательностей.

Задача 4.5. Напишите программу на языке PERL для создания профиля, используя множественное выравнивание последовательностей и оценки веса выравнивания запрашиваемых последовательностей, используя матрицу BLOSUM62. Учтите, что запрашиваемая последовательность уже выровнена до того, как она подается на вход программе.

Задача 4.6. (а) Напишите программу на языке PERL, которая читает две символьные строки выводит на экран все совпадения длиной 5 символов. Протестируйте эту программу на следующих строках:

`My.care.is.loss.of.care,.by.old.care.done, и`

`Your.care.is.gain.of.care,.by.new.care.won`

(б) Усовершенствуйте эту программу, чтобы она расширяла и объединяла найденные совпадения в более длинные участки, содержащие точно совпадающие пентапептиды без делеций с не более чем 25% несовпадений по всей длине последовательности.

Задача 4.7. Продолжите предыдущую задачу написанием программу на языке PERL, чтобы проиллюстрировать, основываясь на точечных матрицах, последовательность действий алгоритма типа BLAST на этапе, когда алгоритм (а) обнаруживает все подстроки длины 5, (б) продолжает их до максимальных длины совпадений, (в) объединяет их, чтобы сформировать совпадающий участок с не более, чем k несовпадений. Вы можете использовать программу PERL для точечных матрицы, приведенной в тексте.

Задача 4.8. Одноцепочечные РНК такие, как тРНК, принимают конформации, содержащие участки стебель-петля, в которых участок цепи накладывается на самого себя для формирования двуцепочечной спирали, образуя комплементарные пары основания с антипараллельными тяжами. Как программа, которая обнаруживает палиндромы, может быть использована для анализа последовательностей РНК тех участков, которые способны формировать идеальные, т. е. без несовпадений, мотивы стебель-петля?

Задача 4.9. Напишите программу для живой иллюстрации последовательности действий алгоритма типа BLAST, как описывается в задаче 4.7. Найдите в Интернете примеры, иллюстрирующие алгоритм поиска строк. (Эта задача требует опыта программирования на компьютере.)

Задача 4.10. Представьте, что у вас есть пара игральных костей, красная и зеленая. Определите состояние этих костей как пары чисел: число, появляющееся на верхней стороне красной кости, а за ним следует число, появляющееся на верхней стороне зеленой. Вместо того, чтобы кидать кости, следуйте от состояния к состоянию, переворачивая каждую кость на 90° в любом направлении с равной вероятностью. За состоянием, когда на верху одной из костей находится 6, может с равной вероятностью следовать 2, 3, 4 или 5 на верхней грани кости. (Игральная кость устроена так, что сумма чисел по каждым трем противоположным граням равно 7. Поэтому вероятность того, что за 1 следует 6 равна 0 потому, что это требует поворота на 180°.) Вероятность порождения последовательности $6, 2, 6, 4$ равна $(1/4)^4 = 1/256$. Вероятность порождения последовательности $6, 2, 5, 4$ равна 0, так как переход из 2→5 не допустим, а вероятность порождения последовательности $6, 6, 2, 3, 4$ равна 0, так как система должна изменяться, т. е. 6 не может следовать за 6.

Процедура определяет марковский процесс первого порядка.

Напишите программу, отвечающую на следующие вопросы: предположим, что сначала подбрасывали 4 красных и 3 зеленых кости. (а) Какова вероятность еще одного состояния, в котором числа складываются в 7, укладываясь в 5 действий? (б) Если первоначальное состояние было бы 8, то какова бы была вероятность другой 8 появиться перед 7?

Задача 4.11. Покажите, что любой ненаправленный граф, обладающий одним из следующих свойств, имеет также другое: (1) имеет единственный путь между двумя узлами; (2) граф не содержит циклов.

Задача 4.12. Сколько путей на рис. 4.8 проходит от старта до финиша? Посчитайте их в каждом из следующих путей:

(а) Грубая сила — перечислите все возможности. Это упражнение фактически не требует умственных затрат. Оно показывает, что это действительно не так трудно, как кажется на первый взгляд. Во вторых, оно показывает, как вы будете ощущать паттерны при выполнении упражнения.

(б) На рис. 4.8 (1) посчитайте число путей от старта до A и от A до финиша. Множество этих чисел вместе дает общее количество путей от старта до финиша, которые проходят через A. (2) Затем посчитайте число путей от cTAPTA до B и от B до финиша. Как эти числа связаны между собой? Множество этих чисел дает общее количество путей от старта до финиша, которые проходят через B. (3) С помощью компьютера посчитайте общее количество путей от старта до финиша, как сумму чисел путей из старта до финиша, которые проходят через A, B, C и D.

(в) Догадайтесь, как добраться от старта до финиша за 6 шагов, обязательно включая 3 поворота налево и 3 поворота направо (вы не можете остановиться в правой половине). Различные выборы поворотов налево и направо соответствуют разным путям. Посчитайте число путей, решив сколько путей получится, если выбрать 3 шага, на которых вы поворачиваете налево (затем вы должны на остальных трех шагах повернуть направо). Назначая 3 первых поворота из 6 шагов, во-первых, вы можете выбрать один из шести шагов для одного поворота налево, затем один из пяти оставшихся шагов для следующего поворота налево и затем один из четырех следующих шагов для последнего левого поворота. Однако результатом этих чисел будет являться значение вероятностей, так как оно включает тот же набор шагов, обусловленный различным порядком. Эти триплеты могут возникнуть

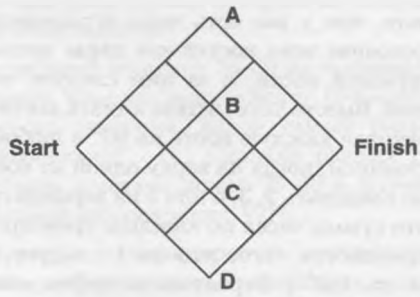


Рис. 4.8. Подсчет путей в конечной решетке

в шести различных путях, их необходимо исправить. Результатом является биномиальный коэффициент $\binom{6}{3} = 6!/(3!3!)$. Объясните происхождение этих результатов.

Задача 4.13. Для конечного дерева в примере 4.7 предоставьте список возможных предков внутренних узлов, выбранных исходя из критерия максимальной экономии. Если ли тут какая-то неопределенность?

Задача 4.14. Удобное условное изображение дерева использует компактный вид записи с помощью круглых скобок, чтобы отражать кластеры. (а) Изобразите следующее в виде укорененного дерева: (A(BC)D). (б) Напишите скобочную формулу для деревьев, показанных в примере раздела «Методы кластеризации» (см. с. 233)

Задача 4.15. Добавьте адекватное количество компонент к программе на PERL для рисования деревьев.

Задача 4.16. Напишите программу на PERL для метода UPGMA построения филогенетического дерева, основываясь на матрице расстояний.

Интернет-задание 4.1. Найдите последовательность гена 6-й субъединицы митохондриальной АТФ-азы из организма атлантической рыбы-ведьмы (*Myxine glutinosa*). Нарисуйте точечную диаграмму с гомологичным геном из организма морской миноги (*Petromyzon marinus*). Прокомментируйте полученное сходство и сравните со сходством между последовательностями миноги и рыбы-собаки, показанными на с. 188.

Интернет-задание 4.2. Используйте BLAST и PSI-BLAST для аминокислотной последовательности папаина из папайи. Какие их гомологий, появляющихся на рис. 4.2а, успешно обнаруживаются с помощью BLAST? Какие — с помощью PSI-BLAST?

Интернет-задание 4.3. Используйте PSI-BLAST для последовательности папаина из папайи (см. предыдущее Интернет-задание). На результатах сравнения с человеческим прокатепсином L, укажите регионы локального выравнивания на фотокопии точечной диаграммы (рис. 4.2б).

Интернет-задание 4.4. Найдите структуры тиоредоксинов, появляющихся на таблице выравниваний во вклейке VII, из организмов отличных от *E. coli*. На фотокопии таблицы выравнивания укажите регионы спиралей и тяжей из листов в соответствии с данными записей PDB и сравните их с регионами тяжей и спиралей тиоредоксина *E. coli*.

Интернет-задание 4.5. Выровняйте аминокислотные последовательности папаина из папайи и гомологов, показанных на рис. 4.2а, используя CLUSTAL-W или T-coffee. Сравните результаты с таблицей выравниваний базы данных Pfam, основанной на скрытых марковских моделях, и со структурными выравниваниями на рис. 4.2а.

Интернет-задание 4.6. Может ли PSI-BLAST идентифицировать гомологию между доменами иммуноглобулинов, доменами эндоглюконазы C *Cellulomonas fimi* и IgA-рецептором *Streptococcus agalactiae*?

Интернет-задание 4.7. (а) Может ли PSI-BLAST идентифицировать родство между уреазой из организма *Klebsiella aerogenes*, фосфотриэстеразой из *Pseudomonas dinimuta* и аденозиндеаминазой мыши? (б) Сравните выравнивания этих трех последовательностей, сделанные с помощью DALI и сделанные с помощью CLUSTAL-W или T-Coffee.

Интернет-задание 4.8. Гормоны роста большинства млекопитающих имеют очень сходные аминокислотные последовательности. (Гормоны роста альпака, собаки, кошки, лошади, кролика и слона — каждый отличается от гормонов свиньи не более чем на 3 позиции из 191). Человеческие гормоны роста отличаются гораздо сильнее — на целых 62 позиции. Эволюция гормонов роста ускорялась резко в той линии, которая вела к человеку. Путем поиска выравнивания последовательностей гормонов роста из организмов близкородственных человеку видов и наших предков было определено, *где* в эволюционном древе человека имело место ускорение эволюции гормонов роста.

Следующая серия Интернет-заданий придумана, чтобы поместить виды рода человек в их биологических контекст путем анализа последовательностей из близких и дальних родственников, чтобы показать некую вариабельность генетической информации, которая была использована для исследования филогенетических родственных связей.

Интернет-задание 4.9. Ныне живущие виды, наиболее близкородственные к людям — это человекообразные обезьяны и мартышки. Alu-элементы — один из типов SINE (короткие распределенные генетические элементы, Short Interspersed Nuclear Elements), полезные в качестве специфических маркеров. Некоторые Alu-элементы участвуют в регуляции экспрессии генов. В частности — паратиреоидного гормона, гематопоетического клеточно-специфичного FceRI-γ рецептора, специфичного для центральной нервной системы никотинового ацетилхолинового рецептора α_3 и специфичного для Т-клеток CD8α. На основе этого наблюдения получите филогенетическое дерево для человека, шимпанзе, гориллы, орангутанга, бабуина, мартышки-резус и макаки.

Интернет-задание 4.10. Люди — приматы, и помимо нас к ним относятся человекообразные обезьяны, мартышки, лемуры и тарсиры. На основе кластера гена β-глобина человека, шимпанзе, европейской обезьяны, американской обезьяны, лемура и долгопята получите филогенетическое дерево этих групп.

Интернет-задание 4.11. Приматы относятся к млекопитающим, и этот класс мы разделяем с сумчатыми и первозверями. Сохранившиеся в настоящее время сумчатые живут преимущественно в Австралии, за исключением опоссума, найденного также в Северной и Южной Америке. Ныне живущие первозвери — это всего два вида животных из Австралии: утконос и ехидна. Используя полные митохондриальные геномы человека, лошади (*Equus caballus*), кенгуру (*Macropus robustus*), американского опоссума (*Didelphis virginiana*) и утконоса (*Ornitborbynchus anatinus*), нарисуйте эволюционное дерево и подпишите длины

ветвей. С кем первозвери более близкородственны: с плацентарными или с сумчатыми?

Интернет-задание 4.12. Млекопитающие — один из классов подтипа позвоночных, к которому также относятся рыбы, акулы, птицы и рептилии, амфибии и примитивные кистперые рыбы (например, латимерия). Используя последовательности цитохромов с и панкреатических рибонуклеаз, постройте эволюционное древо для латимерии (*Latimeria chalumnae*), гигантской белой акулы (*Carchadoron carcharias*), прыгающего тунца (*Katsuwonus pelamis*), морской миноги (*Petromyzon marinus*), лягушки (*Rana pipens*) и нильского крокодила (*Crocodylus niloticus*).

Интернет-задание 4.13. Позвоночные относятся к хордовым, типу, который также включает ланцетников (маленьких рыбоподобных морских организмов, например амфионты) и бесчелюстных рыб (у которых еще нет настоящего позвоночника, например миног). Как и в других организмах с билатеральной симметрией (в том числе у насекомых), HOX-гены позвоночных кодируют семейство белков, связывающихся с ДНК. Экспрессия этих генов меняется в направлении оси, проходящей по телу от головы к хвосту, регулируя план план строения тела. Разумеется, существуют потрясающие карты порядка положения генов в хромосоме, порядка их активации вдоль тела и относительным временем развития начала их активности.

В течение эволюции позвоночных имели место крупномасштабные геномные дупликации, связанные предположительно с развитием гораздо более сложной архитектуры тела, как и предсказывал С. Охно в 1970. Геномы насекомых и ланцетника имеют один HOX-кластер. У рыбы-зебры таких кластеров 7, что можно интерпретировать в терминах дупликаций $1 \to 2 \to 4 \to 8$ как редукцию одного кластера ($8 \to 7$).

Найдите число HOX-кластеров у человека и миноги, постройте множественное выравнивание последовательностей, чтобы иметь соответствия между индивидуальными генами, и, исходя из этих данных, - получите филогенетическое дерево для ланцетника, миноги, рыб и млекопитающих.

Интернет-задание 4.14. Хордовых относят к вторичноротым (см. с. 38) — группе, которая помимо нас включает оболочечников (пример: морские squirts), hemichordates (пример: амфионты) и иглокожих (пример: морская звезда). Есть систематические различия между этими тремя типами в их митохондриальном генетическом коде. Определите примеры организмов из каждого типа, чьи аминокислоты, соответствуют кодонам ATA и AGA. Исходя из этого получите филогенетическое дерево для четырех типов вторичноротых.

5 Структура белков и разработка лекарств

Введение

Огромное разнообразие трехмерных структур и функциональных свойств белков реализуется на базе молекул, подчиняющихся определенным принципам структурной организации. Развивая химические аналогии, белки можно срав-

нить с гирляндами огоньков на рождественской елке: каждый белок представляет собой линейную (т. е. неразветвленную) полимерную цепь, к которой с определенным интервалом присоединены различные боковые радикалы аминокислотных остатков (рис. 1.6). Проводок, нанизывающий огоньки, соответствует основной цепи или полипептидному остову, а череда разноцветных лампочек — очередность индивидуальных свойств боковых радикалов аминокислот.

Аминокислотная последовательность белка определяется последовательностью нуклеотидов соответствующего гена. Трехмерная структура белковой молекулы формируется без дальнейшего участия нуклеиновых кислот и определяется лишь ее собственной аминокислотной последовательностью. Сворачивание полипептидной цепи с образованием нативной конформации белка происходит спонтанно.

Каким же образом аминокислотная последовательность кодирует трехмерную структуру? Каждый из вариантов сворачивания основной цепи приводит к образованию тех или иных контактов между аминокислотными остатками. Взаимодействия боковых радикалов и основной цепи друг с другом и с растворителем, а также ограничения, накладываемые на подвижность боковых радикалов, определяют относительную стабильность различных конформаций. В этом заключается проявление второго закона термодинамики, согласно которому системы при постоянных температуре и давлении приходят к состоянию равновесия, в котором достигается компромисс между энергетическим «комфортом» (низкой энтальпией H) и «свободой» (высокой энтропией S), приводящий к минимуму энергии Гиббса $G = H - TS$ (где T — абсолютная температура). (В человеческих отношениях брак также компромисс подобного рода.)

Эволюция белковых структур шла таким образом, что для произвольного белка существует лишь единственный способ укладки основной цепи в пространстве, который термодинамически гораздо выгоднее остальных возможных конформаций. Это состояние называется нативным. И если бы мы могли достаточно точно рассчитывать энергии и энтропии и исследовать таким образом достаточно большой ансамбль конформаций, чтобы наверняка найти среди них наилучшую, то было бы возможным предсказание структуры белка, исходя только из его аминокислотной последовательности *a priori* на базе физико-химических законов. В этом направлении наблюдается существенный прогресс, хотя окончательная цель еще не достигнута.

В нативном состоянии основная цепь белка образует трехмерную структуру. К настоящему моменту мы знаем структуру примерно 15 000[1)] белков (включая множество идентичных белков, а также точечных мутантов), и видим среди них огромное разнообразие способов пространственной укладки полипептидных цепей. Первая проблема, связанная с анализом этих структур, касается визуализации. На примере небольшого белка ацилфосфатазы (рис. 5.1) показаны трудности интерпретации как полностью детализированной презентации структуры, так и тех упрощенных картинок, которые выдают компьютерные программы при визуализации. Опытный в вопросах молеку-

[1)] На 11 сентября 2007 г. более 42 000 белковых структур. — *Прим. ред.*

лярного моделирования исследователь сможет комбинировать «картинки», чтобы получить различные структурные области с необходимой степенью детализации.

В середине рис. 5.1 в молекуле ацилфосфатазы направление основной цепи. Два участка на переднем плане имеют форму спиралей (похожа на вывеску[1]) на парикмахерской) с практически вертикальными осями. В молекуле имеются четыре тяжа в одной плоскости. Они ориентированы практически вертикальны. Сборка четырех тяжей в одной плоскости стабилизирована взаимодействиями между ними с образованием β-листа. На нижней части рисунка спирали и листы схематически изображены цилиндрами (спирали) и стрелками (плоские фрагменты — тяжи). На верхней части рис. 5.1 приведены детализованные изображения структуры, включая главную цепь и боковые цепи (эти рисунки позволяют понять, почему так важно по возможности упрощать структуру даже в случае малых белковых молекул.

Стабильность и сворачивание (фолдинг) белков

Несмотря на то что пока еще невозможно предсказать структуру белков, исходя только из общих физических принципов, мы уже понимаем природу тех взаимодействий, которые определяют эти структуры.

Чтобы образовать нативную структуру, все взаимодействия между остатками в белке должны быть оптимизированы при заданной геометрии расположения основной цепи в пространстве. Существование предпочтительных конформаций основной цепи обусловливает возникновение в процессе сворачивания периодических структурных паттернов: спиралей, тяжей, которые, взаимодействуя, укладываются в листы, и несколько стандартных типов поворотов.

Графические представления по Сасисекхаран—Рамакришнан—Рамачандран для описания разрешенных конформаций основной цепи

В хорошем приближении конформация основной цепи каждого остатка, если это не глицин, ограничена двумя дискретными состояниями.

Фрагмент линейной полипептидной цепи, общий для всех белковых структур, показан на рис. 5.2. Вращение разрешено вокруг одинарных связей N–Cα и Cα–C в каждом остатке (за исключением пролина). Углы ϕ и ψ вращения вокруг этих связей и угол вращения ω вокруг пептидной связи, определяют конформацию остатка. Сама по себе пептидная связь стремится быть плоской, и для ω имеются два разрешенных значения: *транс*, $\omega \approx 180°$ (обычно) и *цис* $\omega \approx 0°$ (редко, и почти во всех случаях у остатка пролина). Последовательность углов ϕ, ψ и ω для всех остатков белка определяет конформацию белковой цепи.

[1] Распространенная парикмахерская вывеска имеет форму шеста (barber pole), раскрашенного по спирали красными и белыми полосками. — *Прим. перев.*

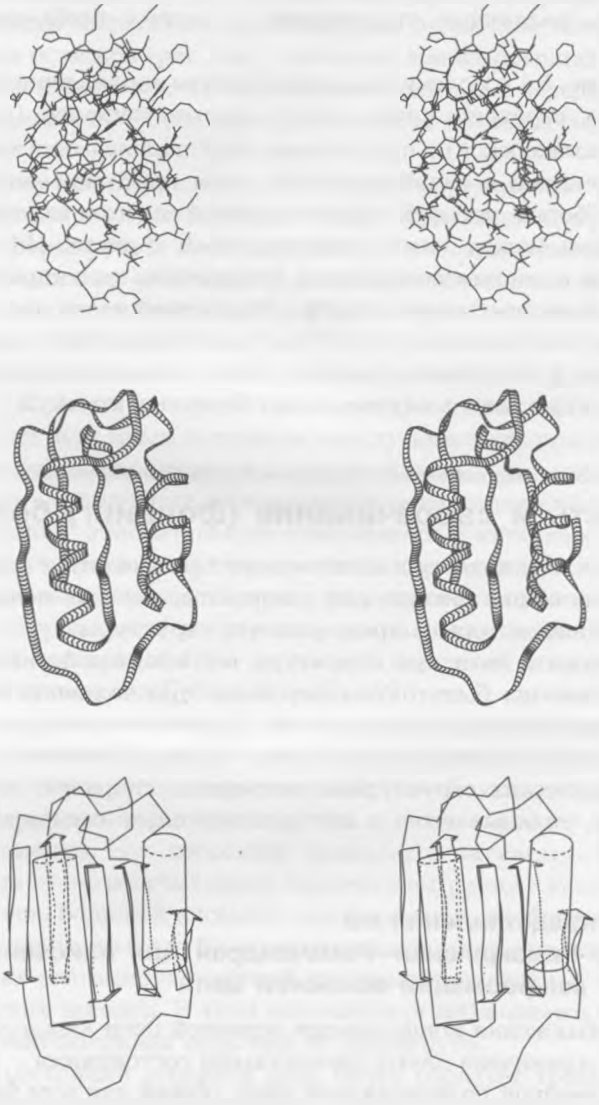


Рис. 5.1. Структура белков достаточно сложна, поэтому для ее представления был разработан специальный аппарат. На данном рисунке изображен достаточно простой белок ацилфосфатаза в трех различных уровнях упрощения. *Сверху*: подробная модель скелета белка. *В центре*: ход основной цепи представлен сглаженной кривой, стрелки обозначают направление цепи. *Внизу*: схема, где цилиндры — спирали, а стрелки — тяжи листов. «Прозрачность» объектов достигается с помощью сплошных и прерывистых (за плоскостью рисунка) линий. Можно попытаться мысленно наложить различные презентации друг на друга, как бы поворачивая страницу на 90°, и представить себе объемные структуры (не слишком при этом напрягаясь)

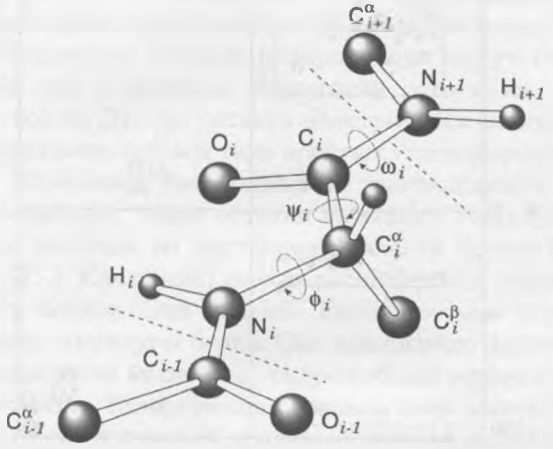


Рис. 5.2. Определение конформационных углов полипептидной цепи

Принцип, согласно которому два атома не могут занимать одно и то же место в пространстве, ограничивает набор возможных значений конформационных углов. Разрешенные значения ϕ и ψ для $\omega = 180°$ попадают в определенные области графика, который называют картой Сасисекхаран—Рамакришнан—Рамачандрана или, обычно сокращая, «картой Рамачандрана» (см. рис. 5.3). Сплошные линии на ней ограничивают энергетически предпочтительные области значений ϕ и ψ, а прерывистые — стерически неразрешенные зоны. Конформации большинства аминокислотных остатков попадают либо в α_R, либо в β-зону. Глицин может быть и в других конформациях. В частности, он может формировать левозакрученную спираль α_L. Рисунок 5.3. показывает типичное распределение конформаций остатков в хорошо разрешенной структуре белка. Большинство остатков либо попало внутрь, либо лежит вблизи разрешенных областей, хотя некоторые при сворачивании оказались в энергетически менее предпочтительных состояниях.

Разрешенные области на карте Рамачандрана соответствуют стандартным конформациям. Чередование остатков в α-конформации (обычно 6–20 в нативном состоянии глобулярного белка) приводит к образованию α-спирали. Чередование остатков в β-конформации приводит к образованию β-тяжа. Два или более β-тяжа могут взаимодействовать в плоскости с образованием β-листа, как в ацилфосфатазе (рис. 5.1). Спирали и листы являются стандартными «заготовками», структурными компонентами, которые образуют пространственную структуру большинства белков. Их стабилизируют относительно слабые взаимодействия — *водородные связи* между атомами основной цепи (рис. 1.7). В некоторых фибриллярных белках все остатки принадлежат к какому-то одному типу: шерсть содержит α-спирали, а шелк — β-листы. Амилоидные фибриллы, возникающие из многих белков при различных болезнях, также содержат огромные β-листы.

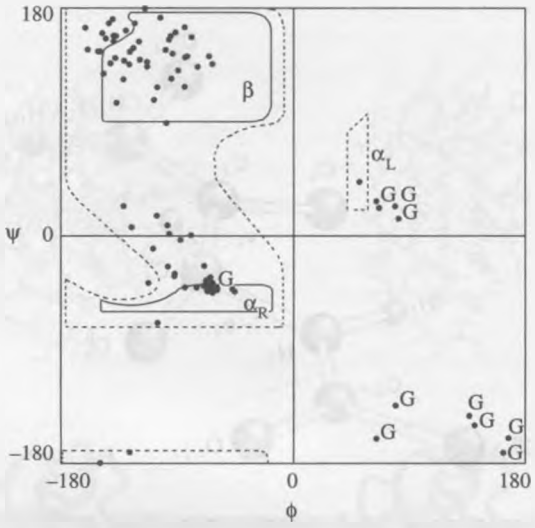


Рис. 5.3. Карта Сасисекхаран—Рамакришнан—Рамачандрана для ацилфосфатазы (PDB-код 2ACY). Обратите внимание, что остатки группируются в α- и β-участках, а большинство исключений приходится на остатки глицина

Типичные глобулярные белки содержат несколько α-спиральных участков и/или β-листов, связанных друг с другом *поворотами*. Обычно конец α-спирального участка или β-листа появляется на поверхностном домене белковой структуры. Они соединены петлями, т. е. участками, в которых цепь меняет направление и уходит внутрь структуры.

Многие, но не все петли образуются короткой цепью, выходящей на поверхность, как правило, содержащей заряженные или полярные остатки.

Как основная цепь выбирает свое структурное состояние возможных допустимых конформаций? Уникальность каждого белка заключается в уникальной последовательности аминокислот (их боковых радикалов). Поэтому взаимодействия между боковыми остатками и определяют конформацию основной цепи.

Боковые остатки

Именно боковые остатки аминокислот создают то разнообразие физико-химических свойств, которое порождает различные способы укладки. Боковые радикалы 20 аминокислот различаются по следующим параметрам:

Размер. Самая маленькая аминокислота глицин содержит в боковой цепи только один атом водорода, а одна из самых больших аминокислот фенилаланин — бензильный радикал.

Электрический заряд. Некоторые боковые остатки несут положительный или отрицательный заряд при нейтральном pH. Asp и Glu заряжены от-

рицательно, Lys и Arg — положительно. (Разноименно заряженные остатки могут попарно взаимодействовать, образовывая *солевые мостики*).

Полярность. Некоторые остатки полярны; они могут образовывать водородные связи как с другими полярными остатками, так и с основной цепью или с водой. Другие остатки электрически нейтральны. Некоторые содержат химические группы типа простых углеводородов, таких как метан или бензол. Поскольку взаимодействие углеводородов с водой термодинамически невыгодно, такие остатки называют гидрофобными. Слипание гидрофобных остатков во внутренней области белков было предсказано Каузманом (W. J. Kauzmann) еще до расшифровки первых пространственных структур белков. Это явление вносит весьма существенный вклад в стабилизацию структуры белка. Оно аналогично формированию капелек масла на поверхности воды. (см. гидрофобный эффект на с. 254).

Форма и жесткость. Конформация боковой цепи зависит от ее химической структуры и от ее внутренней конформационной подвижности.

Стабильность и денатурация белков

Какие химические силы стабилизируют нативную структуру белков? Что это за процессы, с помощью которых белок из множества денатурированных конформаций переходит в единственное нативное состояние?

Чтобы ответить на эти вопросы, биохимики исследовали процесс денатурации белков в ответ на повышение температуры, увеличивающуюся концентрацию мочевины или гидрохлорида гуанидина (обычные методы денатурации). Некоторые измерения носят статический характер: определение содержания нативной и денатурированной форм белка в состоянии равновесия при различных условиях или количество тепла, высвобождаемого в различные моменты при переходе белка из одного состояния в другое. Другие измерения носят кинетический характер: оценка скорости сворачивания или разворачивания белка или идентификация временно образующихся переходных структур.

Важно отметить, что стабильность молекул белков невелика. Нативное состояние глобулярного белка обычно всего на 20–60 кДж/моль (5–15 ккал/моль) стабильнее, чем денатурированное. Это соответствует энергии одной или двух водородных связей между молекулами воды[1)].

Точно не известно, почему белки обладают столь низкой стабильностью. Некоторые полагают, что это способствует круговороту белков. Другие считают, что белки стабильны настолько, насколько это необходимо, поэтому «зачем заботиться» о дальнейшей оптимизации стабилизирующих взаимодействий. Более того, мы знаем, что взаимодействия, которые обеспечивают стабильность нативных белков, действительно способны поддерживать белковые структуры со значительно более высокой стабильностью[2)].

[1)] Необходимо уточнять, к каким условиям относится энергия водородных связей. Энергия в несколько ккал/моль может соответствовать разрыву водородной связи только в сильно неполярной среде или вакууме. — *Прим. ред.*

[2)] Степень термодинамической стабильности связана со скоростью достижения оптимальной структуры — более стабильные белки медленнее приходят к нативной конформации; см. замечательную книгу А. В. Финкельштейна «Физика белка». — *Прим. ред.*

Гидрофобный эффект

Различие между боковыми цепями аминокислотных остатков в их предпочтении водного или масляного окружения — один из главных принципов сворачивания белковой структуры.

Что такое гидрофобный эффект? Разделение фаз в смеси масла с водой — например, в приправе к салату — один из типичных примеров; другим примером может служить уменьшение растворимости газов (в отличие от твердых тел) в воде при повышении температуры. Читатели, имеющие чайник со свистком, слышали низкий звук, предшествующий кипению — это объясняется выходом из воды растворенного в ней воздуха при нагревании.

В чем причина гидрофобного эффекта? Холодная вода — высокоструктурированная жидкость. Ее молекулы образуют между собой множество водородных связей, которые определяют высокую температуру испарения воды и ее низкую плотность. Но молекулы воды располагаются даже более упорядоченно вокруг молекул растворенного вещества, чем в чистой воде. Молекула метана в воде (метан растворим в воде плохо, но все же достаточно, чтобы можно было изучать) окружена оболочкой из молекул воды, называемой клатратным комплексом. Как результат — растворение метана в воде повышает упорядоченность раствора, понижая его энтропию. Естественное стремление системы к состояниям с высокой энтропией препятствует растворению метана в воде. Именно поэтому метан и другие углеводороды лишь слабо растворимы в воде. Растворимость неполярных газов уменьшается при повышении температуры — с уже небольшого значения в холодной воде — потому, что при повышенной температуре энтропия играет еще более важную роль в определении равновесного состояния.

Гидрофобный эффект в водных растворах простых неполярных веществ был хорошо известен физикохимикам, когда Каузман в 1959 г. обнаружил его большое значение для белковой структуры.

Неполярные боковые цепи белков ведут себя в растворе как масло. Их взаимодействия с водой невыгодны. Каузман предположил, что они будут сближены внутри белка, изолированно от растворителя, подобно молекулам в капле масла. Эта модель капли масла была подтверждена с помощью РСА (ренгеноструктурного анализа) структур глобулярных белков. Теперь мы узнали также, что плотность упаковки белков имеет большое значение и что точнее рассматривать свернутый белок как кристалл, чем как органическую жидкость. Но гидрофобный эффект при этом не теряет своей значимости.

Как следствие гидрофобного эффекта, заряженные остатки практически исключаются из внутренностей белковой глобулы; в редких случаях они формируют внутренние солевые мостики. Очевидно, что остов должен проходить внутри белковой структуры и нести с собой полярные атомы N и O, способные взаимодействовать с другими полярными атомами главной цепи и полярными боковыми цепями, такими как, например, треонин или аспарагин. Таким образом, внутри глобула не полностью подобна капле масла. В то же время и не вся поверхность белка несет заряженные или полярные атомы: примерно половина всех ее остатков неполярны.

Представьте, что вы — глобулярный белок в водном растворе, и вы хотите достичь стабильного нативного состояния. Ваша главная проблема — огромная потеря конформационной свободы, связанная с тем, что при переходе в уникальную конформацию множество денатурированных состояний перестает быть вам доступным. Это влечет за собой значительное понижение энтропии, что термодинамически невыгодно. Один из способов компенсации такой потери энергии — сформировать компактную глобулу, большинство остатков которой будет спрятано внутрь, изолировано от контактов с водой. Освобождение молекул воды от взаимодействий с неполярными атомами белка обеспечивает компенсирующее повышение энтропии как результат *гидрофобного эффекта* (см. 254).

Это замечательно, но теперь оказывается, что для образования компактной структуры вы должны скрыть вовнутрь множество полярных атомов, включая атомы азота и кислорода основной цепи (и не только их). В денатурированном состоянии эти атомы образуют водородные связи с водой. При погружении в глобулу их возможность образовывать водородные связи должна быть как-то реализована. (Не забывайте: одна или две некомпенсированные водородные связи — и равновесие нарушено, т. е. нативное состояние потеряет стабильность)[1)]. Достаточно универсальное решение, удовлетворяющее стремление атомов главной цепи к формированию водородных связей, состоит в образовании спиралей или листов.

К тому же формирование спиралей и листов служит свидетельством того, что остов находится в стереохимически приемлемой конформации, в соответствии с картой Сасисекхаран—Рамакришнан—Рамачандрана. Остатки в α-спиралях все находятся в α-конформации; остатки в тяжах β-складок — в β-конформации.

Как решить, какие участки цепи должны формировать спирали или тяжи? По энтальпийным характеристикам α-спирали и β-листы достаточно похожи для большинства остатков. Однако по энтропии некоторые боковые цепи в α-спиралях более заторможены, чем в β-тяжах; эти остатки предпочитают β-листы. Эти эффекты определяют формирование вторичных структур. Специфические последовательности, обеспечивающие образование водородных связей между боковыми цепями и остовом, формируют так называемые «спиральные обрамления», показывающие, где начинаются и заканчиваются α-спирали.

Насколько компактной должна быть белковая глобула? Вы можете достичь исключения всех молекул воды из глобулы при довольно неплотной упаковке белка — так как ни один канал не будет более чем 1.4 Å в радиусе (это размер молекулы воды). Но чем ближе вы сдвинете атомы, тем больше выиграете в вандерваальсовых (vdW) взаимодействиях — основных силах, обеспечивающих притяжение между атомами и придающих материи свойство когезии.

В целом белок упакован достаточно плотно, так что примыкающие друг к другу боковые цепи напоминают составную картинку-загадку (пазл). Однако отдельные кусочки этой картинки (аминокислотные остатки) могут де-

[1)] Это слишком сильное утверждение (*см. прим. ред.* 2) на с. 253). — *Прим. ред.*

формироваться, поэтому процесс сворачивания белков сложнее, чем строгое соответствие деталей в пазле.

Таким образом, вы должны найти такое расположение белковой цепи, которое одновременно удовлетворяло бы следующим требованиям.

1. Все остатки должны находиться в стереохимически допустимой конформации. Это относится как к основной цепи, так и к боковым радикалам. Стерические затруднения повысят энергию структуры и сделают ее нестабильной.
2. Погруженные внутрь глобулы полярные атомы должны образовывать водородные связи с другими полярными атомами. Если случайно не учесть несколько внутренних взаимодействий, модельная структура белка будет стремиться к развернутому состоянию как наиболее выгодному, чтобы дать возможность таким внутренним полярным атомам образовать водородные связи с молекулами растворителя.
3. Для обеспечения термодинамической стабильности достаточная доля гидрофобной поверхности должна быть погружена внутрь глобулы, а соответствующие остатки — хорошо упакованы.

Для большинства белков эти проблемы находят свои уникальные решения, обеспечивая стабильную нативную конформацию. Некоторые белки в процессе своего функционирования могут менять конформацию глобулы при связывании с лигандами или при переходах между своими метастабильными состояниями.

Механизм, обеспечивающий высокую стабильность единственной нативной конформации, сложен, но объясним. По сути, это вопрос оптимизации доступных взаимодействий и отбора тех последовательностей, на которых реализуемый оптимум уникален и существенно ниже по энергии, чем остальные. Для большинства областей локальная структура определяется локальными взаимодействиями. Таким образом, если бы нативное состояние не было уникальным, существовало бы несколько путей сборки заданного набора составляющих элементов. Учитывая ограничения цепи, в ходе эволюции было бы легко избежать подобной ситуации.

Сворачивание (фолдинг) белков

Снова представьте себе, что вы белок, но на этот раз — денатурированный. Теперь, зная, как стабилизировать нативную структуру, как вы будете искать ее? Совершенно очевидно, что перебор всех существующих конформаций вашей глобулы займет очень много времени — много лет назад Левинталь (C. Levinthal) рассчитал, что перебор конформаций (с разумной скоростью вращения связей) займет много, слишком много времени, чтобы когда-то завершиться. Есть два обстоятельства, которые подогревают интерес к изучению сворачивания белков.

Первое обстоятельство следует из того, что белки на самом деле не очень стабильны. Это означает, что любое квазистабильное промежуточное состояние (интермедиат), которое возникает в процессе сворачивания, должно быть еще менее стабильно, чем нативное. В противном случае сворачивание оста-

навливалось бы на стадии соответствующего интермедиата. На самом деле для многих белков определение доли нативных и денатурированных молекул в зависимости от температуры или концентрации денатурата указывает на наличие равновесия лишь между этими двумя конформационными формами. При этом промежуточные состояния присутствуют в недетектируемых количествах. Это подтверждает, что стабильность интермедиатов минимальна, однако затрудняет структурное исследование промежуточных стадий сворачивания белков.

Второе обстоятельство заключается в том, что денатурированное состояние слишком разрозненно по своей структуре. Учитывая, что промежуточные формы характеризуются очень низкой стабильностью, не удается найти приемлемый путь визуализации всего процесса фолдинга.

Сравните фолдинг белка с другими принципами формирования структуры.

1. При сборке мебели строитель последовательно переходит от одного пункта инструкции к другому, от одной промежуточной структуры к другой. Он прикручивает фрагмент А к фрагменту Б. Структура фрагмента А-Б стабилизируется связью между А и Б. Если бы не гравитация, интермедиат А–Б был бы стабилен. Но белкам недоступна такая роскошь — образовывать стабильные интермедиаты.
2. Нельзя построить арку только лишь из клинообразных камней — такая конструкция непрочна; чтобы придать ей стабильность используют краеугольный камень. Недостроенная арка очень неустойчива; для того чтобы в процессе строительства она не развалилась, используют строительные леса. Но белки не могут использовать внешние факторы.

Белки вынуждены работать с нестабильными интермедиатами — не полностью собранной мебелью, которая падает на землю под действием силы тяготения — и делать это так быстро, чтобы они не успели развалиться.

Можно попробовать определить структуру интермедиата методом изотопного обмена. Для этого необходим препарат денатурированного белка, в котором все атомы водорода заменены дейтерием (сигналы этих изотопов различаются при ЯМР-анализе). В ходе нескольких независимых экспериментов по рефолдингу в определенный момент времени (разный для каждого из опытов) образец подвергается обработке протонами. После того как нативная конформация образовалась, определяется, где и в какой момент произошел изотопный обмен D ↔ H. Эти исследования подтверждают механизм сворачивания через образование «расплавленной глобулы», содержащей правильные участки вторичной структуры, которые, однако, не стабилизируются взаимодействиями, присущими третичной структуре. В дальнейшем происходит последовательное, иерархическое сворачивание «расплавленной глобулы» в супервторичную структуру и т. д., постепенно приближаясь к конечной, нативной конформации. Не существует доказательств, что для большинства белков ненативная структура может играть роль интермедиата, хотя ненативные структуры, такие как некорректно свернутые изомеры пролина, могут иногда образовываться, тем самым замедляя процесс сворачивания.

Вывод: локальная структура участков полипептидной последовательности определяется в основном взаимодействиями в пределах этих участков. Несмотря на то что эти взаимодействия могут быть недостаточными для того, чтобы выделить эти локальные участки, они все же пригодны для реализации низкоэнергетического маршрута структурной сборки.

Применения гидрофобности

Используя шкалу гидрофобности, в которой каждой аминокислоте ставится в соответствие некоторое число, можно составить карту изменения гидрофобности вдоль аминокислотной последовательности белка. Ее называют *профилем гидрофобности*. Анализ профилей гидрофобности использовался для предсказания позиций поворотов между элементами вторичной структуры, положений экспонированных на поверхности и погруженных в глобулу остатков, мембранно-протяженных сегментов, сайтов антигенности.

ПРИМЕР 5.1.

Использование профилей гидрофобности для предсказания позиций переходов между α-спиралями и тяжами β-листов.

На рис. 5.4, *а* показан профиль гидрофобности лизоцима из белка куриного яйца. Отчетливые минимумы на графике соответствуют остаткам с номерами 17, 44, 70, 93 и 117. На рис. 5.4, *б* представлена структура лизоцима белка куриного яйца, в которой можно проследить корреляцию поворотов (между α-спиралями и тяжами β-листов) с позициями аминокислотных остатков, которым соответствуют минимумы на профиле гидрофобности.

Четыре основных минимума на профиле гидрофобности лежат в окрестностях позиций поворотов. Еще один минимум соответствует участку, экспонированному на поверхности, но в структуре это скорее один из тяжей β-листа, а не поворот. Один из минимумов находится внутри α-спирали. В свою очередь, ряд переходов не соответствует минимумам на графике гидрофобности. Таким образом, из профилей гидрофобности можно почерпнуть полезную информацию, но по ним нельзя однозначно предсказать все повороты в структуре белка.

ПРИМЕР 5.2.

Спиральное кольцо. О. Б. Птицын обнаружил, что у α-спиралей в глобулярных белках часто есть гидрофобное «лицо», повернутое внутрь, т. е. к внутренней части белка, и гидрофильное «лицо», повернутое во внешнюю среду, к растворителю. Каждый остаток в α-спирали повернут относительно предыдущего на 100°. Из эффекта Птицына следует вывод,

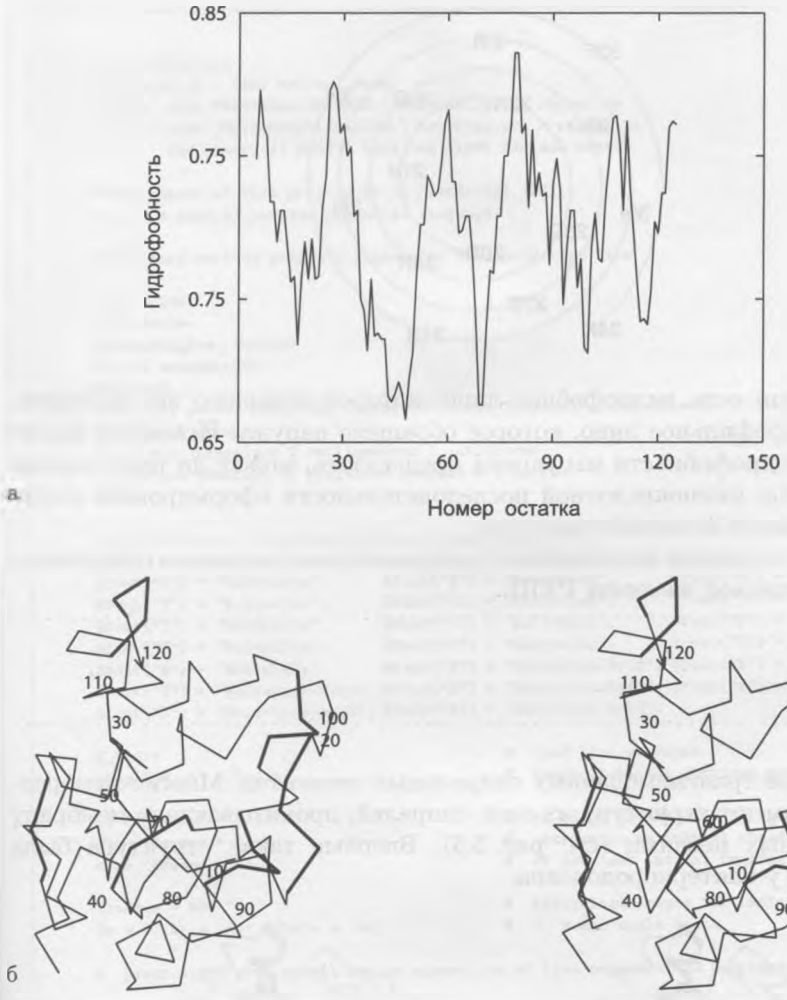
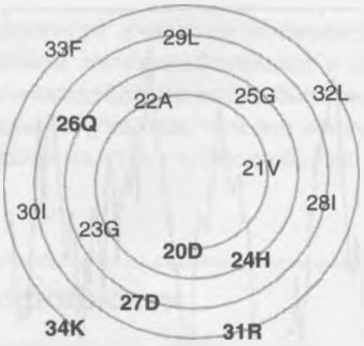


Рис. 5.4. (а) Профиль гидрофобности лизоцима из белка куриного яйца (получен при помощи утилит анализа первичной структуры, доступных на сайте `http://www.expasy.ch`.). (б) Структура лизоцима белка куриного яйца. Участки, соответствующие минимумам на графике гидрофобности, выделены жирными линиями

что в белковой последовательности гидрофобные и гидрофильные остатки должны сменяться с периодичностью, примерно равной четырем.

Для проверки этого утверждения спроецируем остатки на плоскость, перпендикулярную оси спирали, и получим диаграмму, называемую *спиральным кольцом*. В этом примере использована аминокислотная последовательность α-спирали миоглобина сперматозоида кита. Заряженные и полярные остатки обозначены полужирным шрифтом; остальные — обычным шрифтом.



У спирали есть гидрофобное лицо, которое обращено внутрь структуры, и гидрофильное лицо, которое обращено наружу. Исходя из такого принципа гидрофобности мы можем предсказать, может ли рассматриваемый участок аминокислотной последовательности сформировать α-спираль в нативной белковой структуре.

На с. 261 приведен исходный код программы для рисования спиральных колец, написанной на языке PERL.

ПРИМЕР 5.3.

Обнаружение трансмембранных спиральных сегментов. Многие мембранные белки имеют структуру из семи спиралей, пронизывающих мембрану и соединенных петлями (см. рис. 5.5). Впервые такая структура была обнаружена у бактериородопсина.

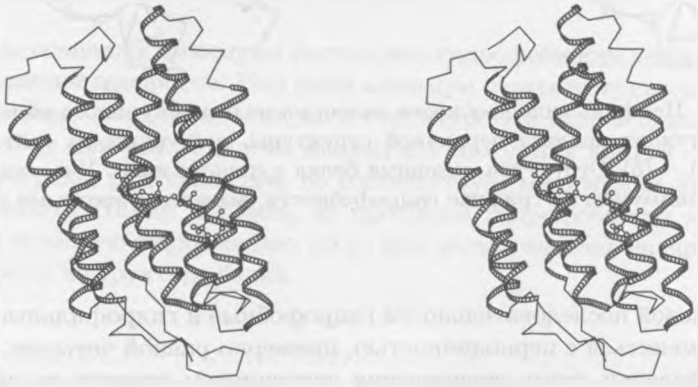


Рис. 5.5. Обнаруженный в мембране бактерии *Halobacterium salinarium* (ранее *Halobacterium halobium* |2BRD) бактериородопсин. Лиганд, показанный в виде шариков/палочек (ball-and-stick), — хромофор ретиналь

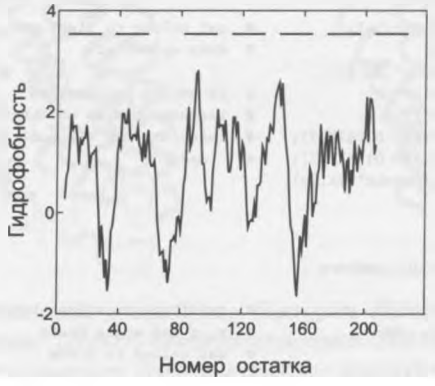
```
#!/usr/bin/perl
#helwheel.pl - draw helical wheel
#usage: echo DVAGHGQDILIRLFKSH | helwheel.prl > output.ps
# or    echo 20DVAGHGQDILIRLFKSH | helwheel.prl > output.ps
#       the numerical prefix sets the first residue number

# The output of this program is in PostScript (TM),
#      a general-purpose graphical language

# The next section prints a header for the PostScript file

print <<EOF;
%!PS-Adobe-
%%BoundingBox: (atend)
%1 0 0 setrgbcolor
%newpath
%37.5 161 moveto 557.5 161 lineto 557.5 681 lineto 37.5 681 lineto
%closepath stroke
297.5 421. translate 2 setlinewidth 1 setlinecap
/Helvetica findfont 20 scalefont setfont 0 0 moveto
EOF

#  Define fonts to associate with each amino acid

$font{"G"} = "Helvetica";      $font{"A"} = "Helvetica";      $font{"S"} = "Helvetica";
$font{"T"} = "Helvetica";      $font{"C"} = "Helvetica";      $font{"V"} = "Helvetica";
$font{"I"} = "Helvetica";      $font{"L"} = "Helvetica";      $font{"F"} = "Helvetica";
$font{"Y"} = "Helvetica";      $font{"P"} = "Helvetica";      $font{"M"} = "Helvetica";
$font{"W"} = "Helvetica";      $font{"H"} = "Helvetica-Bold"; $font{"N"} = "Helvetica-Bold";
$font{"Q"} = "Helvetica-Bold"; $font{"D"} = "Helvetica-Bold"; $font{"E"} = "Helvetica-Bold";
$font{"K"} = "Helvetica-Bold"; $font{"R"} = "Helvetica-Bold";

$_= <>;                                  #  read line of input
chop();$_ =~ s/\s//g;                    #  remove terminal carriage return and blanks

if ($_ =~ s/^(\d+)//)                    #  if input begins with integer
     {$resno = $1;}                      #  extract it as initial residue number
else {$resno = 1}                        #  if not, set initial residue number = 1

$radius = 50;                            #  initialize values for radius,
$x = 0; $y = -50; $theta = -90;          #  x, y and angle theta

#  print light gray spiral arc as succession of line segments, 10 per residue

$npoints = 10*(length($_) - 1);

print "0.8 0.8 0.8 setrgbcolor\n";       #  set colour to light gray
print "newpath\n";                       #  draw spiral arc
printf("%8.3f %8.3f moveto\n",$x,$y);
foreach $d (1 .. $npoints) {             #  10 points per residue
    $theta += 10; $radius += 0.6;        #  increase radius and theta
    $x = $radius*cos($theta*0.01747737); #  calculate new value of x
    $y = $radius*sin($theta*0.01747737); #    and y
    printf("%8.3f %8.3f lineto\n",$x,$y);
}
print "stroke\n";

#  print residues and residue numbers

$radius = 50;                            #  reinitialize values for radius,
$x = 0; $y = -50; $theta = -90;          #  x, y and angle theta
print '0 setgray\n'                      #  set colour to black

foreach (split ("",$_)) {                #  loop over characters from input line
```

```
        print "/$font{$_} findfont";          #  set font appropriate
        print "20 scalefont setfont\n";       #  for this amino acid
        printf("%8.3f %8.3f moveto\n",$x,$y); #  move to current point
        print " ($resno$_) stringwidth";      #  adjust position to center residue
        print " pop -0.5 mul -7 rmoveto\n";   #     identification on point on spiral
        print " ($resno$_) show\n";           #  print residue number and id
        print "% $theta $resno$_\n";
        $theta += 100; $radius += 6;          #  set new values of angle, radius
        $x = $radius*cos($theta*0.01747737);  #  compute new values of x
        $y = $radius*sin($theta*0.01747737);  #     and y
        $resno++;                             #  increase residue number
    }

    print "showpage\n";                       #  postscript signals to
    print "%%BoundingBox:";                   #  print
    $xl = 297.5 - 1.05*$radius;               #   x
    $xr = 297.5 + 1.05*$radius;               #     and
    $yb = 421.  - 1.05*$radius;               #      y
    $yt = 421.  + 1.05*$radius;               #       limits
    printf("%8.3f %8.3f %8.3f %8.3f\n",$xl,$xr,$yb,$yt);

    print "showpage\n";
    print "%%EOF\n";                          #  and wind up
```

ПРИМЕР 5.3 *Продолжение*

Трансмембранные сегменты состоят почти исключительно из гидрофобных аминокислотных остатков, так как вся спираль погружена в неводную среду. Такие сегменты отделены друг от друга участками, содержащими полярные аминокислоты. Обычная длина трансмембранных спиралей равна 15–30 остаткам.

На карте гидрофобности аминокислотной последовательности бактериородопсина *H. salinarum* видны семь областей максимумов, соответствующие семи трансмембранным спиралям (их позиции обозначены горизонтальными линиями).



WEB-РЕСУРСЫ: ПРЕДСКАЗАНИЕ ТРАНСМЕМБРАННЫХ СПИРАЛЕЙ

TMHMM (A. Krogh and E. Sonnammer) — предсказание на основе скрытой марковской модели (HMM):

`http://www.cbs.dtu.dk/krogh/TMHMM/`

PHDhtm (B. Rost):

`http://dodo.bioc.columbia.edu/predictprotein`

Membrane protein explorer (S. White):

`http://blanco.biomol.uci.edu/mpex/`

Ряд программ, доступных в сети, позволяют предсказывать участки трансмембранных спиралей в аминокислотной последовательности.

Совмещение структур и структурные выравнивания

Некоторые аспекты структурного анализа сводятся непосредственно к анализу последовательностей, некоторые задачи требуют обобщения, некоторые — вообще не имеют аналогов.

Как и в случае последовательностей, фундаментальным вопросом структурного анализа является выбор меры сходства и разработка метода ее вычисления. Если две молекулы имеют идентичные или очень похожие структуры, легко представить себе их наложенными друг на друга так, что соответствующие точки находятся на максимально близком расстоянии. Тогда мерой сходства структур является среднее расстояние между соответствующими точками. На практике обычно указывают среднеквадратичное отклонение Δd соответствующих атомов:

$$\Delta d = \sqrt{\sum d_i^2 / n},$$

где d_i^2 — расстояние между i-й парой точек после оптимальной «подгонки» структур, а n — число точек.

Вышесказанное подразумевает, что мы заранее установили соответствие между совмещаемыми точками.

Если же о соответствии ничего неизвестно, то сначала его нужно установить, и только потом рассчитать среднеквадратичное отклонение выровненных подструктур. Если каждая точка соответствует атому в следующих один за другим остатках структуры белка или нуклеиновой кислоты ($C\alpha$-атомы белков или атомы фосфора в нуклеиновых кислотах), задача буквально сводится к задаче выравнивания (т. е. это задание соответствий

остаток—остаток) (см. с. 265). В самом деле, установление соответствий остатков при структурных выравниваниях двух или более белков является мощным методом выравнивания последовательностей этих белков. Поскольку пространственная структура, как правило, консервативнее аминокислотной последовательности, структурное выравнивание по сравнению с обычным парным выравниванием последовательностей — более эффективная стратегия определения гомологии и выравнивания последовательностей белков с низкой степенью родства.

ПРИМЕР 5.4.

Структурное выравнивание γ-химотрипсина и эпидермолитического токсина А из *Staphylococcus aureus*.

Химотрипсин и эпидермолитический токсин А из *S. aureus* относятся к протеиназам семейства химотрипсина. На рис. 5.6 изображено наложение моделей γ-химотрипсина (PDB код 8GCH)(жирными линиями) и эпидермолитического токсина А из *S. aureus* (PDB код 1AGJ)(пунктиром). Обе молекулы имеют общий фолд, характерный для всех сериновых протеиназ семейства химотрипсина и каталитическую триаду Ser-His-Asp (самые толстые линии).

Выравнивание последовательностей белков, полученное из совмещения их структур:

```
8gch CGVPAIQPVLIVNG---------------EEAVP-GS--WPWQVSLQ-DKTG
1agj -------EVSAEEIKKHEEKWNKYYGVNAFNLPKELFSKVDEKDR-QKYPYNTIGNVFVK-G-

8gch FH-FCGGSLINE-NWVVTAAHC-GV-T--T-SDVVVAGEFDQG--SSSEKI-QKLKIAKVFK-NS-
1agj -QTSATGVLIG-KNTVLTNRHIAK-FANGDPSKVSFRPSI-NTDDNGNT-E-TPYGEYEVKEILQEP-F

8gch KYNSLTINNDITLLKLST---AAS-FSQTVSAVCLPSASD-DFAAGTTCVTTGWG-LTRYNTPD-R
1agj GAG---VDLALIRLKPDQNGVSL-GDK--ISPAKIGT--SNDLKDGDKLELIGYPFDH--KVNQ

9gch LQQASLPLL-SNTNCKKYWGTKIKDAM-ICAGASGV-SSCMGDSGGPLVCKKNGAWTLVGIVSWGSSTC
1agj MHRSEIELTTLS--------RGLRYY--GFTVPGNSGSGIFNSN--GELVGIHSSK--

8gch STST-----PGVYARVTA-LVNWVQQTLAAN-
1agj --VSHLDREHQINYGVGIGNYVKRIINEKN--E
```

Сходство этих двух последовательностей весьма слабое, оно не обнаруживается стандартным методом парного выравнивания одних лишь последовательностей.

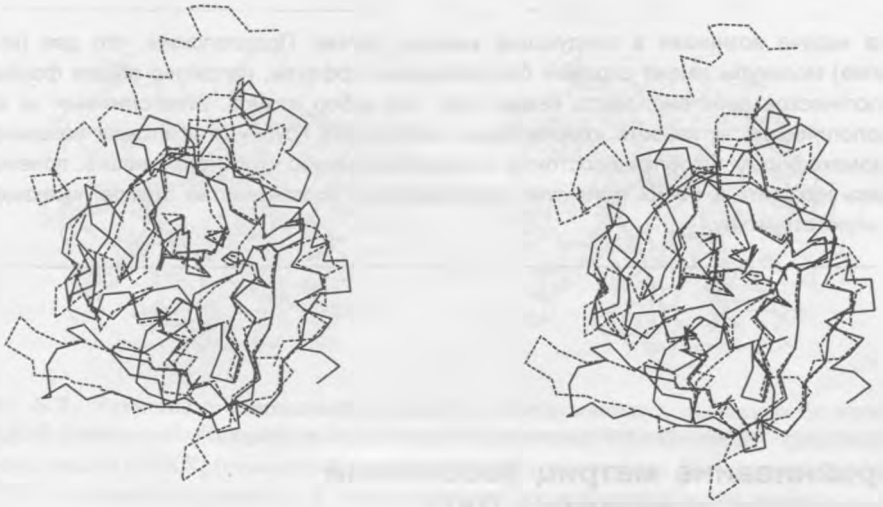


Рис. 5.6. Совмещение структур γ-химотрипсина [8GCH] (жирные линии) и эпидермолитического токсина А из *S. aureus* (пунктир). Показаны боковые цепи остатков каталитических триад. Обратите внимание на то, что область активного центра является наиболее консервативной областью всего белка

Определение сходства и выравнивания в вычислительной химии

1. Мера сходства двух наборов атомов с известным соответствием:

$$p_i \longleftrightarrow q_i, \quad i = 1, \ldots, N.$$

   В случае последовательностей аналог — расстояние Хэмминга только как мера несовпадения.

2. Мера сходства двух наборов атомов с неизвестным соответствием, но молекулярная структура которых, а именно линейный порядок расположения остатков, ограничивает число возможных соответствий. В случае белков или нуклеиновых кислот мы ограничены теми соответствиями, при которых сохраняется порядок вдоль цепи:

$$p_{i(k)} \longleftrightarrow q_{j(k)}, \quad k = 1, \ldots, K \leqslant N, M$$

   при условии, что $k_1 > k_2 \Rightarrow i(k_1) > i(k_2) и j(k_1) > j(k_2)$. Это принято называть расстоянием Левенштейна, или выравниванием со вставками. Результатом такого расчета будет выравнивание частей последовательностей.

3. Мера сходства двух наборов атомов с неизвестным соответствием (без ограничений на соответствия):

$$p_{i(k)} \longleftrightarrow q_{j(k)}$$

Эта задача возникает в следующем важном случае. Предположим, что две (или более) молекулы имеют сходные биологические эффекты, например общее фармакологическое действие. Часто бывает так, что набор атомов, ответственных за их биологическую активность, относительно небольшой. Группу этих атомов называют *фармакофором*. Проблема состоит в их идентификации: чтобы это сделать, полезно уметь находить в обеих молекулах максимальные подмножества атомов, имеющих схожую структуру.

Выравнивание матриц расстояний с помощью программы DALI

С эволюцией белков меняется и их структура. К деталям, ведущим себя наиболее консервативно в ходе эволюции, относятся паттерны контактов между остатками. Таким образом, если два остатка контактируют в одном белке, то, скорее всего, соответствующие им остатки родственного белка также контактируют между собой. Это верно для отдаленных гомологов, а также даже если изменяется размер вовлеченных в эволюцию остатков. Мутации, изменяющие размер остатков, упакованных внутри структуры, приводят к подходящим изменениям в расположении спиралей и листов относительно друг друга.

Л. Холм и К. Сандер применили эти наблюдения к решению проблемы структурного выравнивания белков. Если в отдаленно родственных белках паттерн контактирующих остатков консервативен, то становится возможным *идентифицировать* эти белки с помощью детектирования консервативных паттернов контактов.

В вычислительном плане нужно сделать матрицы паттернов контактов двух белков (это очень просто), а дальше найти в них максимально совпадающие блоки (это уже сложнее). Используя тщательно подобранные аппроксимации, Холм и Сандер написали эффективную программу, названную DALI (Distance-matrix-ALIgnment), которую часто используют для идентификации белков со сходными паттернами упаковки по отношению к заданной структуре. Это достаточно быстрая программа выполнения обычных скринингов всей базы Protein Data Bank для поиска структур, схожих с новой структурой, и даже для классификации структур доменов белков путем сравнения всех последовательностей относительно друг друга. Холм и Сандер нашли несколько неожиданных сходств между белками, которые не детектировались ранее при помощи парного выравнивания последовательностей.

Примером того, что DALI может определить даже очень отдаленно родственные белки, может служить идентификация гомологии между аденозиндеаминазой мыши, уреазой *Klebsiella aerogenes* и фосфотриэстеразой *Pseudomonas diminuta* (см. рис. 5.7).

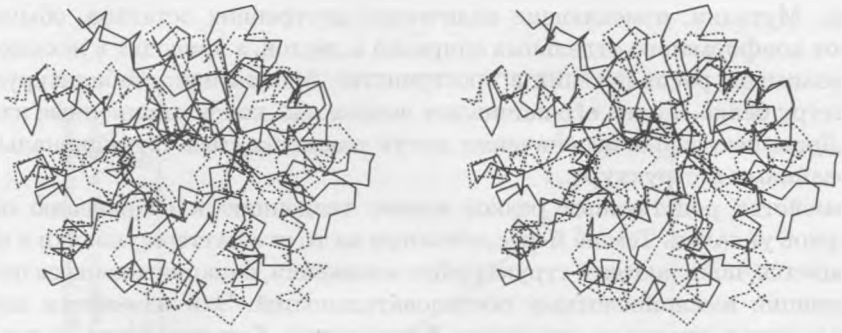


Рис. 5.7. Регионы с одинаковой укладкой, определенные с помощью программы DALI Л. Холма и С. Сандера, в белках с укладкой типа TIM-бочонка — аденозиндеаминазы мыши [1FKX] (сплошные линии) и фосфотриэстеразы *Pseudomonas diminuta* [1PTA] (прерывистые линии). В выравнивании последовательности имеется только 13% идентичных остатков — ближе к неидентифицируемой зоне гомологий

DALI доступен в Интернете. Вы можете загрузить интересующие вас структуры по адресу `http://www2.ebi.ac.uk/dali/`, и получить набор схожих структур и их выравнивание относительно запрашиваемой структуры.

Эволюция белковых структур

15 000[1)] известных сейчас белковых структур образуют несколько семейств, объединяемых базовым паттерном фолда, при вариации в последовательностях от почти одинаковых до имеющих 20% идентичности. Примерами могут служить семейства сериновых протеаз (γ-химотрипсин и эпидермолитический токсин А из *S. aureus*; рис. 5.6) и аденозиндеаминазы—фосфотриэстеразы (рис. 5.7).

Следствием мутаций обычно является изменение в структуре. И это характерное свойство биологических систем — принимать определенную форму вследствие эволюции родственных, но не идентичных структур. В таком случае структура должна быть достаточно устойчивой по отношению к возможным мутациям. И мы можем использовать это свойство устойчивости: идентифицируя и сравнивая родственные объекты, можно разделять вариабельные и консервативные участки и таким образом идентифицировать остатки, отвечающие за структуру и функцию.

Вариации, встречающиеся в семействах гомологичных белков с одинаковой функцией, показывают, как структура приспосабливается к изменениям в аминокислотной последовательности. Остатки на поверхности белка, не влияющие на функцию, обычно свободно мутируют. Внешние петли обычно могут адаптироваться к изменениям путем локальных структурных пере-

[1)] На 11 сентября 2007 г. — более 42 000 белковых структур, согласно данным `www.rcb.org/pdb`. — *Прим. ред.*

строек. Мутации, изменяющие количество внутренних остатков, обычно не меняют конформацию отдельных спиралей и листов, а приводят к искажениям в их взаимном расположении в пространстве. Механизмы, стабилизирующие структуру белка, также ограничивают возможные конформационные изменения. Дополнительные ограничения могут накладываться функциональными требованиями к структуре.

Семейства родственных белков имеют тенденцию к сохранению общих паттернов укладки. Тем не менее, несмотря на то что паттерн укладки в целом сохраняется, появляются и структурные искажения, накапливающиеся по мере дивергенции аминокислотных последовательностей. Эти изменения неоднородно распределяются по структуре. Как правило, большое центральное ядро структуры сохраняет качественно ту же укладку, в то время как другие участки структуры изменяют конформацию более радикально. Представьте буквы B и R. Как и структуры, они имеют одно ядро, соответствующее букве P. Вне этого одинакового ядра они различаются: снизу справа у буквы B петля, а у R — диагональная линия.

Систематическое изучение структурных различий между парами родственных белков выявило количественную связь между расхождением аминокислотной последовательности ядра структуры семейства и расхождением структуры. С расхождением последовательностей увеличивается и расхождение в конформации основной цепи, а доля остатков в ядре обычно уменьшается. Пока доля идентичных остатков в последовательности не станет ниже 40–50%, эти эффекты относительно несущественны. Почти вся структура напоминает консервативное ядро, и деформации в атомах основной цепи в среднем не превышают 1 Å. С дальнейшим расхождением в последовательности некоторые участки полностью перестраиваются, уменьшая размер ядра, а структурные деформации внутри ядра увеличиваются в амплитуде.

Корреляция между расхождением последовательности и структуры применима ко всем семействам белков. На рис. 5.8, *a* показаны изменения в ядре, представленные в виде среднеквадратичного отклонения атомов основной цепи после оптимального наложения, в зависимости от расхождения последовательности (процент консервативных аминокислот в ядре при оптимальном выравнивании). Точки соответствуют парам гомологичных белков из нескольких семейств. (Точки, соответствующие структурам со 100% идентичностью, — те, для которых структура была разрешена в двух или более кристаллографических экспериментах; и разброс показывает, что силы в упакованном кристалле — а для некоторых также растворитель и температура — все же незначительно влияют на структуру белка.) На рис. 5.8 показано изменение в количестве остатков, образующих ядро, как функция от расхождения белковых последовательностей. Количество остатков в ядрах отдаленно родственных белков может сильно варьировать: в некоторых случаях количество остатков в ядре остается высоким, в других — падает ниже 50%.

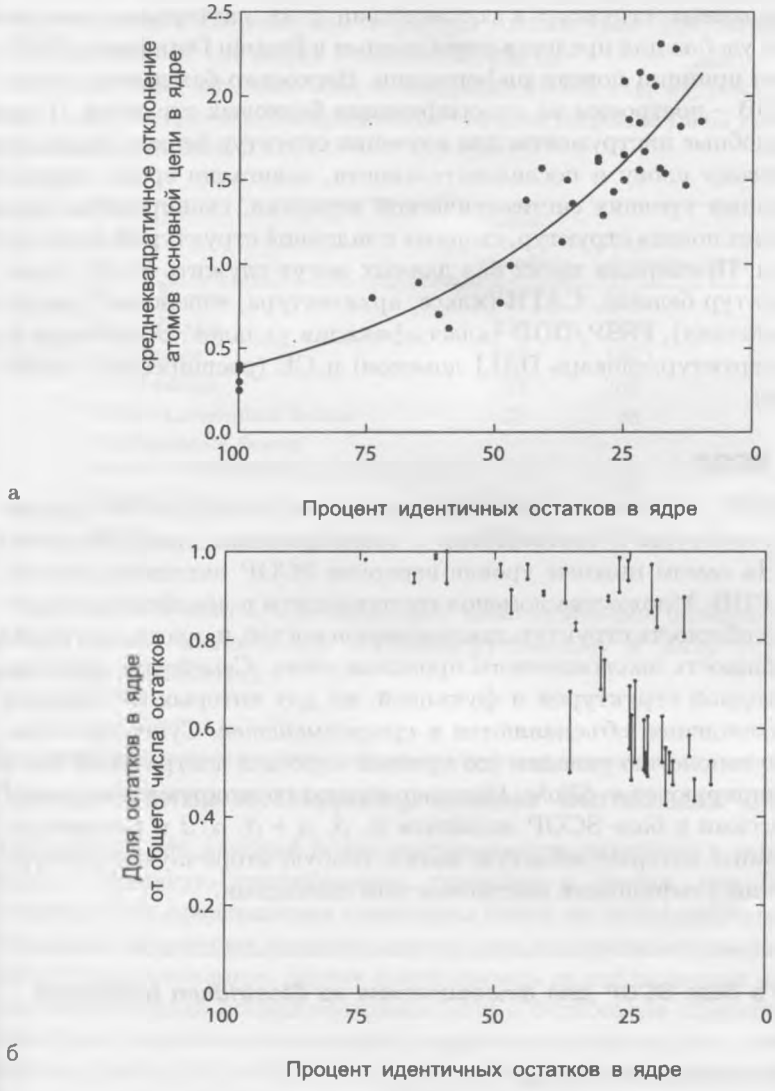


Рис. 5.8. Связи между расхождением в аминокислотной последовательности и трехмерной структурой ядра в эволюционно связанных белках. *а* — зависимость среднеквадратичного отклонения от процента идентичных атомов в ядре *б* — зависимость размера ядра от процента идентичных остатков в ядре. Графики отражают результаты обсчета 32 пар гомологичных белков разных структурных типов. (Адаптировано из Clothia, C, и Lesk, A. M. (1986) «Связь между расхождением последовательностей и структурой белков» The EMBO Journal 5, 823–6.)

Классификация структур белков

Организация белковых структур в соответствии с их паттернами укладки логически очень удобна для представления данных в Protein Data Bank (PDB). На этом основан принцип поиска информации. Несколько баз данных — производных от PDB — построены на классификации белковых структур. В них предлагаются удобные инструменты для изучения структур белков, такие как поиск по ключевому слову и последовательности, навигация среди сходных структур на разных уровнях систематической иерархии, сканирование базы данных на предмет поиска структур, сходных с заданной структурой, и ссылки на другие сайты. Примерами таких баз данных могут служить SCOP (классификация структур белков), CATH (класс, архитектура, топология, гомология в суперсемействах), FSSP/DDD (классификация укладок, основанная на выравнивании структур/словарь DALI доменов) и CE (расширенный комбинаторный метод).

База данных SCOP

База данных SCOP (structural classification of proteins) иерархически организует белковые структуры в соответствии с эволюционным происхождением и структурой. На самом нижнем уровне иерархии SCOP находятся *домены*, выделенные из PDB. Множества доменов группируются в *семейства* гомологов, для которых общность структур, последовательностей, и, иногда, функций указывает на общность эволюционного происхождения. Семейства, содержащие белки со сходной структурой и функцией, но для которых не очевидна общность происхождения, объединяются в *суперсемейства*. Суперсемейства, имеющие общую топологию укладки (по крайней мере для центральной части структуры), группируются в *фолды*. Наконец, фолды группируются в *классы*. Основными классами в базе SCOP являются α, β, $\alpha+\beta$, α/β и разнообразные «малые белки», которые зачастую имеют слабую вторичную структуру и объединяются дисульфидными мостиками или лигандами.

Классификация в базе SCOP для флаводоксина из *Clostridium beijerinckii*

1. *Корень:* SCOP
2. *Класс:* Alpha and beta proteins (α/β)
   В основном параллельные β-листы ($\beta-\alpha-\beta$ единицы)
3. *Фолд:* Flavodoxin-like
   3 слоя, $\alpha/\beta/\alpha$; параллельные β-листы из 5 тяжей, порядок 21345
4. *Суперсемейство:* Flavoproteins
5. *Семейство:* Flavodoxin-related binds FMN
6. *Белок:* Flavodoxin
7. *Организм: Clostridium beijerinckii*

из: `http://scop.mrc-lmb.cam.ac.uk/scop`

Здесь приведены классификация в базе SCOP флаводоксина из *Clostridium beijerinckii* (цветная иллюстрация II). Иллюстрации степени сходства белков на разных уровнях иерархии и обсуждение других схем классификации приведены в книге [Introduction to protein Architecture: The Structural Biology of Proteins (Oxford University Press, 2001), гл. 4].

Релиз SCOP от июля 2001 г. содержит 13 220 структур из PDB, разделенных на 31 474 домена. Распределение записей по различным уровням иерархии приведено в таблице.

Класс	Число		
	семейств	суперсемейств	фолдов
Все α белки	337	224	138
Все β белки	276	171	93
α/β белки	374	167	97
$\alpha + \beta$ белки	391	263	184
Мультидоменные белки	35	28	28
Мембранные белки и белки клеточной поверхности	28	17	11
Малые белки	116	77	54
Всего	1557	947	605

Другие Web-сайты, предлагающие классификацию белковых структур см.: `http://www.bioscience.org/urllists/protdb.htm` и `http://www2.ebi.ac.uk/msd/Links/fold.shtml`.)

Предсказание и моделирование белковых структур

Наблюдение, что каждый белок сворачивается спонтанно в уникальную трехмерную нативную конформацию, приводит к мысли, что Природа имеет алгоритм для предсказания структуры белка из аминокислотной последовательности. Некоторые попытки понять этот алгоритм основывались на общих физических принципах; другие основывались на наблюдениях над известными аминокислотными последовательностями и белковыми структурами. Доказательством полноты и правильности нашего понимания была бы возможность воспроизвести этот алгоритм в виде компьютерной программы, которая бы могла предсказывать структуру белка по ее аминокислотной последовательности.

Большинство попыток предсказания структуры белка на основе только физических принципов пробуют воспроизвести межатомные взаимодействия в белках, чтобы вычислить энергию каждой конформации. С вычислительной точки зрения проблема предсказания структуры белка сводится к поиску глобального минимума конформационной энергии. Это подход до сих пор не привел к успеху отчасти потому, что методы минимизации находят локальные минимумы.

Другие попытки предсказания белковых структур были основаны на упрощениях задачи путем выделения существенных особенностей.

Альтернативой к априорным методам являются подходы, основанные на компоновке структуры исследуемой последовательности с помощью поиска сходства с известными структурами. Этот эмпирический основанный на знании подход оказывается весьма продуктивным.

Мы уже почти достигли такого момента в своих знаниях, когда известны многие, но еще не все возможные способы укладки белков с известной структурой. Это было объявлено как цель проекта структурной геномики (см. с. 273). Когда будут перечислены все возможные способы укладки (фолды) и последовательности и будут созданы мощные методы для установления соответствия между ними, тогда эмпирические методы приведут к практическому решению многих задач. Какой эффект мы ожидаем от возможности предсказывать структуру белков *a priori*? Интеллектуальный вызов все еще останется — научиться предсказывать структуру белка без поиска в базах данных. Но, к сожалению, это уже не будет иметь такого большого интереса, поскольку практическая задача уже будет решена.

Однако здесь есть парадокс: методы, разработанные для идентификации способа укладки по последовательности не более, чем упражнение по подгонке параметров в функции веса. Они являются экспериментальными и показывают важные свойства аминокислотных последовательностей, которые определяют структуру белков. Когда эти методы приведут к успеху, мы получим более прочную основу к пониманию взаимоотношений последовательность—структура, чем у нас есть сейчас. Может так случиться, что такое апостериорное знание поможет решению проблемы априорного предсказания белковых структур.

Методы предсказания структуры белков по аминокислотной последовательности включают в себя:

- Попытку предсказания вторичной структуры белка без укладки ее в пространственную структуру. В результате получается список сегментов, для которых предсказано, что они формируют α-спирали или тяжи β-листов.
- Моделирование по гомологии: предсказание трехмерной структуры белка на основе известной структуры одного или нескольких гомологичных белков. В результате получается полный список всех координат всех атомов как главной цепи, так и боковых радикалов. При этом точность предсказания сравнима с экспериментальной структурой, полученной методами малого разрешения.
- Распознавание способа укладки: в данной библиотеке известных структур определить, какие из них могут быть наиболее похожими на структуру нового белка. Если белок не соответствует ни одному из фолдов библиотеки, то метод также должен это распознать. Результатом является отнесение нашего белка к одному из известных фолдов или утверждение, что такого фолда в библиотеке нет.
- Предсказание новых фолдов, в том числе и с помощью априорных методов, основанных на знании. В результате получается полный набор координат

Структурная геномика

По аналогии с проектами по секвенированию полных геномов структурная геномика ставит перед собой цель сделать доступными структуры всех белков. Методами РСА и ЯМР будет разрешен достаточно «плотный набор» белковых структур, такой чтобы структуры всех остальных белков могли быть получены с помощью моделирования по гомологии на основании одной или большего числа достаточно близких и экспериментально доступных структур. Проекты структурной геномики опираются на данные о гораздо большем количестве организмов в сравнении с проектами по секвенированию геномов. В этом плане белки системы человеческого комплемента, как и белки специфические для определенных болезнетворных микроорганизмов, безусловно, представляют особый интерес.

Задачи структурной геномики стали реалистичными отчасти благодаря развитию экспериментальных методов, открывшему доступ к высокопроизводительным работам по расширению структур; и отчасти благодаря продвижению в нашем понимании белковой структуры, что позволило ставить адекватные цели перед экспериментаторами и фокусироваться на определенных молекулярных мишенях.

Теория и практика моделирования по гомологии показывают, что между последовательностями исследуемого белка и экспериментально разрешенной структурой должно быть не менее 30% идентичности. Это означает, что необходимо экспериментально определять структуры для представителей всех семейства последовательностей, включая многие структуры, обладающие схожим фолдом, т. е. число определенных экспериментально доменов должно быть почти к 10 000. В 2000 г. в PDB было помещено 2297 структуры[1)]. Таким образом, производительность науки соответствует требованиям.

Методы биоинформатики помогают в выборе мишени для экспериментального определения структуры.

Цели выбора мишени включают:

- отсев вырожденных мишеней — белков, слишком схожих с уже известной структурой.
- поиск последовательностей с недетектируемым сходством с белками с уже известной структурой.
- идентификация последовательностей, схожих только с белками, чьи функции неизвестны, либо:
- с неизвестными структурами белков, обладающими «интересными» функциями; например, человеческие белки, связанные с заболеваниями, или бактериальные белки, участвующие в формировании устойчивости к антибиотикам.
- белки со свойствами, предпочтительными для расшифровки их структур, — предпочтительно растворимые, с повышенным содержанием метионина (который ускоряет решение фазовой проблемы РСА), и т. д.

«Машина» моделирования уже функционирует. MODBASE собирает гомологичные модели белков при помощи комплексного пакета PSI-BLAST, а MODELLER (A. Sali и коллеги) строит модели по гомологии.

Проекты по структурной геномике поддерживаются инициативами Национального института здоровья (NIH) США и различными частными предприятиями.

атомов как минимум для основной цепи и, иногда, для боковых цепей. Модель стремится предсказать способ укладки, но при этом не ожидается, что ее предсказание количественно сравнимо с экспериментальными результатами. Д. Джонс (D. Jones) сравнил различие между априорным моделированием и распознаванием фолда с различием между сочинением и тестированием с выбором ответов из заданного списка на экзамене.

Критическая оценка предсказаний структуры (CASP)

Программы CASP (critical assessment of structure prediction) были кратко рассмотрены в гл. 1. CASP организует «слепые» тесты по предсказанию белковых структур, в которых кристаллографы и ЯМР-спектроскописты публикуют аминокислотные последовательности белков, структуры которых они расшифровали, но сами структуры держат в тайне до тех пор, пока «предсказатели» не будут готовы представить модели этих белков. Каждые два года последовательности публикуются весной, а к осени предсказания должны быть уже готовы. В конце года авторы предсказаний собираются на торжественной конференции для обсуждения текущих результатов и оценки успехов.

Предсказания в CASP делятся на три основные категории: (1) сравнительное моделирование (в сущности, моделирование по гомологии), (2) распознавание укладки (фолда) и (3) моделирование новых фолдов:

Категория CASP	Свойства мишени
Сравнительное моделирование	Доступны близкие гомологи с известной структурой; применимы методы гомологичного моделирования
Распознавание фолда	Структуры со схожими фолдами доступны, но недостаточно близки для моделирования по гомологии; задача состоит в том, чтобы идентифицировать структуры со схожей топологией
Новый фолд	Структуры с таким фолдом неизвестны; требуется a priori достоверный или основанный на знании метод, который сможет учитывать особенности нескольких структур

Три эксперта-оценщика, по одному на каждую категорию, сравнивают предсказанную и экспериментально расшифрованную структуры и оценивают качество предсказаний. В число докладчиков входят организаторы, эксперты-оценщики и предсказатели, включая тех, кто только отчасти добился успеха и тех, кто разработал интересный новый метод моделирования.

Последняя программа CASP проходила в 2000 г.[1]. Было представлено 43 мишени. 163 группы предсказателей во всех категориях выдвинули в общей сложности 11 136 моделей. Это примерно равно числу структур в PDB на то время.

Предсказание вторичной структуры

Кажется очевидным, что (1) вторичную структуру легче предсказать, чем третичную, и (2) что наиболее точный способ предсказания третичной структуры состоит в нахождении спиралей и листов, с последующим объединением их в фолд (укладку). Независимо от того, верны эти предположения или нет, многие доверяют и следуют им. По аминокислотной последовательности белка с неизвестной структурой делаются предсказания вторичной структуры — отнесение участков последовательности к спиралям или тяжам листов.

На программе CASP в 2000 г. сервер PROF (B. Rost) достиг хороших результатов в предсказании структуры домена белка репарации MutS из *Thermus aquaticus*. Для оценки качества предсказания аминокислотные остатки экспериментально расшифрованной трехмерной структуры были отнесены к трем категориям (спираль = H, тяж = E и другие = –). Процент остатков, предсказанных правильно, был обозначен как Q3. Для предсказания Б. Роста аминокислотная последовательность Q3 составил 81%:

```
                                    10        20        30        40        50
                                     |         |         |         |         |
Аминокислотная последовательность  ALVEDPPLKVSEGGLIREGYDPDLDALRAAHREGVAYFLELEERERERTG
Предсказание                        HH------EEE---HHHHHHHHHH-HHHHHHHHHHHHHHH-
Эксперимент                        -E-------E---HHHHHHHHHHHHHHHHHHHHHHHHHHHH-

                                    60        70        80        90       100
                                     |         |         |         |         |
Аминокислотная последовательность  IPTLKVGYNAVFGYYLEVTRPYYERVPKEYRPVQTLKDRQRYTLPEMKEK
Предсказание                       -EEEEEEEEEEEEEEE------EEEEEEEE-EEEE-HHHHHH
Эксперимент                        --EEEEE--EEEEEEEHHHHHH---EEEEE--EEEEE-HHHHHH

                                   110       120
                                    |         |
Аминокислотная последовательность  EREVYRLEALIRRREEEVFLEVRERAKRQ
Предсказание                       HHHHHHHHHHHHHHHHHHHHHHHHHHHH-
Эксперимент                        HHHHHHHHHHHHHHHHHHHHHHHHHHHH-
```

На рис. 5.9 показана экспериментальная структура, на которой отмечены предсказанные программой элементы вторичной структурой. За исключением короткой 3_{10} спирали и незначительных расхождений в позициях начала и конца, элементы вторичной структуры были предсказаны правильно. (Другие схемы оценки, которые осуществляют проверку участков совпадения,

[1] Последняя встреча CASP была в 2004 г. — *Прим. ред.*

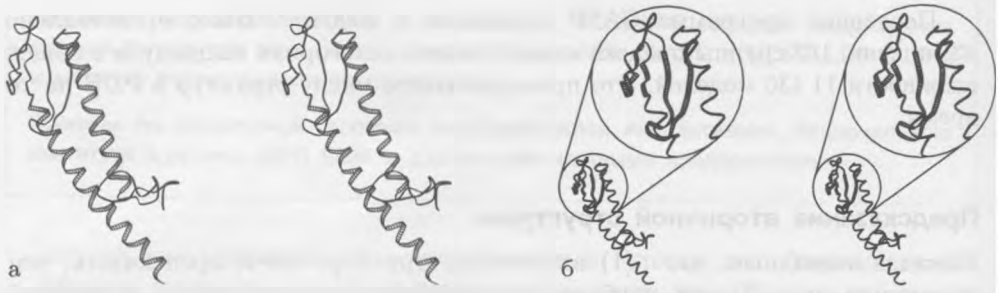


Рис. 5.9. Структура белка репарации MutS из *Thermus aquaticus* [1EWQ]. (а) Участки, предсказанные сервером Роста PROF как спиральные, изображены более широкими лентами. В предсказанной структуре недостает только короткой 3_{10} спирали (слева вверху). (б) Участки, опознанные программой как листы, обозначены более широкими лентами

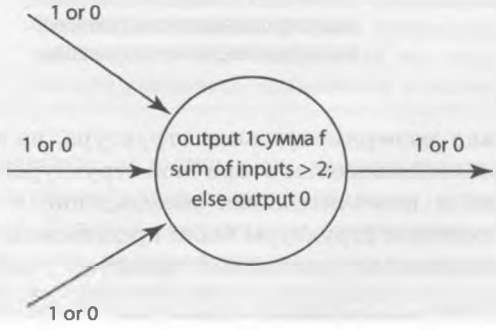
менее чувствительны к концевым эффектам.) Качество этого результата очень высоко, но такой результат не редкость. Структура моделированного Ростом белка была классифицирована экспертами CASP как объект средней сложности. В настоящее время PROF работает в среднем с точностью Q3≈77%. Другие методы предсказания вторичной структуры работают также сравнительно хорошо.

Наиболее мощные методы предсказания вторичной структуры основаны на нейронных сетях.

Нейронные сети

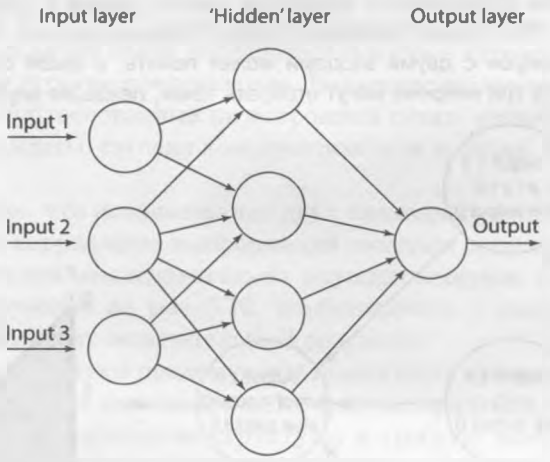
Нейронные сети — это класс общих вычислительных структур, которые моделируют анатомию и физиологию биологических нервных систем. Они с успехом применяются к широкому спектру задач распознавания образов, классификации и задачам принятия решений.

В вычислительной схеме одиночный нейрон является вершиной графа с одним или несколькими входящими ребрами (входами) и одним исходящим ребром (выходом):



Используя физиологическую метафору, можно сказать, что нейрон испускает сигнал, если на выходе у него 1, и не испускает сигнал, если на выходе 0. Модельные нейроны могут различаться по количеству входов и выходов и по формуле, которая вычисляет выход (см. с. 277).

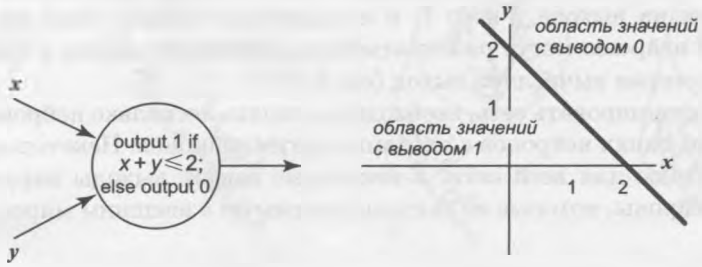
Чтобы сформировать сеть, необходимо создать несколько нейронов и соединить выходы одних нейронов с входами других нейронов. Некоторые вершины содержат входы для всей сети, а некоторые имеют выходы наружу. Кроме того есть нейроны, которые не связаны напрямую с внешним миром (скрытые слои).



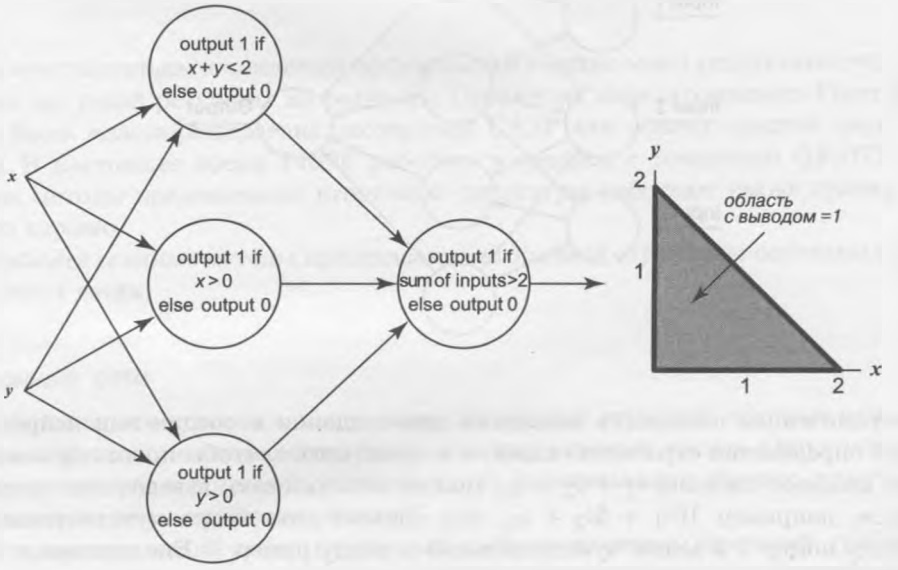
Неограниченная сложность возможна при создании и соединении нейронов и при определении строгости связей, т. е. вместо того, чтобы просто суммировать входные сигналы $i_1 + i_2 + i_3$, можно использовать взвешенные суммы входов, например $10i_1 + 5i_2 + i_3$, что сделает сеть более чувствительной к входу номер 1 и менее чувствительной к входу номер 3. Биологически это соответствует изменению строгости синапсов.

Логика нейронных сетей

Для одиночного нейрона линейная функция отклика может иметь геометрическую интерпретацию в терминах линий и плоскостей. Нейрон на рисунке имеет два входа. Если мы будем интерпретировать вход (x, y) как точку на плоскости, нейрон принимает решение, на какой стороне от линии находится вход. Выход будет 1, если и только если $x + y \leqslant 2$ (если точка лежит под линией $x + y = 2$).



Нейронная сеть определяется топологией связей, весами и формулой принятия решения в узлах. Сеть может принимать более сложные решения, чем один нейрон. Так, если один нейрон с двумя входами может понять, с какой стороны от линии находится точка, то три нейрона могут отобрать точки, лежащие внутри треугольника.



Нейронная сеть будет более мощной и ясной, если выход является гладкой функцией от входов. Для тренировки нейронных сетей полезно, чтобы выход был дифференцируемой функцией от параметров. Поэтому резкую функцию отсечения заменяют на гладкую сигмовидную.

ступенчатая функция сигмоидная функция

Свойство нейронной сети, которое определяет ее мощь, заключается в том, что веса могут рассматриваться как переменные и могут вычисляться в процессе обучения для частных случаев. Чтобы обучить нейронную сеть, ее применяют к разным примерам и сравнивают выход с правильным решением. Если ответ не совпадает, то проводят уточнение параметров. В процессе обучения топология сети остается неизменной, при этом, если вес какой-либо связи оказывается равным 0, то это эквивалентно разрыву соответствующей связи.

Тип нейронной сети, которая может быть применена к распознаванию вторичной структуры, показан на рис. 5.10.

Важной информацией, которая может быть использована при предсказании вторичной структуры, является эволюционная информация. Множественное выравнивание содержит в себе гораздо больше информации, чем одна последовательность. Сохранение вторичной структуры в родственных белках означает наличие связи последовательность — структура, и это позволяет делать более строгие предсказания. Большинство методов предсказания вторичных структур, основанных на нейронных сетях, имеют на входном слое не только информацию о степени консервативности позиции, но и профильные веса.

Показано также, что использование двух тандемных нейронных сетей позволяет учитывать корреляцию конформаций соседних остатков. Предсказания состояний нескольких последовательно идущих остатков с помощью сети, аналогичной показанной на рис. 5.10, комбинируется с помощью еще одной сети, которая формирует окончательный результат.

Тест на зрелость методов предсказания может быть проведен полностью автоматически. Некоторые вычислительные методы продуцируют только предварительное грубое предсказание структуры и требуют вмешательства человека для формирования окончательного результата. Другие методы полностью автоматические. Есть множество Web-ресурсов, которые принимают последовательность и возвращают предсказание. PROF — одна из таких систем, использовавшихся для предсказания вторичной структуры MutS.

Непрерывный полностью автоматический анализ Web-серверов предсказаний структур (включая предсказание вторичных структур, но не только их) называется EVA. Это плод сотрудничества групп из Нью-Йорка и Мадрида. PDB представляет последовательности до опубликования структур, и программы реализованные в EVA посылают их на серверы предсказаний и анализируют результаты. Это можно рассматривать как непрерывный CASP, ограниченный методами, которые могут быть проверены автоматически. См. `http://cubic.bioc.columbia.edu/eva`

Целью проекта EVA является мониторинг прогресса в этой области и выработка рекомендаций пользователям для использования различных серверов предсказания структуры в разных категориях. EVA имеет доступ к гораздо большему набору данных, чем CASP. Поэтому ее решения менее подвержены статистическим флуктуациям и трудностям выбора задач в CASP.

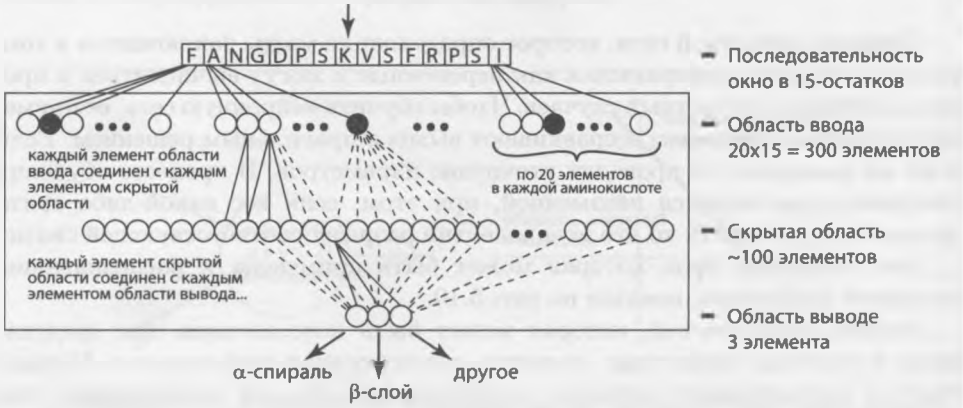


Рис. 5.10. Нейронная сеть для предсказания вторичной структуры белка из трех областей

- Входная область сканирует последовательность окном в 15 остатков, т. е. анализируется фрагмент последовательности размером 15 букв. Предсказание относится к центральному остатку (наверху, отмечен стрелкой). Затем окно сдвигается на одну позицию вправо по последовательности и делается следующее предсказание. Каждой из 15 позиций в окне соответствует 20 нейронов, один из которых активен.
- Скрытая область содержит ≈100 нейронов, соединенных с вводом и выводом. Каждый нейрон в скрытой области соединен с каждым нейроном областей ввода и вывода (показаны не все связи).
- Область вывода состоит только из трех нейронов, которые просто фиксируют предсказание — спираль, лист, или ни то ни другое

Моделирование по гомологии

Построение модели по гомологии — полезный метод для предсказания структуры белка по известной последовательности, когда исследуемый белок состоит в родстве хотя бы с одним белком, для которого известны и последовательность, и структура. Если белки являются близкими родственниками, то известные белковые структуры (называемые родительскими) могут служить основой для модели исследуемого белка. И хотя качество модели зависит от степени сходства последовательностей, оценить это качество возможно до экспериментальной проверки (см. рис. 5.8). Поэтому знание того, какое качество модели требуется приложением, для которого она предназначена, позволяет с высокой долей вероятности предсказать успешность выполнения задачи.

Шаги моделирования по гомологии:

1. Выровнять аминокислотные последовательности исследуемого белка и белка (белков) с известной структурой. В большинстве случаев вставки и делеции будут наблюдаться в петлях между α-спиралями и β-тяжами.
2. Определить сегменты основной цепи, содержащие вставки или делеции. Сшивание этих участков с основной цепью известного белка (матрицы) создает модель основной цепи исследуемого белка[1)].

[1)] Под сшивкой понимается: а) удаление участков основной цепи, присутствующих в матрице и отсутствующих в исследуемом белке; б) вставка участков, которые есть в исследуемом белке, но отсутствуют в матрице. — *Прим. ред.*

3. Заменить боковые цепи мутировавших остатков. Для немутировавших остатков сохранить конформацию боковых цепей. Мутировавшие остатки склонны сохранять конформацию боковой цепи, и это можно использовать при моделировании. При этом сейчас также доступны вычислительные методы поиска подходящей конформации боковой цепи среди возможных комбинаций.
4. Проверить модель (и визуально, и с помощью программ), чтобы выявить значительные наложения атомов. Насколько возможно, устранить подобные наложения вручную.
5. Уточнить модель методом ограниченной минимизации энергии. Роль этого шага — установить точное геометрическое расположение в тех местах, где были соединены участки главной цепи, и позволить боковым цепям слегка перемещаться, чтобы занять удобное положение. На самом деле эффект этой процедуры только косметический: минимизация энергии не устранит серьезных ошибок в такой модели.

По сути, данная процедура выдает то, что «вы получаете за бесплатно» в том смысле, что модель нового белка строится путем внесения минимальных изменений в доступную структуру матрицы. К сожалению, существенно улучшить такую модель непросто. Эмпирическое правило (см. рис. 5.8) гласит, что если две последовательности идентичны хотя бы на 40–50%, описанная процедура дает модель, достаточно точную для использования во многих приложениях. Если же сходство ниже, то ни описанная процедура, ни какой-либо другой доступный на данный момент алгоритм, не дадут детально точной модели, исходя из доступных структур родственных белков.

Структуры большинства белковых семейств содержат как относительно постоянные, так и более вариабельные участки. Ядро структуры семейства сохраняет топологию укладки, хотя и может быть искажено, периферия же может быть целиком сложена заново. Располагая единственной прародительской структурой, можно относительно достоверно моделировать консервативную часть исследуемого белка, но построить модель вариабельной части уже не удастся. Более того, непросто предсказать, какие участки являются вариабельными, а какие — консервативными. Более предпочтительна ситуация, когда несколько родственных белков с известной структурой могут выступать в качестве родителей для моделирования по гомологии. Они выявляют внутри семейства участки с консервативной и вариабельной структурой. Наблюдаемое распределение структурной вариабельности среди родительских структур диктует подходящее распределение ограничений применительно к моделированию.

Развитое программное обеспечение для моделирования по гомологии доступно. SWISS-MODEL — это Web-сайт, который принимает на входе аминокислотную последовательность исследуемого белка, определяет, существует ли подходящая для моделирования по гомологии родительская структура (структуры), и, если существует, выдает на выходе набор координат исследуемого белка. SWISS-MODEL разработали T. Schwede, M. C. Peitsch и N. Guex в Женевском институте биомедицинских исследований.

WEB-РЕСУРСЫ: МОДЕЛИРОВАНИЕ ПО ГОМОЛОГИИ

SWISS-MODEL (автоматическое моделирование по гомологии):
`http://www.expasy.ch/swissmod/SWISS-MODEL.html`

MODBASE, база данных сравнения моделей белков из полных геномов:
`http://guitar.rockefeller.edu/modbase/`

Описание Web-сайтов по структурной геномике см. Wixon, J. (2001) 'Structural genomics on the web', *Comp. Funct. Genomics* 2, 103–13.

Примером автоматического предсказания с помощью SWISS-MODEL является предсказание структуры нейротоксина красного скорпиона (*Buthus tamulus*) на основе известной структуры нейротоксина родственного североамериканского желтого скорпиона (*Androctonus australis hector*). Эти два белка имеют 52% идентичных остатков в выравнивании последовательностей. При таком высоком сходстве неудивительно, что модель очень близка к экспериментальным результатам даже в том, что касается конформаций боковых цепей (рис. 5.11).

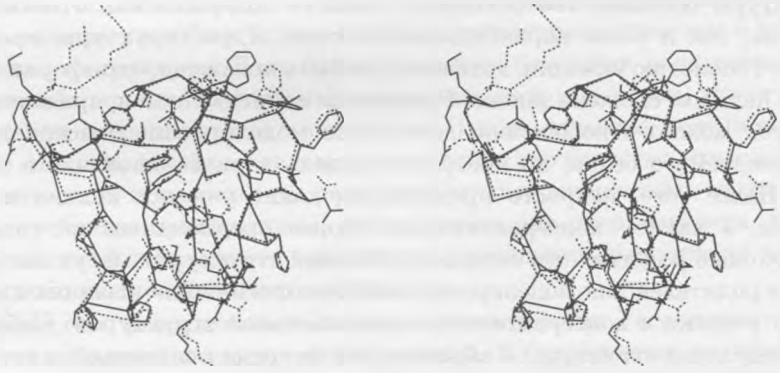


Рис. 5.11. SWISS-MODEL предсказывает структуру нейротоксина красного скорпиона [1DQ7] (сплошные линии), исходя из структуры близкородственного белка [1PTX]. Предсказание (пунктирные линии) сделано автоматически. Обратите внимание на то, что большинство спрятанных внутрь боковых цепей не мутировали и имеют очень близкие конформации. Некоторые боковые цепи на поверхности имеют иные конформации и С-конец главной цепи расположен по другому (слева вверху). Не показана сеть дисульфидных мостиков, скрепляющих структуру. Тем не менее, для двух настолько близких белков столь высокое качество модели должно было быть ожидаемо, даже без таких дополнительных скреплений

Распознавание фолда

Поиск последовательности в базе данных последовательностей и поиск структуры в базе данных структур — это задачи, имеющие решения. Смешанные задачи (поиск по структуре в банке последовательностей или по последовательности в банке структур) менее очевидны. Они требуют метода для оценки совместимости данной последовательности с данным способом (паттерном) укладки (фолда).

Цель состоит в выделении существенного набора последовательностей и структур. Ожидается, что белки, имеющие один и тот же паттерн, имеют схожие структуры.

3D-профили

Мы обсудили паттерны и профили, получаемые из множественных выравниваний последовательностей, и их применение для выявления далеких гомологов. Один из способов повысить мощность этих методов, используя доступную структурную информацию, — это такой тип профиля, который получается из доступных последовательностей *и* структур белкового семейства.

J. U. Bowie, R. Lüthy и D. Eisenberg проанализировали *окружение* каждой позиции в известных белковых структурах и соотнесли с набором предпочтений двадцати аминокислот в структурном контексте.

Имея белковую структуру, можно классифицировать окружение каждой аминокислоты по трем отдельным категориям:

1. водородные связи основной цепи, т. е. вторичная структура;
2. степень погруженности внутрь или экспонированности на поверхность белковой глобулы;
3. полярная/неполярная природа окружения.

Возможны три варианта вторичной структуры: *α-спираль* (helix), *β-слой* (sheet) и *иное*. Боковая цепь считается погруженной, если площадь доступной поверхности составляет менее 40 Å^2, частично погруженной — от 40 Å^2 до 114 Å^2 и экспонированной — более 114 Å^2. Доля площади боковой цепи, приходящейся полярные атомы, также измеряется. Авторы определяют 6 классов аминокислот на основе доступности и полярности окружения. Боковые цепи каждого из этих шести классов могут быть в любом из трех типов вторичной структуры. Таким образом, всего получается 18 классов. Если отнести каждую боковую цепь к одному из 18 классов, то можно, пользуясь алфавитом из 18 букв, создать описание белковой структуры, называемое профилем трехмерной структуры (*3D-профилем*). К «последовательностям», в которых таким образом закодированы структуры, можно применить алгоритмы, разработанные для поиска последовательностей. Например, можно попытаться выровнять две далекие друг от друга родственные последовательности путем выравнивания их 3D-профилей, а не самих аминокислотных последовательностей. Метод 3D-профилей превращает белковые структуры в одномерные объекты, не сохраняющие точно ни последовательность, ни структуру молекул, из которых они были получены.

Далее, как можно соотнести 3D-профиль с набором известных последовательностей и структур? Ясно, что некоторые аминокислоты «не рады» находиться в определенных местах; например, заряженная боковая цепь не может быть спрятана внутри совсем неполярного окружения. Остальные предпочтения не столь четки, поэтому необходимо составить таблицу предпочтений на основе статистического обзора библиотеки, содержащей белковые структуры высокого качества.

Предположим теперь, что у нас есть последовательность и мы хотим оценить вероятность того, что она принимает, скажем, фолд глобина. Из 3D-профиля известной структуры миоглобина кашалота мы знаем класс окружения у каждой позиции в последовательности. Рассмотрим частичное выравнивание неизвестной последовательности с миоглобином кашалота и предположим, что первому остатку миоглобина соответствует в неизвестной последовательности остаток фенилаланина. В 3D-профиле класс окружения первого остатка миоглобина следующий: экспонированная боковая цепь, нет вторичной структуры. Можно оценить вероятность нахождения фенилаланина в этом классе окружения, используя таблицу предпочтений отдельных аминокислот для этого класса 3D-профилей. (Тот факт, что первым остатком в миоглобине кашалота является валин, не использован; эта информация недоступна алгоритму. Миоглобин кашалота представлен только последовательностью классов окружения своих остатков, а таблица предпочтений усреднена для белков, имеющих различные способы укладки.) Распространение этих подсчетов на все позиции и на все возможные выравнивания (не допускающие разрывов внутри участков, имеющих вторичную структуру) дает число, которое оценивает, насколько хорошо данная неизвестная последовательность подходит профилю миоглобина кашалота.

Особое преимущество этого метода состоит в том, что он может быть автоматизирован. Новую последовательность можно сравнивать с каждым 3D-профилем в библиотеке известных фолдов по сути таким же способом, какой отработан для сравнения новой последовательности с библиотекой известных последовательностей.

Использование 3D-профилей для определения качества структур

3D-профиль, полученный из структуры, весьма опосредованно зависит от аминокислотной последовательности. Поэтому можно спросить, не только насколько возможно отнести другую аминокислотную последовательность к данному фолду, но также использовать вес 3D-профиля для родной последовательности в качестве меры совместимости стуктуры и последовательности. Естественно, если реальные последовательности не показывают высокого веса сопоставления с собственной структурой, то можно придти к выводу, что на других последовательностях тем более будет получен плохой результат. Есть два интересных наблюдения. (1) Белковые структуры, хорошо соответствующие собственным профилям на родственных белках, также дают высокий вес. Профиль является абстрактным свойством семейства, а не только индивидуально белка. (2) Когда родственная последовательность плохо со-

ответствует профилю, полученному из экспериментальной структуры этой последовательности, то по-видимому в структуре есть ошибка. Позиции, где профиль не соответствует последовательности могут указывать на область, где находится ошибка.

Трединг

Трединг (threading — протягивание) — это метод распознавания фолда. Если есть библиотека известных структур, и запрашиваемая последовательность, то соответствует ли последовательность одному из фолдов в библиотеке? Библиотека фолдов может содержать некоторые (или все) структуры из PDB и даже гипотетические фолды.

Основная идея трединга состоит в том, чтобы построить много грубых моделей для данной последовательности, используя всевозможные выравнивания с последовательностями, для которых известна структура. Систематическое исследование множества возможных выравниваний определило название метода. Представьте, что вы пытаетесь аккуратно протянуть вашу последовательность через известную трехмерную структуру. При этом допустимы вставки и делеции, но если протягивание достаточно мягкое, то метафора «протягивания» остается в силе.

Как трединг, так и моделирование по гомологии, имеют дело с трехмерными структурами, индуцированными выравниваниями искомой последовательности и последовательности, для которой трехмерная структура определена. Моделирование по гомологии концентрируется на множестве выравниваний и имеет целью построение детальной структуры. Трединг использует множество различных выравниваний и работает только с грубыми моделями, иногда даже не построенными явно:

Моделирование по гомологии	Трединг
Сначала найди гомологов	Проверь всех возможных партнеров
Построй оптимальное выравнивание	Проверь все возможные выравнивания
Оптимизируй одну модель	Оцени много грубых моделей

Для успешного распознавания с использованием трединга требуется:

1. Метод для оценки моделей, позволяющий выбрать одну.
2. Метод калибровки весов, чтобы можно было понять насколько выбранная модель хороша.

Было испытано несколько аппроксимаций взвешивания. Одна из наиболее эффективных основана на эмпирической оценке близости аминокислотных остатков, полученной из анализа известных структур. Наблюдения над межостаточными расстояниями в известных структурах для всех 20×20 пар типов остатков. Для каждой пары остатков было построено распределение вероятностей пространственных расстояний между остатками. Например, для пары Leu–Ile были рассмотрены все Leu и Ile во всех структурах и были вычислены пространственные расстояния между C_β-атомами и расстояния по

последовательности. Коллекция этих данных позволила построить оценку, насколько хорошо расстояния в модели соответствуют расстояниям в известных структурах.

Распределение Больцмана связывает энергию и вероятность. Обычно применение распределения Больцмана начинается с вычисления энергии и вероятность. (Стандартный пример — предсказание плотности атмосферы как функция высоты с использованием потенциальной энергии молекул.) При тредине наоборот — из вероятности выводится энергия. Эта энергетическая функция используется для оценки качества модели.

Для каждой структуры из библиотеки процедура находит соответствие остатков, доставляющее минимум энергии. Хотя это и является задачей выравнивания, нелокальность взаимодействий не позволяет применить здесь метод динамического программирования.

Распознавание фолда в CASP 2000

Наилучшие методы предсказания фолда единообразно эффективны. Они включают методы, основанные на тредине, но не ограничиваются им.

На рис. 5.12 и 5.13 показаны предсказания, сделанные А. Г. Мурзиным и группой Бонно, Цай, Ружински и Бэкер для белка из программы CASP 2000. Оба белка с неизвестной функцией из *H. Influenzae*.

а Целевая структура

б Предсказание А. Г. Мурзина

в Известная структура

Рис. 5.12. Предсказание структуры белка из *H. Influenzae*. (а) Целевая структура. (б) Предсказание А. Г. Мурзина. (в) Структура ближайшего гомолога с известной структурой: чувствительный к N-этилмалеимиду белок, вовлеченный в везикулярный транспорт (PDB 1NSF). Топология, предложенная А. Г. Мурзиным ближе к целевому белку, чем известная структура ближайшего гомолога

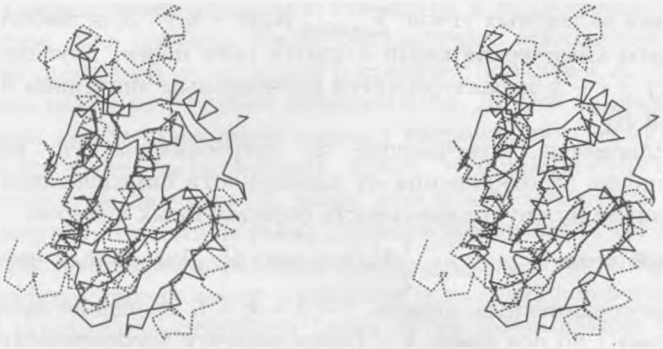


Рис. 5.13. Предсказание группы Бонно, Цай, Ружински и Бэкер для другого белка из *H. Influenzae*, основанное на глицин-N-метилтрансферазе [1XVA]. Экспериментальная структура показана сплошными линиями, предсказанная — пунктиром. Отметим, что значительная часть структуры хорошо соответствует экспериментальной структуре. Некоторые другие участки имеют похожую локальную структуру, но неправильную ориентацию и упаковку относительно основной части белка

Вычисление конформационной энергии и молекулярная динамика

Белок состоит из многих атомов. Взаимодействие атомов создает уникальное состояние максимальной стабильности. Надо только его найти!

Эта задача сложна для вычислений, из-за того что (а) модель межатомных взаимодействий недостаточно полна и недостаточно точна, (б) даже если удастся построить адекватную модель, мы столкнемся с проблемой оптимизации нелинейной целевой функции в очень большом пространстве переменных с нелинейными ограничениями, что порождает весьма сложную поверхность с множеством локальных минимумов. Эта задача весьма трудная, так же, как поле для гольфа с множеством лунок.

Взаимодействия между атомами могут быть разбиты на два вида:

(a) первичные химические связи — прочные взаимодействия, заставляющие атомы находиться на малом расстоянии друг от друга. Они рассматриваются как постоянные взаимодействия, которые не разрушаются при структурных перестройках белковых молекул, а сохраняются во всех конформациях.

(b) более слабые взаимодействия, зависящие от конформации цепи. Они могут быть значительными в одних конформациях и незначительными — в других. Они возникают, когда атомы при различных упаковках приближаются друг к другу.

Конформацию белка можно задать списком названий и координат его атомов, а также набором химических связей между ними (эту информацию достаточно достоверно можно получить из аминокислотной последовательности). При оценке конформационной энергии учитывают следующие величины:

- Растяжение связей: $\sum_{\text{bonds}} K_r(r - r_0)^2$, где r_0 — равновесное межатомное расстояние, а K_r — константа растяжения связи. r_0 и K_r зависят от типа химической связи.

- Деформация валентных углов: $\sum_{\text{angles}} K_\theta(\theta - \theta_0)^2$. Для любого i-го атома, образующего химические связи с двумя (или более) другими атомами j и k, угол $j - i - k$ характеризуется равновесным значением θ_0 и силовой константой K_θ.
- Прочие слагаемые, отвечающие за стереохимическую корректность, и штрафующие за отклонения от планарности определенных групп, или удерживающие нужную хиральность определенных центров.
- Торсионный угол: $\sum_{\text{dihedrals}} \frac{1}{2} V_n[1 + \cos n\phi]$. Для любых четырех последовательно соединенных атомов, $i - j - k - l$, вращение атома l относительно атома i по оси связи $k - l$ определяется энергетическим барьером с периодическим потенциалом. V_n — высота барьера внутреннего вращения; n — количество барьеров, встречающихся при повороте на 360°. Пример: торсионные углы ϕ, ψ и ω между атомами основной цепи полипептида (рис. 5.2).
- Вандерваальсовы (vdW) взаимодействия: $\sum_i \sum_{j<i} (A_{ij} R_{ij}^{-12} - B_{ij} R_{ij}^{-6})$. Для каждой пары несвязанных атомов i и j первое слагаемое характеризует силы близкодействующего отталкивания, а второе — силы дальнодействующего притяжения. Параметры A и B зависят от типа атомов. R_{ij} — расстояние между атомами i и j.
- Водородные связи: $\sum_i \sum_{j<i} (C_{ij} R_{ij}^{-12} - D_{ij} R_{ij}^{-10})$. Водородная связь — это слабое химическое/электростатическое взаимодействие между двумя полярными атомами. Энергия такого взаимодействия зависит как от расстояния между атомами, так и от угла. Приведенная формула энергии взаимодействия не отражает в явном виде зависимость от угла связи. Другие формулы могут включать этот «геометрический» параметр.
- Электростатические взаимодействия: $\sum_i \sum_{j<i} Q_i Q_j / (\varepsilon R_{ij})$. Q_i и Q_j — эффективные заряды на атомах i и j, R_{ij} — расстояние между ними, а ε — диэлектрическая константа. Эта формула является лишь приближением применительно к средам, которые, как и белки, не являются непрерывными и изотропными.
- Растворитель: взаимодействия с растворителем, водой и другими компонентами раствора (такими, как соли и сахара) оказывают большое влияние на термодинамику структуры белка. Рассмотрение растворителей как непрерывных сред, характеризуемых диэлектрической проницаемостью в качестве основного параметра, является лишь приближением. С развитием вычислительной техники стало возможным производить расчеты белка в ячейке с явно заданными молекулами растворителя (воды).

Существует большое количество потенциалов, описывающих конформационную энергию, и много усилий было приложено к уточнению параметров, используемых в расчетах. Энергия данной конформации рассчитывается суммированием взаимодействий различных типов по всем атомам системы.

Правильность функции потенциальной энергии — необходимое, но не достаточное условие удачного предсказания структуры белка. Один из способов проверить это утверждение — взять в качестве стартовой конформации экспе-

риментально определенную белковую структуру и попробовать минимизировать ее энергию. Как правило, среднеквадратичное отклонение такой минимизированной структуры от исходной составляет порядка 1 Å. Эту величину можно назвать мерой разрешения силового поля. Другой вариант состоял бы в минимизации конформационной энергии неправильно свернутой белковой глобулы. Это позволило бы определить, лежит ли минимум энергии правильно свернутой структуры значительно ниже локального минимума неправильно свернутой глобулы. Результаты таких расчетов показывают, что на основании одних лишь расчетов конформационной энергии нельзя достоверно отличить нативную конформацию белка от множества остальных конформеров.

Попытки предсказать структуру белка путем минимизации конформационной энергии до сих пор не привели к разработке метода, позволяющего делать предсказания, отталкиваясь от одной лишь аминокислотной последовательности. Для преодоления обеих проблем — попадания в ложный локальный минимум, а также построения хорошей модели взаимодействия с растворителем была разработана методология молекулярной динамики. Белок вместе с явно заданным растворителем рассчитываются — посредством задания силового поля — в рамках ньютоновской механики. Данный метод хорош тем, что и вправду позволяет исследовать большие участки фазового пространства. Хотя как метод для априорного предсказания структуры по аминокислотной последовательности, эта методология еще не созрела. В то же время молекулярная динамика, как, пожалуй, никакой другой метод в этой области, зависит от развития компьютерных технологий, и поэтому появление более мощных процессоров, возможно, изменит положение вещей в лучшую сторону.

В настоящее время молекулярная динамика может существенно облегчить экспериментальное разрешение структуры белка такими методами, как РСА (обычно помогает) и ЯМР (помогает всегда). Каким образом молекулярная динамика может быть интегрирована в процесс определения структуры? Для каждой из конформаций можно рассчитать отклонение модели от реальных экспериментальных данных. В случае кристаллографии экспериментальными данными являются абсолютные значения коэффициентов Фурье-образа электронной плотности молекулы. В случае ЯМР, используя экспериментальные данные, можно рассчитать расстояния между определенными аминокислотными остатками. Но в обоих случаях экспериментальные данные недоопределяют структуру белка. Для того чтобы полностью разрешить структуру, необходимо отыскать такой набор координат, который бы минимизировал как отклонение от экспериментальных данных, так и конформационную энергию структуры. С этой задачей успешно справляется молекулярная динамика, грамотно сканируя конформационное пространство и сходясь к корректной структуре путем минимизации отклонений от доступных экспериментальных данных.

Программа ROSETTA

Программа ROSETTA — продукт лаборатории D. Baker для предсказания структуры белка по аминокислотной последовательности; программа использует информацию об уже расшифрованных структурах. ROSETTA показала хорошие результаты на конкурсе CASP в 2000 г. в категории предсказания нового фолда. В настоящий момент данный программный продукт опережает свои аналоги на несколько корпусов.

ROSETTA, используя данные об уже имеющихся структурах, сначала предсказывает структуру отдельных фрагментов, объединяя их впоследствии в единую структуру. Вначале последовательность разбивается на фрагменты от 3 до 9 аминокислот и происходит поиск схожих фрагментов в белках с известной структурой. Поскольку фрагменты достаточно короткие, то никаких предположений о родственных связях между белками не делается. Исходя из возможных вариантов структуры отдельных фрагментов, рассчитываются возможные варианты структуры белка в целом.

ROSETTA использует для анализа таких комбинаций метод Монте-Карло (см. с. 290). Функция энергии включает в себя члены, характеризующие компактность, спаренность β-листов и погруженность гидрофобных остатков. Процедура производит 1000 независимых вычислений для структур, отобранных на основании заранее генерированного паттерна распределения конформаций фрагментов. Схожие структуры, полученные в результате независимых симуляций Монте-Карло, объединяются в кластеры; при этом центральные структуры наибольших кластеров принимаются за модели исследуемого белка. Таким образом, основная идея заключается в том, что структура, полученная наибольшее количество раз в ходе независимых испытаний Монте-Карло, и будет наиболее правдоподобной моделью.

Алгоритмы Монте-Карло

Методы Монте-Карло широко используются в расчетах белковых структур, для эффективного перебора конформаций, для поиска минимума сложной функции, а также во многих других оптимизационных задачах. Простые минимизационные процедуры, основанные на движении вниз по градиенту энергии, неэффективны, так как зачастую расчеты сходятся к локальному минимуму, далекому от нативной структуры.

Идея метода заключается в том, чтобы использовать случайные числа в решении вопроса, ответ на который трудно вычислить точно. Название метода было придумано J. von Neumann со ссылкой на самое известное казино, использовавшее генераторы случайных чисел.

Для того чтобы с помощью метода Монте-Карло найти минимум функции многих переменных — например, функции энергии белка, зависящей от переменных, определяющих его конформацию. Будем считать, что эта функция зависит от набора переменных x, и для любого набора значений этих переменных мы можем подсчитать

значение функции — энергию конформации белка $\mathcal{E}(x)$. (В качестве x может выступать набор координат атомов белка или торсионных углов основных и боковых цепей аминокислот.)

Метод Монте-Карло применяют в соответствии с процедурой Метрополиса (придумана за обедом в Лос-Аламос в 1953 г.):

1. Случайным образом генерируются начальные условия — набор переменных x. Подсчитывается энергия полученной конформации $\mathcal{E} = \mathcal{E}(x)$.
2. Создается возмущение переменной $x \to x'$, для того чтобы попасть в соседнее состояние.
3. Подсчитывается энергия нового состояния $\mathcal{E}(x')$.
4. Дается ответ на вопрос, принять новую, возмущенную конформацию, или вернуться к прежней, исходной:
   (а) если энергия уменьшилась, $\mathcal{E} = \mathcal{E}(x) > \mathcal{E}(x')$ необходимо всегда принимать новое состояние. Оно будет использоваться в качестве начальных условий в следующем цикле, $x' \to x$ и $\mathcal{E} = \mathcal{E}(x')$.
   (б) если энергия увеличилась или осталась неизменной, т. е. $\mathcal{E}(x) \leqslant \mathcal{E}(x')$ то можно отвергнуть новое состояние или принять его с вероятностью $\exp[-\Delta/(kT)]$, где $\Delta = \mathcal{E}(x') - \mathcal{E}(x)$, k — константа Больцмана, а T — эффективная температура.
5. Вернуться к пункту 2.

Вся изюминка в пункте 4б. Он позволяет выбираться из «ловушек» локальных минимумов. Вероятность того, что такое движение вверх по энергии будет принято, определяется эффективной температурой T. В данном случае T не физический, а лишь числовой параметр, управляющий расчетом. Для любого значения температуры, вероятность принять новую структуру с более высокой энергией тем меньше, чем больше эта разница в энергии. Для любого значения $\mathcal{E}$, если температура низкая, то значение $\mathcal{E}(x)/(kT)$ будет высоким, а значение $\exp[-\mathcal{E}(x)/(kT)]$ — низким. И наоборот, если температура высокая, то значение $\mathcal{E}(x)/(kT)$ будет низким, а экспонента $\exp[-\mathcal{E}(x)/(kT)]$ — близка к единице. С увеличением температуры увеличивается вероятность того, что возмущенная структура с большей энергией будет принята.

Эта достаточно простая идея оказалась крайне эффективной для многих приложений к расчетам белковых структур и в других областях.

Читатели могли также слышать о «симулируемом отжиге» (simulated annealing) — развитии метода Монте-Карло, в котором варьируется значение параметра T: вначале температура делается высокой для того, чтобы разрешить различные конформационные изменения, а потом — понижается, для того чтобы свести систему к минимуму энергии.

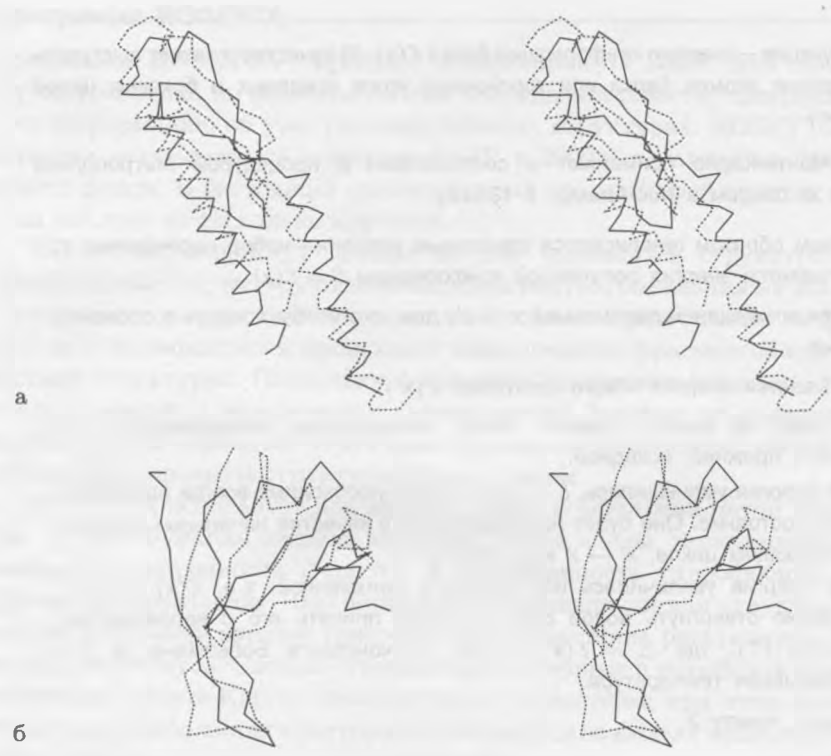


Рис. 5.14. Предсказание при помощи программы ROSETTA: (а) гипотетический белок H. influenzae, (б) N-концевая половина домена 1 белка человека Xrcc4 из системы репарации ДНК; показана выбранная подструктура, содержащая 55 N-концевых аминокислотных остатков из 116. Экспериментальные структуры показаны сплошными линиями, предсказанные — пунктиром

На рис. 5.14 показаны результаты успешного предсказания двух структур с помощью программы ROSETTA на конкурсе CASP в 2000 г.

Программа LINUS

LINUS (local independently nucleated units of structure) — программа предсказания структуры белка по аминокислотной последовательности, разработанная G. D. Rose и R. Srinivasan. Данная процедура полностью априорна, т. е. работает только с самой последовательностью, не опираясь ни на экспериментальные данные, ни на известные структурные корреляции. В LINUS реализован «иерархический» алгоритм — сворачивание начинается с коротких фрагментов, постепенно объединяя их в более длинные.

Основная идея LINUS заключается в том, что структура локальных участков белка — коротких аминокислотных фрагментов — определяется локальными взаимодействиями внутри этих участков. В процессе фолдинга, каждый сегмент предпочтительно будет принимать наиболее энергетически выгодные

конформации. Однако эти предпочтительные конформации и даже самые выгодные из них — те, что с неизбежностью реализуются в нативной структуре белка, — лишь ненамного превышают порог термодинамической стабильности. Локальная структура будет претерпевать множество переходов, до тех пор пока не отыщется подходящее стабилизирующее взаимодействие. Но компьютер волен предвосхищать появление результатов. Те из локальных структур, что появляются достаточно часто в ходе расчетов, могут передавать свои структурные свойства последующим генерациям и таким образом влиять на конечный результат. Процедура использует принцип известного в технике храпового колеса с тем, чтобы направить расчеты по наиболее эффективному пути.

LINUS начинает строить полипептидную цепь отталкиваясь от структуры тяжа. В ходе расчетов производится возмущение случайно отобранных троек последовательно расположенных остатков, и оценивается энергия результирующей конформации. Стерически затрудненные структуры отбраковываются; остальные энергетические вклады рассчитываются только для локальных взаимодействий. Метод Монте-Карло (см. с. 290) используется для того, чтобы сделать выбор: принять возмущенную структуру, или вернуться к ее предшественнику. В LINUS эти стадии многократно повторяются; в процессе работы создается статистика структурных предпочтений для всех аминокислот.

Далее локальные фрагменты собираются в более крупные, основываясь на рассчитанной статистике структурных предпочтений. При этом границы зоны, внутри которой рассчитываются взаимодействия, постоянно расширяются — от локальных фрагментов до целой глобулы.

Представление процесса сворачивания белка в программе LINUS реалистично по своей сути, хотя и приближенно. В моделировании участвуют все атомы белка, кроме атомов водорода, и функция энергии при этом рассчитывается приблизительно, а динамика упрощена. Функционал энергии учитывает: (1) стерическое отталкивание атомов, (2) кучность погруженных гидрофобных остатков, (3) водородные связи, и (4) солевые мостики. В настоящее время LINUS в большинстве случаев успешно предсказывает структуры маленьких фрагментов (размером между супервторичной структурой и доменом) и в некоторых случаях может интегрировать их в правильную общую структуру. На рис. 5.15 представлены результаты предсказаний, сделанных с помощью LINUS для структуры С-концевого домена белка ERp29 эндоплазматического ретикулума крысы, одной из задач конкурса CASP 2000 г.

Определение белковых структур в геномах

Геномная последовательность — это полное выражение возможностей живого. Соотнесение структур с продуктами определенных генов — первый шаг к пониманию того, каким образом организмы реализуют свою геномную информацию.

Мы хотим понимать структуры молекул, закодированных в геноме, их индивидуальные функции и взаимодействия между собой, а также организацию

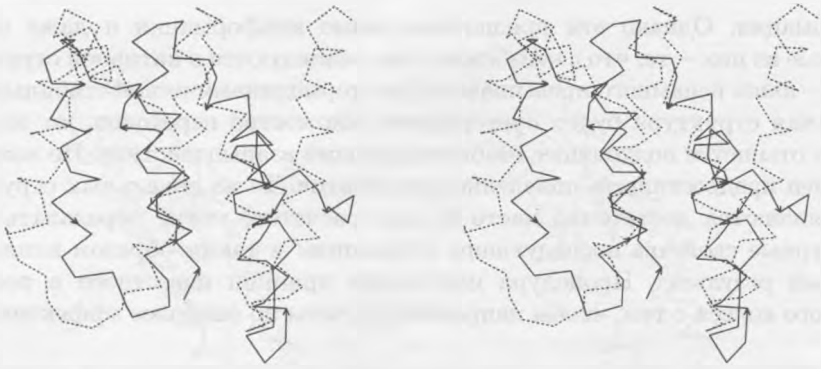


Рис. 5.15. Применение программы LINUS для предсказания С-концевого домена белка ERp29 эндоплазматического ретикулума крысы, представленное на конкурс CASP в 2000 г. Экспериментальная структура обозначена сплошной линией, предсказанная — пунктиром

их деятельности и взаимодействий в пространстве и времени в ходе всего жизненного цикла организма. Мы хотим понимать взаимоотношения молекул, закодированных в геноме индивидуальной особи, а также взаимоотношения молекул из разных особей и видов.

Для конкретного белка его структура — основа понимания механизма его функционирования и взаимодействия с другими молекулами. Для целого организма его строение говорит лишь о том, как используется набор возможных белковых структур и насколько такое использование распространено среди разных функциональных семейств в разных видах. При межвидовых сравнениях может быть установлено родство белковых структур, которое плохо различимо из-за большого различия в последовательностях.

Для аннотирования структуры используется несколько методов:

- *Экспериментальное определение структуры.* Самый лучший метод!
- *Выявление гомологии в последовательностях.* С помощью таких нетривиальных методов, как PSI-BLAST или HMM (скрытые марковские модели), можно обнаружить родство белков не только из организмов одного вида, но и из организмов разных видов. Если из эксперимента известна структура какого-нибудь из гомологов, то можно, по крайней мере, предположить структуру фолда, характерную для всего семейства.
- *Методы распознавания фолда* могут применяться даже в отсутствие гомологии.
- *Специализированные методы* распознают мембранные и сверхспиральные[1)] белки.

Структурное аннотирование позволяет провести по крайней мере частичную инвентаризацию белков в различных геномах, а для семейств, содержащих достаточно близкородственные белки с известной структурой, построить

[1)] Белковый домен, состоящий из параллельно переплетенных α-спиралей. В англоязычной литературе — *D. coiled coil.* — *Прим. ред.*

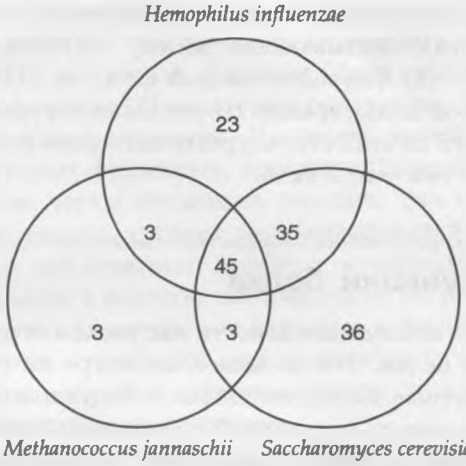


Рис. 5.16. Общие фолды белков археи *Methanococcus jannashii*, бактерии *Haemophilus influenzae* и эукариотического организма *Saccharomyces cerevisiae*. Взято из: [Gerstein, M. (1997), A structural census of genomes: comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure. J. Mol. Biol. 274, 562–76]

детальные пространственные модели. Количество аннотированных структур растет очень быстро, преимущественно благодаря быстрому увеличению числа секвенированных последовательностей и структурных данных. Таблица отражает состояние дел на текущий момент[1)].

Вид	Количество последовательностей	Количество аннотированных структур	%
E. coli	4289	916	21
M. jannaschii	1773	262	14
S. cerevisiae	6289	1109	17
D. melanogaster	13687	2990	21

(Взято из: GeneQuiz, `http://jura.ebi.ac.uk:8765/ext-genequiz/`)

Что говорят эти результаты об освоении возможного белкового репертуара? В настоящий момент известно 350 структурных классов белков, из общего числа, оцениваемого в 1000. Сравнение структур, взятых из геномов археи *Methanococcus jannashii*, бактерии *Haemophilus influenzae* и эукариотического организма *Saccharomyces cerevisiae*, выявило 148 различных фолдов, 45 из которых общие для всех трех видов и, как предполагается, общие для большинства форм жизни. Археа *M. jannashii* имеет наименьшее количество уникальных вариантов укладки (см. рис. 5.16).

Инвентаризация структур, общих для всех трех видов, показала, что пять наиболее общих структурных паттернов доменов таковы: (1) фолд НТФ

[1)] Автор имеет в виду 2002 г. — *Прим. ред.*

(нуклеозидтрифосфат)-гидролазы, содержащий P-петлю, (2) НАД (никотинамидадениндинуклеотид)-связывающий домен, (3) фолд триозофосфатизомеразного бочонка, (4) флаводоксиновый фолд, и (5) тиамин-связывающий фолд. На цветной иллюстрации III показана структура и упрощенная схема топологии первого из этих структурных паттернов (см. также Интернет-задания 5.3 и 5.4). Все они — α/β типа.

Предсказание функции белка

В идеальном случае из последовательности мы узнаем структуру, а затем из структуры — функцию белка. Тем не менее, несмотря на уверенность в том, что схожие аминокислотные последовательности будут иметь схожие белковые структуры, отношения между структурой и функцией более сложные. Белки с одинаковой структурой и даже с одинаковой последовательностью могут использоваться для реализации совершенно разных функций. Очень сильно различающиеся белки могут выполнять одинаковые функции. Более того, как неродственные последовательности могут иметь схожую структуру, так и неродственные белки с различной структурой могут выполнять одну и ту же функцию.

В ходе эволюции белки могут:

1. сохранять функцию и специфичность,
2. сохранять функцию, но менять специфичность,
3. менять функцию на подобную или ту же, но в другом метаболическом «контексте»,
4. переключаться на совершенно другую функцию.

Часто возникает вопрос: насколько должна измениться белковая последовательность или структура для того, чтобы изменилась функция? Ответ таков, что некоторые белки полифункциональны, так что им вовсе нет нужды меняться!

- У утки активная лактатдегидрогеназа и енолаза выполняют функцию кристалликов в хрусталике глаза, хотя и не встречают своих субстратов *in situ*. В других случаях кристаллики очень близки к ферментам, хотя некоторая дивергенция с потерей каталитической активности уже произошла (что доказывает, что ферментативная активность не является необходимой в хрусталике глаза).
- Белок из *E. coli*, называемый Do, DegP или HtrA, действует как шаперон (катализирует сворачивание белков) при низких температурах, а при 42° C превращается в протеиназу. Объясняется это, похоже, тем, что при нормальных условиях или при умеренном тепловом стрессе задача Do — защитить белки, у которых возникли трудности со сворачиванием; а при более жестком тепловом стрессе, когда сохранить белки невозможно, их надо утилизировать.
- Мы уже упоминали о липоатдегидрогеназе *E. coli*, которая также является необходимой субъединицей в пируватдегидрогеназном, 2-оксоглутаратдегидрогеназном комплексах и комплексе расщепления глицина.

Эти примеры соотношения структуры и функции находятся на границах широкого диапазона возможного поведения белков.

Проблемой является то, что непросто определить различия в функциях количественно. Когда две разные функции более похожи друг на друга, чем две другие разные функции? В некоторых случаях под измененной функцией может скрываться общий механизм. Например, семейство енолаз содержит несколько гомологичных ферментов, которые катализируют различные реакции, сохраняя общие черты механизма реакции. Эта группа включает саму енолазу, манделатрацемазу, муконат-лактонизирующий фермент I и D-глюкоратдегидратазу. Все они отрывают протон с α-углеродного атома карбоновой кислоты, образуя енолят в качестве интермедиата. Но последующие превращения и природа продукта варьируют от фермента к ферменту. Эти белки имеют в целом очень схожие структуры типа триозофосфатизомеразного бочонка (TIM-barrel). Разные остатки в активном центре служат причиной того, что ферменты катализируют разные реакции.

Дивергенция функций: ортологи и паралоги

Семейство химотрипсинподобных сериновых протеаз включает близкородственные ферменты, сохранившие свою функцию, и далеко разошедшиеся гомологи, которые приобрели новые функции. Трипсин — это пищеварительный фермент млекопитающих, который катализирует гидролиз пептидных связей, примыкающих к положительно заряженному остатку Arg или Lys. («Карман» специфичности — расщелина на поверхности активного центра — комплементарна по форме и распределению заряда боковой цепи остатка, соседнего с расщепляемой связью). Ферменты со сходными последовательностями, структурами, функциями и специфичностью существуют у многих видов, включая человека, корову, атлантического лосося и даже *Streptomyces griseus* (рис. 5.17). Сходство фермента *S. griseus* с трипсинами позвоночных предполагает горизонтальный перенос генов. Для трех ферментов из позвоночных каждая пара последовательностей имеет $\geqslant 64\%$ идентичных остатков в выравнивании, а бактериальный гомолог — только $\geqslant 30\%$ идентичных остатков с ними; все они очень похожи по структуре. Эти ферменты были названы *ортологами* — гомологичными белками в разных видах. (Последовательности других бактериальных гомологов сильно различаются.)

Эволюция создала также родственные ферменты с разной специфичностью в рамках организмов одного вида. Химотрипсин и панкреатическая эластаза — также пищеварительные ферменты, которые, как и трипсин, разрезают пептидные связи, но следующие в контексте разных аминокислотных остатков: химотрипсин разрезает связи, соседние с большими плоскими гидрофобными остатками (Phe, Trp), а эластаза — смежные с маленькими остатками (Ala). На изменение специфичности влияют мутации остатков в специфическом кармане. Другой гомолог, эластаза лейкоцитов (объект поиска по базе данных в гл. 3), необходима для фагоцитоза и защиты против инфекций. В определенных условиях она ответственна за повреждение легких, приводящее к эмфиземе.

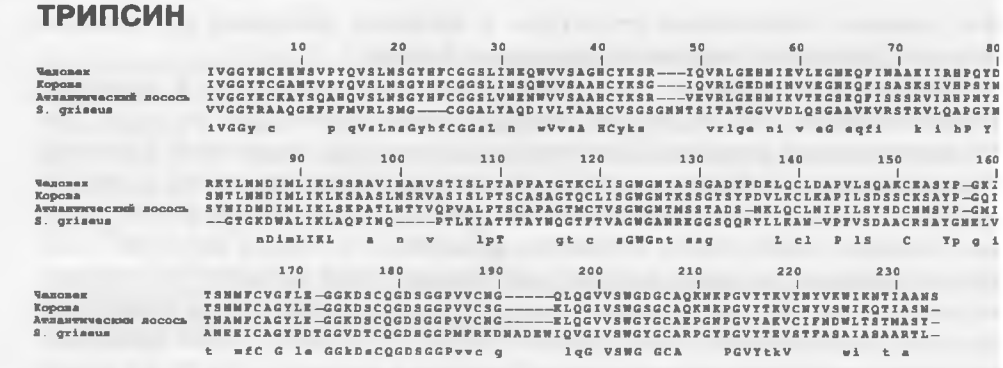


Рис. 5.17. Выравнивание последовательностей трипсинов из человека, коровы, атлантического лосося и *Streptomyces griseus*. В линиях под блоками выравнивания заглавными буквами указаны абсолютно консервативные остатки, маленькими — консервативные в трех из четырех последовательностях (в большинстве случаев, но не во всех, *S. griseus* — исключение)

Некоторые гомологи трипсина приобрели принципиально новые функции:

- Гаптоглобин — гомолог химотрипсина, который потерял протеолитическую активность. Он действует как шаперон, предотвращающий нежелательную агрегацию белков. Гаптоглобин формирует компактный комплекс с фрагментами гемоглобина, вышедшими из эритроцитов, реализуя ряд полезных эффектов, например предотвращение потери иона железа.
- Сериновая протеиназа риновирусов развила специальную, независимую функцию, формируя инициаторный комплекс при синтезе РНК, используя при этом остатки на стороне молекулы, противоположной активному центру протеолиза. Это не модификация активного центра, а создание нового.
- Субъединицы, гомологичные сериновым протеиназам, появляются в плазминогенподобных факторах роста. Роль этих субъединиц в активности фактора роста еще неизвестна, однако они не могут выполнять протеолитическую функцию, потому что были утеряны необходимые каталитические остатки
- «Иммунный» белок насекомых сколексин — далекий гомолог сериновых протеиназ, который вызывает свертывание гемолимфы в ответ на инфекцию.

В семействе химотрипсина мы видим сохранение структуры и функции в близкородственных белках и прогрессивную дивергенцию функций в некоторых, но не всех далеких гомологах.

Идея состоит в том, что общий паттерн сворачивания белка является ненадежным ориентиром для предсказания его функций, особенно для очень

далеких гомологов. Для правильного предсказания функции в далеких белках необходимо обращать внимание на активный центр. Например:

- Вирусные 3С протеиназы были определены как далекие гомологи химотрипсина, несмотря на то что серин их каталитической триады заменен на цистеин.
- Далекая гомология между ретровирусной и аспартатной протеиназами была обнаружена по консервативным остаткам Asp, Thr и Gly.

Так же, как во многих библиотеках мотивов типа PROSITE, эти подходы отталкиваются от характерных паттернов остатков активного центра, и переходят к консервативной функции, даже в отсутствие экспериментальной структуры.

Фокусируясь на активном центре, есть возможность использовать методы, сходные с теми, что используются в дизайне лекарственных препаратов, для предсказания субстратов, которые могли бы связаться с данным белком. Важно использовать и другие экспериментальные данные, такие как паттерны экспрессии белка в ткани, и каталоги взаимодействующих белков. Попытки определить функцию напрямую, например с помощью «нокаута гена», иногда могут быть успешными, однако непродуктивными в случае летального фенотипа при такой мутации или если функцию выполняет группа белков.

Вклад биоинформатики в предсказание функции белка по его аминокислотной последовательности и структуре, по-видимому, не будет исчерпываться одним лишь алгоритмом, дающим однозначный ответ (как, например, существует надежда, что когда-нибудь можно будет предсказывать структуру белка на основе его последовательности). Более разумной целью видится возможность планировать информативные эксперименты и интерпретировать их результаты. Это того стоит!

Открытие и разработка лекарств

Резонно было бы спросить группу студентов, сколько их осталось бы в живых, если бы каждый не прошел курса лечения лекарствами в период серьезной болезни (без учета заболеваний, предотвращенных вакцинацией). Или спросить студентов, сколько их бабушек и дедушек пришлось бы наслаждаться жизнью без регулярного лечения медикаментами. Следует ожидать впечатляющих ответов, которые внушают страх перед новыми штаммами инфекционных микроорганизмов, устойчивых к антибиотикам. Необходимо развивать новые виды лекарств, которые, вкупе с геномной информацией, способной улучшить их специфичность, продлят и улучшат нашу жизнь.

Но как все-таки непросто быть лекарством! Химическое соединение, претендующее на звание лекарства, должно быть:

1. безопасным
2. эффективным
3. стабильным — как химически, так и метаболически
4. транспортабельным — оно должно всасываться и направляться к месту воздействия

5. доступным — добываться из натуральных продуктов или синтетическим путем
6. новым, т. е. патентно чистым.

Этапы развития нового лекарства описаны на с. 301. Этот процесс включает в себя научные исследования, клинические испытания, доказывающие эффективность и безопасность, проработку юридических аспектов, включая вопросы патентования, а также оценку возврата инвестированных денежных средств.

Для создания лекарственного препарата прежде всего необходимо определиться с заболеванием. Понадобится изучить всю информацию о возможных причинах этого заболевания, симптоматике, генетических аспектах, эпидемиологии, взаимосвязи с другими заболеваниями людей и животных, а также обо всех известных путях лечения. Предполагая, что потенциальная польза лекарства оправдывает затрату средств, времени и усилий на его разработку, вы можете начинать.

Вы должны разработать подходящую методику, чтобы с ее помощью могли бы документировать успех уже на ранних стадиях. Если мишенью является известный белок, то связывание препарата с ним может измеряться непосредственно. Потенциальное антибактериальное средство может быть протестировано по влиянию на рост патогена. Некоторые вещества могут быть протестированы по их влиянию на рост эукариотических клеток в тканевых культурах. Если лабораторные животные восприимчивы к исследуемому заболеванию, то испытания могут проводиться на них. Следует иметь в виду, однако, что препараты могут оказывать различное воздействие на животных и на человека. Например, тамоксифен, ныне широко используемый как лекарство при раке груди, первоначально был разработан как противозачаточное средство. На самом деле, это не только эффективный контрацептив для крыс, но и вещество, вызывающее овуляцию у женщин.

Лидерное соединение (Лид)

Основной целью на ранних этапах разработки лекарства является идентификация одного или более лидерных соединений (лидов). Лид — это любое соединение, которое проявляет искомую биологическую активность. Оно должно удовлетворять хотя бы некоторым из требуемых (см. выше) критериев.

Есть несколько способов найти лиды:

1. Озарение: классический пример — пенициллин.
2. Обзор естественных источников. «Перелопатить и найти!» — вот девиз медицинских химиков. Иногда средства народной медицины указывают на источник соединений, проявляющих активность. Например, дигиталис был экстрагирован из листьев наперстянки, которая издавна применялась при сердечной недостаточности. (Почему бы не продолжить просто использовать традиционное средство? Выделение активного компонента делает возможным регуляцию дозировки и исследование прочих вариаций.)
3. Изучение уже известных данных о субстратах, ингибиторах и механизме действия белка-мишени, выбор потенциально активных веществ, исходя из этих характеристик.

Этапы развития нового лекарственного препарата

1. Понимание биологической основы и симптомов заболевания. Вызвано ли оно
   - инфекционным агентом — бактерией, вирусом или чем-то другим?
   - ядовитым веществом небиологического происхождения?
   - белком-мутантом из организма больного?
2. Разработка методики тестирования препарата. Выбирая кандидата в лекарство, можно ли протестировать его на
   - влияние на рост микроорганизмов?
   - влияние на клеточный рост в культуре тканей?
   - влияние на больных животных?
   - связывание с известным белком-мишенью?
3. Существует ли эффективный препарат, применяющийся в народной медицине? Если да, следует перейти к этапу 6.
4. Идентификация специфической молекулярной мишени, обычно белка. Определение его структуры экспериментально или посредством моделирования.
5. Осознание того, какой тип соединения будет соответствовать сайту связывания мишени. Существует ли известный субстрат или ингибитор?
6. Нахождение лида (устоявшийся англоязычный термин — lead compound) — вещества с детектируемой биологической активностью. Лид — всего лишь плацдарм для дальнейших действий; его открытие и последующая модификация (с целью создания лекарства) — принципиально разные вещи.
7. Проработка лида: широкое изучение производных, с целью придать соединению нужные свойства и активность.
8. Предклинические испытания in vitro или на животных, доказывающие эффективность и безопасность. На этой стадии лекарство может быть запатентовано. (На самом деле принято откладывать патентование насколько это возможно из-за конечного времени жизни патента, ведь впереди еще много длительных этапов, предшествующих продаже лекарства.)
9. В США предусмотрена подача документов в виде заявки на новое исследуемое лекарство (Investigational New Drug Application) на рассмотрение в Управление лекарств и продуктов питания (Food and Drug Administration, FDA). За этим следуют три фазы клинических испытаний.
10. Первая фаза клинических испытаний. Тестирование медикамента на безопасность, проводимое на здоровых добровольцах. Исследуется, каким образом лекарство воздействует на организм: как оно адсорбируется, распространяется, метаболизируется и выделяется. Результаты дают основание для определения дозировки.
11. Вторая фаза клинических испытаний. Тестирование лекарства на эффективность приблизительно на 200 больных добровольцах. Лечит ли оно болезнь или лишь смягчает симптомы? Производится уточнение дозировки.
12. Третья фаза клинических испытаний. Тестируется примерно 2000 пациентов для доказательства того, что лекарство лучше, чем предыдущие методы лечения. Это достигается в ходе двукратного тестирования «вслепую», чтобы определить, не является результат эффектом плацебо, и улучшает ли новый препарат результаты уже существующих лекарств. Эти тесты являются очень дорогими;

часто бывает, что до испытаний в третьей фазе дело не доходит, если вторая фаза выявила какие-то побочные эффекты.

13. Запись в FDA, содержащая данные, подтверждающие безопасность и эффективность. Утверждение FDA дает право на продажу лекарства.
14. Четвертая фаза разработки, соразмерная с утверждением FDA и с маркетингом, которая включает продолжительный мониторинг на тему эффективности препарата, отражая широкий опыт его применения. Новые побочные эффекты могут проявиться у некоторых пациентов, приводя к ограничению применения или даже запрещению препарата.

4. Проверка эффективности уже существующих лекарств, применяемых при аналогичных заболеваниях.
5. Широкомасштабный скрининг. Методы комбинаторной химии позволяют проводить параллельное тестирование большого числа сходных веществ. Фаговый дисплей — специальная технология, пригодная для тестирования полипептидов.
6. Использование побочных эффектов уже существующих лекарств. Миноксидил (2,4-диамино-6-пиперидино-пиримидин-3-оксид), первоначально разрабатываемый как антигипертензивный препарат, проявил себя как индуктор роста волос. Другой пример — виагра, изначально разрабатывался как лекарство от сердечных заболеваний.
7. Скрининг. Национальный институт рака в США проверил десятки тысяч препаратов. (Поиск среди аналогов очень важен, даже после нахождения лидерного соединения).
8. Компьютерный скрининг и *ab initio* (от начала) компьютерный дизайн.

Открытие лида дает начало другим видам исследовательской деятельности. Многие из производных лидов должны быть протестированы для улучшения его эффективности и придания других важных свойств. Например, соединение, которое связывается со своей мишенью, не является лекарством, если оно не доставлено к своей мишени. Доставка лекарства к мишени внутри организма требует его всасывания и транспортировки. Необходимо, чтобы лекарство было достаточно водорастворимо, чтобы оно могло всасываться, но не настолько, чтобы оно сразу выводилось из организма. Оно также должно быть достаточно жирорастворимо, чтобы проникать через мембраны, но не настолько, чтобы откладываться в жировых запасах.

Уточнение лида: количественное соотношение структура—активность (QSAR)

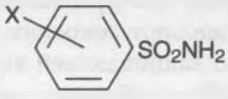
Вещества, подобные соединению, имеющему фармакологическую активность, обычно обладают аналогичной активностью, но отличаются своей эффективностью и специфичностью. Начиная с лида, химики должны проверить множество похожих молекул для того, чтобы оптимизировать фармакологические свойства. Для систематического поиска было бы очень полезно понимать,

какие изменения структуры и физико-химических свойств семейства молекул коррелируют с фармакологическими свойствами. Проблема заключается в том, что молекулы характеризуются большим количеством дескрипторов. Они включают структурные особенности строения молекулы, такие как природа и расположение заместителей, экспериментальные свойства, такие как растворимость в водной и органической среде, или дипольный момент. Важны также расчетные параметры, такие как заряд каждого атома.

Система QSAR предлагает методы для предсказания фармакологической активности множества веществ, основываясь на соотношениях между свойствами молекулы и ее активностью, полученных на тестовой выборке. Метод был разработан К. Ханшем (C. Hansch) и коллегами в 1960-е годы и широко использовался.

Группа ученых C. Hansch, J. McClarin, T. Klein и R. Langridge применяли методы QSAR для исследования карбоангидраз. Карбоангидразы катализируют реакцию разложения угольной кислоты $CO_2 + H_2O \rightleftharpoons H^+ + HCO_3^-$. В клинической практике ингибиторы карбоангидраз применяются как диуретики, для лечениея высокого внутриглазного давления при глаукоме с помощью подавления водянистых выделений из глаза, и как антисептики. Альпинисты при высотных восхождениях используют ингибиторы карбоангидраз для подавления симптомов горной болезни.

Изучение комплексообразования карбоангидразы с 29 лигандами — фенилсульфонамидами:



(где X — разные заместители в кольце, которые варьировали как по структуре, так и по положению) показывает, что константа образования зависит от константы Хамметта σ, которая характеризует электронодонорные или электроноакцепторные свойства заместителя, а также от коэффициента распределения P неионизированной формы лиганда в системе вода-октанол и положения заместителя в кольце (*орто* или *мета*):

$$\log K \approx 1.55\sigma + 0.65 \log P - 2.07I_1 + 3.28I_2 + 6.94$$

где K — константа образования; $I_1 = 1$, если X в *мета*-положении, иначе 0; $I_2 = 1$, если X находится в *орто*-положении, иначе 0; X — алкил, $^-$COO-алкил, $^-$CONH-алкил.

Поэтому при разработке нового лекарства можно применять следующие приемы:

1. Многие вещества можно «изучать» с помощью компьютера, что позволяет подобрать несколько лучших кандидатов для дальнейшей экспериментальной проверки.
2. По рассчитанным параметрам можно предположить структуру сайта связывания:
   - Если коэффициент σ положительный, то заместитель должен проявлять электронакцепторные свойства; поэтому в ионизированной форме $^-SO_2NH_2$ связывается с ионом цинка в активном центре карбоангидразы.

- Положительный коэффициент $\log P$ предполагает гидрофобные взаимодействия белка с лигандом.
- Отрицательные коэффициенты I_1 и I_2 означают стерические затруднения заместителей в *мета-* или *орто-* положении.

Структуры карбоангидразы с лигандами подтверждают эти предположения (см. Интернет-задание 5.8).

Компьютерный дизайн лекарств

ПРИМЕР 5.5.

Дизайн лекарственного препарата при помощи компьютера: специфические ингибиторы простагландинциклооксигеназы 2.

Простагландины относятся к природным веществам — медиаторам очень многих физиологических процессов. Фармакологические применения простагландинов основаны на свойствах собственно простагландинов, так и на препаратах, блокирующих их синтез. Простагландин E_2 (динопростон) используется в акушерстве для облегчения родов. Аспирин, ибупрофен, ацетаминофенол (тайленол) и другие нестероидные противовоспалительные препараты (НСПВП) эффективны при артритах и связанных заболеваниях (см. с. 307). Их активность основана на ингибировании ферментов пути синтеза простагландинов, в частности, простагландинциклооксигеназ. Хорошо известен побочный эффект аспирина — кровотечение стенки желудка. Это происходит потому, что простагландины (продукция которых блокируется аспирином) подавляют секрецию кислоты в желудке и активируют формирование слизи, защищающей внутреннюю стенку желудка.

Аспирин и другие НСПВП ингибируют две родственные циклооксигеназы, называемые COX-1 и COX-2. (К сожалению, та же аббревиатура используется для цитохром оксидаз 1 и 2). COX-1 экспрессируется конститутивно в эпителии желудка, а COX-2 является индуцируемым ферментом, экспрессия которого повышается при воспалении. Из этого следует, что лекарство, которое будет ингибировать COX-2, но не будет ингибировать COX-1, сохранит желательную активность НСПВП, но не будет приводить к нежелательным побочным эффектам.

Аминокислотные последовательности и кристаллические структуры COX-1 и COX-2 известны. (Последовательности этих белков идентичны на 65%). На рис. 5.18 показана часть структуры белка COX-1, ацетилированного аналогом аспирина 2-бромацетоксибензойной кислотой (аспирин, бромированный по метильной группе ацетильного остатка). Салицилатная группа находится рядом. Такое соединение блокирует вход в активный сайт. Большинство препаратов группы НСПВП связываются на активном сайте, но не модифицируют фермент ковалентно.

На рис. 5.19 показана та же картинка с наложением соответствующего участка COX-2. Можно увидеть структурные различия, которые могут дать ключ к созданию селективного лекарства. На рис. 5.20 показан участок COX-2 с селективным ингибитором SC-558 (1-фенилсульфонамид-3-трифторметил-5-*пара*-бромфенилпиразол, разработанный компанией Searle). Из рис. 5.21 мы видим, почему SC-558 не может ингибировать

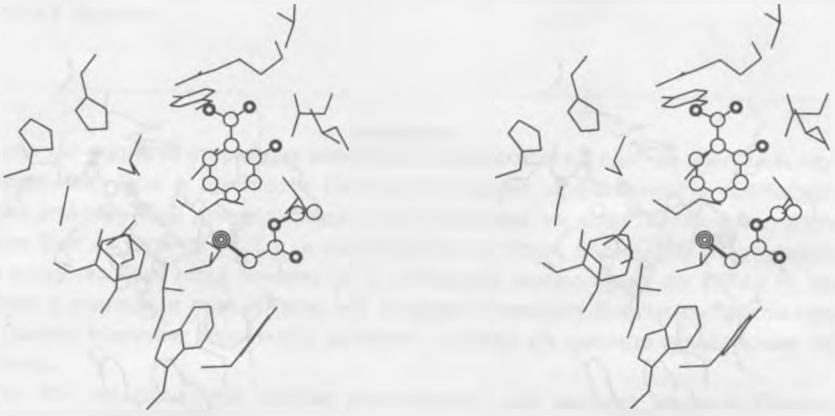


Рис. 5.18. Сайт связывания COX-1 с аналогом аспирина 2-бромацетоксибензойной кислотой. Лиганд реагирует с ферментом, перенося бромацетильную группу на боковую цепь серина 530. Белок показан скелетной моделью. Аналог аспирина показан шариковой моделью

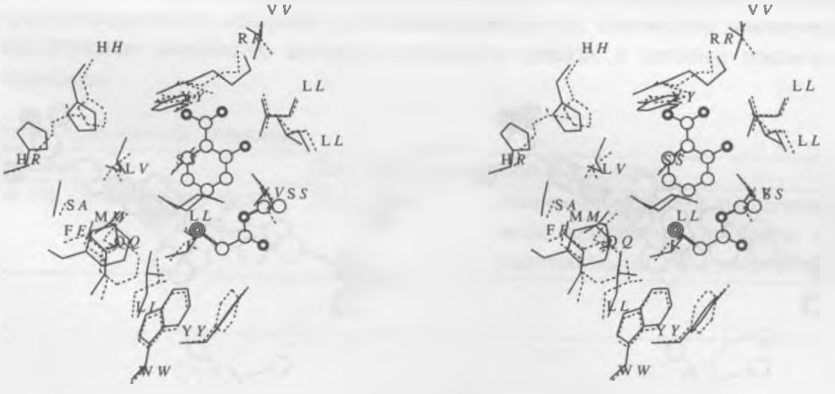


Рис. 5.19. Сайт связывания COX-1 с аналогом аспирина 2-бромацетоксибензойной кислотой. Сплошными линиями и прямыми метками отмечена структура COX-1. Пунктирными линиями и курсивом показана структура COX-2. Видите ли вы незанятое пространство в этом сайте, которое могло бы разместить лиганд большего размера? Видите ли вы разницу в последовательности, которую можно было бы использовать для дизайна ингибитора, который бы связывался с COX-2 (пунктир), но не связывался бы с COX-1 (сплошные линии)?

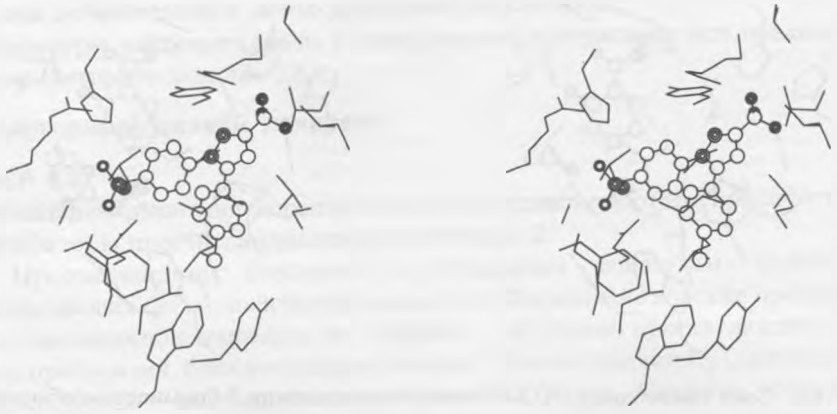


Рис. 5.20. Сайт связывания в COX-2 для селективного ингибитора COX-2 SC-558 (1-фенилсульфонамид-3-трифторметил-5-*пара*-бромфенилпиразола)

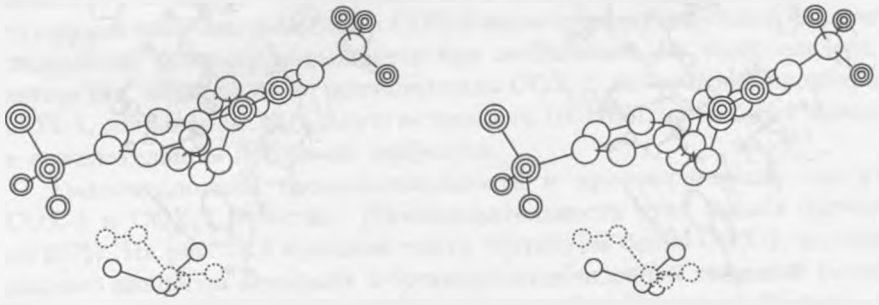


Рис. 5.21. SC-558 и аминокислотные остатки в COX-2 (сплошные линии, валин) и в COX-1 (пунктирные линии, изолейцин), которые обуславливают селективность. SC-558 не может связываться с COX-1, поскольку возникают стерические ограничения при взаимодействии с изолейцином

COX-1. При связывании возникали бы стерические затруднения из-за взаимодействия с боковой цепью изолейцина, который в COX-2 соответствует остатку валина.

Аспирин

Аспирин — одно из старейших всем известных лекарств и одно из новейших научно обоснованных. Еще в древности Гиппократ отмечал эффективность препарата из листьев или коры ивы для ослабления боли и понижения жара. Активный компонент салицин был выделен в 1828 г. и синтезирован в 1859 г. Колбе. Механизм действия этого вещества был тогда неизвестен и оставался неизвестным до 1970-х гг., когда Дж. Вэйн с коллегами обнаружили, что аспирин блокирует биосинтез простагландинов. Однако незнание механизма действия никогда не мешало применению этого лекарства.

Уже 100 лет салицилат натрия используется для лечения артрита. Поскольку раздражение желудка было серьезным побочным эффектом, Ф. Хоффман стремился уменьшить кислотность соединения и предложил ацетилсалициловую кислоту, или аспирин. Аспирин был первым синтетическим лекарством, положившим начало современной фармакологической индустрии. (Название салицин происходит от латинского названия ивы — *salix*, а название аспирин произошло от слов ацетил и спир — *spirea* — растение, которое было другим природным источником салицина).

Аспирин обладает жаропонижающим эффектом и ослабляет головную боль. В больших дозах он эффективен против артрита. Его также используют для профилактики сердечных приступов и инсультов. Применение в сердечно-сосудистой медицине связано с подавлением свертываемости крови, что обусловлено ослаблением простагландинового контроля слипания тромбоцитов. Множество применений аспирина отражает множество физиологических процессов, в которые вовлечены простагландины.

Множество применений аспирина

Малые дозы	Средние дозы	Большие дозы
Влияние на свертывание крови	Жар, боль	Уменьшение болей при воспалениях, вызванных артритом или родственными заболеваниями

Литература

Сворачивание (фолдинг) белков

Baldwin, R. L. and Rose, G. D. (1999) 'Is protein folding hierarchic? I. Local structure and peptide folding. II. Folding intermediates and transition states', *Trends in Biochemical Sciences* 24, 26–32; 77–83. [Введение в современное понимание стабильности и фолдинга белков.]

Совмещение структур и структурные выравнивания

Holm, L. and Sander, C. (1995) 'Dali: a network tool for protein structure comparison', *Trends in Biochemical Sciencs* 20, 478–80. [Описывает DALI и соответствующее применение в выравнивании структур.]

Smith, T. F. (1999) 'The art of matchmaking: sequence alignment methods and their structural implications', *Structure with Folding and Design* 7, R7–R12. [Описание работы, объединяющей анализ последовательности и структуры. Взаимосвязи между последовательностями и структурами.]

Das, R., Junker, J., Greenbaum, D., and Gerstein, M. B. (2001) 'Global perspectives on proteins: comparing genomes in terms of folds, pathways and beyond', *The Pharmacogenomics Journal* 1, 115–25. [Интегральный взгляд на геномные исследования.]

Koonin, E. V. (2001) 'Computational genomics', *Current Biology* 11, R155–R158. [Точка зрения о том, куда мы идем.]

Моделирование по гомологии

Guex, N., Diemand, A. and Peitsch, M. C. (1999) 'Protein modelling for all', *Trends in Biochemical Sciences* 24, 364–7. [Описание SWISS-MODEL.]

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Šali, A. (2000) 'Comparative protein structure modeling of genes and genomes', *Annual Review of Biophysics and Biomolecular Structure* 29, 291–325. [Состояние дел в моделировании по гомологии и его использовании в структурной геномике.]

Peitsch, M. C., Schwede, T., and Guex, N. (2000) 'Automated protein modelling — the proteome in 3D', *Pharmacogenomics* 1, 257–66. [Что потребуется для решения проблем структурной геномики.]

Другие методы предсказания структуры белков

Bonneau, R. and Baker, D. (2001) 'Ab initio protein structure prediction: progress and prospects', *Annual Review of Biophysics and Biomolecular Structure* 30, 173–189. [Обзор методов предсказания структуры авторами наиболее успешных работ в этой области.]

Классические публикации, которые стоит почитать

Kauzmann, W. (1959) 'Some factors in the interpretation of protein denaturation', *Advances in Protein Chemistry* 14, 1–63.

Richards, F. M. (1977) 'Areas, volumes, packing and protein structure', *Annual Review of Biophysics and Bioengineering* 6, 151–76.

Chothia, C. (1984) 'Principles that determine the structure of proteins', *Annual Review of Biochemistry* 53, 537–72.

Richards, F. M. (1991) 'The protein folding problem', *Scientific American* 264(1), 54–7, 60–3.

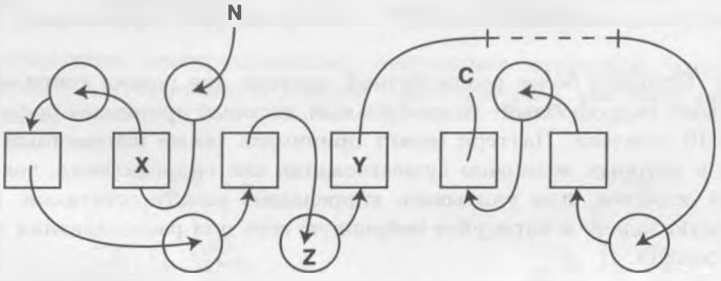
Упражнения, задачи и компьютерные задания

Упражнение 5.1. При температуре замерзания теплота возгонки льда 51 кДж/моль. В твердом состоянии каждая молекула воды имеет две водородные связи. Чему равна энергия одной водородной связи вода–вода?

Упражнение 5.2. Какие пары являются ортологами, какие — паралогами, какие — ни тем ни другим?

(а) Человеческий гемоглобин α и человеческий гемоглобин β.
(б) Человеческий гемоглобин α и гемоглобин лошади α.
(в) Человеческий гемоглобин α и гемоглобин лошади β.
(г) Человеческий гемоглобин α и человеческий гемоглобин γ.
(д) Протеазы: человеческий химотрипсин и человеческий тромбин.
(е) Протеазы: человеческий химотрипсин и актинидин из киви.

Упражнение 5.3. На копии цветной иллюстрации VIIa отметьте положение структурных элементов X, Y, Z.



Упражнение 5.4. На копии рис. 5.9, *б* отметьте спираль 3_{10}, которая не была предсказана как спираль.

Упражнение 5.5. Какая из показанных топологий отражает правильную структуру (порядок тяжей и ориентацию), показанную на рис. 5.9, *б*?

(а) ↑ ↑ ↑ ↑ (1 2 3 4) (б) ↑ ↓ ↑ ↓ (3 4 2 1) (в) ↑ ↑ ↓ ↑ (1 3 2 4)

Упражнение 5.6. Рассмотрите структурные предсказания гипотетического белка из *H. Influenzae*, показанные на рис. 5.12: (а) В чем разница в способе укладки целевого белка и экспериментальных «родителей»? (б) Чем отличается предсказание А. Г. Мурзина от укладки целевого белка? (в) Чем отличается предсказание А. Г. Мурзина от структуры родственного белка? Почему предсказание А. Г. Мурзина лучше, чем укладка родственного белка?

Упражнение 5.7. Нарисуйте химическую структуру 2-бромацетоксибензойной кислоты.

Упражнение 5.8. Многие белки патогенов имеют гомологов у человека. Предположим у вас есть метод сравнения детерминант специфичности сайтов связывания для гомологичных белков. Как вы могли бы использовать этот метод для того, чтобы отобрать подходящую мишень для разработки лекарств?

Упражнение 5.9. Рассмотрите нейронную сеть, показанную на с. 277 внизу. Сколько параметров — весов и порогов — необходимо подобрать в предположении о линейности функции отклика?

Упражнение 5.10. Какова геометрическая интерпретация поведения нейрона, который имеет два входа x, y и откликается, если и только если $x + 2y \geqslant 2$?

Упражнение 5.11. Нарисуйте нейрон с двумя входами x, y, которые могут принимать любые значения и который откликается, если первое значение больше второго. Какова геометрическая интерпретация этого нейрона?

Задача 5.1. На множественном выравнивании последовательностей доменов ETS (см. задачу 1.1): (а) какие белки являются наиболее близкими и какие наиболее далекими членами семейства? (б) предположим, что пространственная структура известна только для первого белка. Для каких других белков этого семейства вы предполагаете возможность предсказания структуры со средним отклонением $\leqslant 1.0$Å для 90% остатков?

Задача 5.2. Нарисуйте нейронную сеть, получающую 8 входных сигналов, которые могут принимать значения 0 или 1, и соответствующую 8 остаткам в последовательности из 8 аминокислот. Для i-го входа 0 означает, что соответствующий остаток гидрофильный, 1 — гидрофобный. Сеть должна продуцировать выход 1, если есть спиральный паттерн — для простоты PPHHPPHH, где H означает гидрофобный (незаряженный) остаток, P — полярный (или заряженный), и 0 — если иное.

Задача 5.3. Опишите более реалистичный паттерн для поиска спиралей в алфавите символа гидрофобный—гидрофильный, который принимает последовательность из 10 остатков. Паттерн может принимать также неизвестный символ — позиции, в которых возможно существование как гидрофобного, так и гидрофильного остатков, или учитывать корреляцию между остатками. Обобщите предыдущую задачу и нарисуйте нейронную сеть для распознавания этой более сложной задачи.

Задача 5.4. Мы, как и компьютеры, можем производить арифметические и логические операции. Определим 1 = истина (True, T) и 0 = ложь (False, F). Нарисуйте нейроны с двумя входами, каждый из которых может принимать значения 0, 1, и линейное правило реализации основных логических операций: (а) логическое И, (б) логическое ИЛИ, (в) какова будет простейшая нейронная сеть для реализации логической операции ИСКЛЮЧАЮЩЕЕ ИЛИ. Можно ли эту операцию осуществить на одном нейроне? Если нет, то какое наименьшее число слоев необходимо?

Задача 5.5. Модифицируйте PERL-программу для рисования спиральных кругов (с. 261) так, чтобы разные аминокислотные остатки представлялись одним шрифтом, но отмечались разными цветами: GAST — голубым; CVILFYPMW — зеленым; HNQ — лиловым; DE — красным; KR — синим.

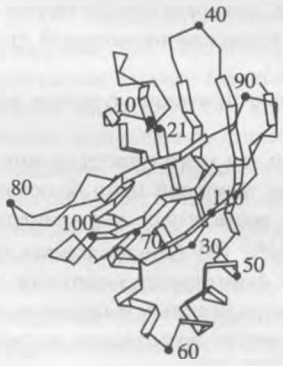
Задача 5.6. Анализ гидрофобных кластеров. Предположим, что некий фрагмент белка образует α-спираль. Чтобы представить ее поверхность, представьте себе последовательность, скрученную в α-спираль (даже если на самом деле они образуют β-тяжи или петли). Затем «соедините» поверхность спирали и прокатите ее по листу бумаги так, чтобы имена остатков «отпечатались» на нем. Если прокатить спираль дважды, то вся поверхность будет видна. На такой диаграмме хорошо идентифицируются гидрофобные пятна на поверхности. Таким путем можно пытаться предсказать участки последовательности, которые формируют спирали в нативной структуре. Сравнение гидрофобных кластеров может также помочь определить дальнее родство белков. Напишите PERL-программу, которая печатает подобную диаграмму.

Задача 5.7. На конкурсе CASP в 2000 г., в категории мишеней, для которых аналоги со схожими фолдами неизвестны, одним из объектов был N-концевой домен (аминокислотные остатки с 1 по 116) белка человека XRCC4, присоединяющегося к концу ДНК из системы репарации ДНК.

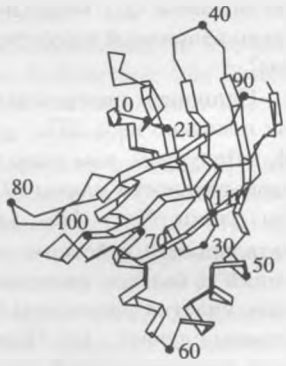
Вторичная структура, предсказанная Б. Ростом с использованием метода PROF (основанное на профилях предсказание с помощью нейронных сетей), получилась следующей (здесь буква H под аминокислотным остатком означает, что для данной позиции предсказано спиральное состояние, E — предсказание тяжа, «–» означает «другое»):

```
                          1         2         3         4         5         6
                          0         0         0         0         0         0
Последовательность  MERKISRIHLVSEPSITHFLQVSWEKTLESGFVITLTDGHSAWTGTVSESEISQEADDMA
Предсказание        --EEEEEEE---HHHHHH-HHHHHHH-EEEEEE---EE--HHHHHHHHHHH

                                                  1         1
                          7         8         9         0         1
                          0         0         0         0         0
Последовательность  MEKGKYVGELRKALLSGAGPADVYTFNFSKESCYFFFEKNLKDVSFRLGSFNLEKV
Предсказание        HHH-HHHHHHHHHHH---EEEEEE---EEEEE---EEEE---HHHH
```

Ниже изображена экспериментально полученная структура этого домена (PDB-код 1FU1), открытая для доступа уже после того, как было предложено предсказание:



HUMAN XRCC4 [1fu1] domain1



HUMAN XRCC4 [1fu1] domain1

Распределение остатков по элементам вторичной структуры согласно структуре из PDB оказалось следующим:

Элемент вторичной структуры	Номера позиций
Спираль	27–29, 49–59, 62–75
Лист 1	2–8, 18–24, 31–37, 42–48, 114–115
Лист 2	84–88, 95–101, 104–111

(a) Подсчитайте значение Q_3: процент оснований, правильно отнесенных к спирали (H), тяжу (E) или другим элементам вторичной структуры (-).

(б) На копии рисунка, изображающего белок XRCC4, раскрасьте различными цветами области, для которых *предсказано* нахождение в составе спиралей или тяжей.

(в) Используя результат задания б: Сколько предсказанных спиралей совпадают с экспериментально определенными? Сколько предсказанных тяжей совпадают с экспериментально определенными?

Задача 5.8. На конкурсе CASP в 2000 г. Bonneau, Tsai, Ruczinski и Baker осуществили предсказание полной пространственной структуры белка XRCC4 (аминокислотные остатки с 1 по 116). Предсказание вторичной структуры, полученное из их модели, выглядит следующим образом: (H — спираль, E — тяж, «–» — остальное):

```
                  1         2         3         4         5         6
                  0 .       0         0         0         0         0
Sequence    MERKISRIHLVSEPSITHFLQVSWEKTLESGFVITLTDGHSAWTGTVSESEISQEADDMA
Prediction  --E-EEEE--EEEE-EHHHHHHHH--EEEE-EEEE---HHHHHHHHHHHH

                  7         8         9         0         1
                  0         0         0         0         0
Sequence    MEKGKYVGELRKALLSGAGPADVYTFNFSKESCYFFFEKNLKDVSFRLGSFNLEKV
Prediction  HHH--HHHHHHHHHHH---EEEEEEE-EEEEEEE---HHHH--HHHH
```

(a) Каково значение Q_3 для этого предсказания? (b) Какой из методов в данном случае дал лучшие результаты для предсказания вторичной структуры (исходя из значений Q_3): метод нейронных сетей, который осуществляет предсказание только вторичной структуры, или метод предсказания полной трехмерной структуры?

Задача 5.9. Напишите программы PERL, которые применяют метод нейронных цепей, как показано на с. 277.

Задача 5.10. Допустим, вам надо оценить, используя метод протягивания, может ли последовательность длины M иметь укладку основной цепи белка известной структуры длины $N > M$. (a) Сколько из возможных выравниваний этих последовательностей действительно возможны[1)]? (b) Предположим, что половина позиций в белке с известной структурой формирует α-спирали и внутри спирального участка запрещены бреши. Сколько различных выравниваний существует в таком случае? (c) Подсчитайте количество возможных выравниваний при каждом из этих условий, если $N = 200$, и $M = 150$?

Задача 5.11. Напишите программу PERL для подсчета приближенного значения числа π методом Монте-Карло: квадрат на плоскости с вершинами в точках $(0,0)$, $(0,1)$, $(1,0)$ и $(1,1)$ имеет площадь, равную 1. Создайте серию *пар* случайных чисел (x,y) из интервала $[0,1]$, и таким образом генерируйте точки, случайно распределенные внутри этого квадрата. Подсчитайте число точек, которые лежат внутри круга радиуса 0,5, вписанного в этот квадрат. Отношение числа точек, которые попали внутрь этого круга, к общему количеству точек, равно отношению площади круга к площади квадрата, т. е. $\pi/4$. Определите соотношение между выбранным количеством точек и числом верных цифр в подсчитанном значении числа π. Оцените количество точек, необходимое для определения числа π с точностью 50 знаков после запятой.

[1)] Совместимы со структурным протягиванием. — *Прим. ред.*

Задача 5.12. Для того чтобы конвертировать сигнал нейрона из ступенчатой функции в гладкую (см. с. 278), можно заменить утверждение «Пусть X есть некоторая взвешенная сумма входных данных; тогда на выходе получаем 1, если $X > 0$, иначе 0» на утверждение следующего вида «Пусть X есть некоторая взвешенная сумма входных данных; тогда на выходе получаем $1/(1+e^{-X})$.» (а) Проверить, что при $X \to -\infty$, $1/(1+e^{-X}) \to 0$, при $X \to +\infty$, $1/(1+e^{-X}) \to 1$, и при $X = 0$, $1/(1+e^{-X}) = 0.5$. (б) Предположим, что нейронная сеть для определения: лежит ли точка в пределах треугольника, преобразована так, что выходной сигнал каждого нейрона скорее описывается гладкой функцией $1/(1+e^{-X})$, чем ступенчатой, и что точка считается внутри принятой площади, если на выходе сеть выдает > 0.5. Напишите на PERL программу, которая определяет, какая область задана в таком случае.

Интернет-задание 5.1. Бактерия *Pseudomonas fluorescens* и гриб *Curvularia inaequalis* содержат хлорпероксидазу — фермент, катализирующий реакцию галогенирования. Имеют ли эти ферменты одинаковый паттерн укладки?

Интернет-задание 5.2. Вычислите профиль гидрофобности для бычьего родопсина, определите на его основе номера аминокислотных остатков, которые формируют трансмембранные спирали, и сравните результат с экспериментальными спиралями, полученными из рентгеноструктурного анализа (PCA).

Интернет-задание 5.3. На цветной иллюстрации III показана структура тиаминсвязывающего домена, определенная M. Gerstein как одна из пяти наиболее общих паттернов фолда среди архей, бактерий и эукариот. Используя средства, доступные в SCOP, нарисуйте четыре другие структуры.

Интернет-задание 5.4. Используя результаты Интернет-задачи 5.3, или изображения, доступные в книге *Introduction to Protein Architecture: The Structural Biology of Proteins*, нарисуйте (по аналогии с цветной иллюстрацией III), упрощенные диаграммы топологии для остальных четырех структур.

Интернет-задание 5.5. Кодирует ли ген человеческого глобина θ_1 активный глобин? Или он в действительности является псевдогеном? Отправьте аминокислотную последовательность человеческого глобина *θ1* на SWISS-MODEL, включив в запрос отчет программы WhatCheck. Что вы можете заключить из полученного результата относительно статуса гена человеческого глобина θ_1?

Интернет-задание 5.6. Сравните число аннотаций в SCOP в различных категориях, перечисленных на с. 271, с тем, что SCOP содержит на данный момент.

Интернет-задание 5.7. Сделайте выравнивание последовательностей γ-химотрипсина и эпидермолитического токсина A из *S. aureus*, используя методы парного выравнивания. Сравните результат со структурным выравниванием, показанным в тексте.

Интернет-задание 5.8. Сделайте выравнивание последовательностей эластазы нейтрофила человека и эластазы из *C. elegans*. (а) Сколько идентичных аминокислотных остатков в оптимальном выравнивании? (б) Разумно ли строить модель эластазы из *C. elegans*, начиная со структуры эластазы нейтрофила человека?

Интернет-задание 5.9. S. Chakravarty и K. K. Kannan определили строение карбоангидразы с бензолсульфонамидом в качестве лиганда (PDB-код 1CZM). Изобразите участок связывания, отобразив природу взаимодействий белка и лиганда. Опишите природу взаимодействия, полученную в результате QSAR-анализа.

Заключение

Какой нам видится биоинформатика в будущем? Очевидно, что сбор экспериментальных данных продолжится и будет только ускоряться. Возрастающие компьютерные мощности будут применяться для хранения данных, распространения и анализа. Усовершенствованные алгоритмы будут разработаны для анализа и интерпретации получаемой информации, а также для ее преобразования в знания и мудрость.

Одна из вех будет достигнута, когда наши знания последовательностей и структур станут практически полными, т. е. будет собрано достаточно плотное подмножество данных относительно ныне живущих видов. (Речь, конечно же, не идет о том, чтобы знать все.) Это будет означать, что при исследовании случайного участка генома или разрешении новой белковой структуры станет гораздо более вероятным, что натолкнешься на что-либо уже известное, нежели будешь открывать что-то совсем новое. В конце концов природа является системой с неограниченными возможностями, но с конечным набором реализованных вариантов.

Приложения биоинформатики станут более реалистичны и будут гораздо быстрее созревать от стадии заоблачных научных изысканий до использования в промышленности и клинической практике. Некоторые виды передачи биологической информации на более высоком уровне — такие как программы генетического развития индивида на протяжении его жизни или программы активности человеческого ума — попадут в число процессов, которые мы можем описать количественно и анализировать на уровне молекул и их взаимодействий.

На плафоне Сикстинской Капеллы фреска Микеланджело, где Змей предлагает Еве фрукт с дерева знания, а ноги закручены вокруг дерева как двойная спираль. Мы можем надеяться, что наше искушение новым знанием, воплощенное в другую двойную спираль, будет иметь более счастливые последствия.

Оглавление

ИМЕЕТСЯ В ПРОДАЖЕ

Примроуз С. ***Геномика. Роль в медицине*** */ С. Примроуз, Р. Твайменн ; пер. с англ. — 2008. — 277 с. : ил.*

В учебном издании обсуждается роль достижений биотехнологии, а также нового направления биологии — геномики — в развитии современной медицины. Рассмотрены подходы при крупномасштабных структурных и функциональных исследованиях полных геномов различных организмов. Представлены сведения об этиологии, патогенезе и диагностике инфекционных заболеваний, рассмотрены новые пути борьбы с бактериальными, вирусными, грибковыми и протозойными инфекциями. Изложены представления о молекулярных механизмах возникновения наследственных заболеваний и рака, описаны методы установления взаимосвязи между указанными заболеваниями и нарушениями в хромосомах и отдельных генах. Рассмотрены основные подходы к скринингу новых лекарственных препаратов, показан вклад геномики в развитие новых видов терапии.

Для студентов, изучающих молекулярную биологию, генную инженерию, геномику, молекулярную медицину, а также для научных работников.

ИЗДАТЕЛЬСТВО
«БИНОМ
Лаборатория знаний»

125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272
e-mail: binom@Lbz.ru, http://www.Lbz.ru
Оптовые поставки:
(499) 174-7616, 171-1954, 170-6674

ИМЕЕТСЯ В ПРОДАЖЕ

Ребриков Д. В. ***ПЦР «в реальном времени»*** */ Д. В. Ребриков, Г. А. Саматов, Д. Ю. Трофимов и др. ; под ред. д. б. н. Д. В. Ребрикова. — 2009. — 215 с. : ил.*

Рассмотрены различные варианты и особенности оборудования для проведения ПЦР «в реальном времени», даны рекомендации по выбору амплификатора. Разобраны особенности систем флуоресцентной регистрации накопления ДНК. Рассмотрены ключевые факторы, определяющие выбор последовательности олигонуклеотидов и параметры программ амплификации. Уделено внимание подготовке проб и особенностям анализа получаемых данных, что необходимо для получения наиболее достоверных результатов. Отдельные главы посвящены применению ПЦР «в реальном времени» для решения различных задач: определения уровня представленности транскриптов, вирусной нагрузки, нуклеотидного полиморфизма, относительного содержания нуклеиновых кислот на примере ГМО.

Для сотрудников генно-инженерных и медицинских диагностических лабораторий, а также для преподавателей и студентов, специализирующихся в области молекулярной биологии.

ИЗДАТЕЛЬСТВО
«БИНОМ
Лаборатория знаний»

125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272
e-mail: binom@Lbz.ru, http://www.Lbz.ru
Оптовые поставки:
(499) 174-7616, 171-1954, 170-6674

На стыке биологии и информатики родилась новая научная область – биоинформатика.

Интенсивное развитие биоинформатики совпало по времени с победным шествием компьютерных технологий. Биоинформатика в очень большой степени зависима от ресурсов Интернета и успешно развивается благодаря этим ресурсам.

Наши новые представления о строении, механизмах функционирования и регуляции живых систем во многом основаны на результатах, полученных методами биоинформатики. Потенциал биоинформатики играет большую роль в развитии медицинских и других технологий на благо человечества.

Автор этой книги доктор Артур Леск, опираясь на свой опыт и знания, знакомит читателя с приемами биоинформатики и ее приложениями в научных исследованиях. В книге показано, как, не владея теоретическими основами информатики и сложными приемами программирования, можно эффективно использовать вычислительные методы биоинформатики. Значительное внимание уделено развитию практических навыков и решению типовых биологических задач.

Для студентов и научных работников.

ISBN 978-5-94774-501-6