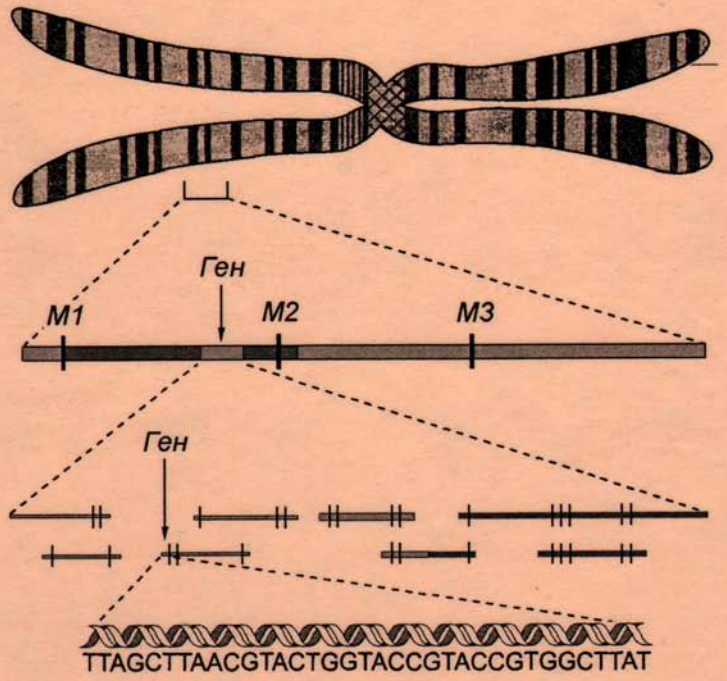
А. Н. Огурцов

ВВЕДЕНИЕ В БИОИНФОРМАТИКУ



Учебное пособие

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ УКРАИНЫ

НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«Харьковский политехнический институт»

А. Н. Огурцов

ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Учебное пособие
по курсу «Биоинформатика и информационная биотехнология»

для студентов направления подготовки
051401 «Биотехнология»,
в том числе для иностранных студентов

Утверждено
редакционно-издательским
советом университета,
протокол № 2 от 01.12.2010 г.

Харьков НТУ «ХПИ» 2011

ББК 28.071.3
О-39
УДК 577.3

Рецензенты:

Ю.В. Малюкин, д-р физ.-мат. наук, проф., зав. отд. нанокристаллических материалов, зам. директора ИСМ НАН Украины

В.А. Карачевцев, д-р физ.-мат. наук, зав. отд. молекулярной биофизики ФТИНТ им. Б.И. Веркина НАН Украины

Навчальний посібник містить матеріали з основних питань першого розділу курсу «Біоіформатика та інформаційна біотехнологія» відповідно до програми підготовки студентів напряму «Біотехнологія».

Призначено для студентів спеціальностей біотехнологічного профілю всіх форм навчання.

Огурцов, А. Н.

О-39 Введение в биоинформатику : учеб. пособие [по курсу «Биоинформатика и информационная биотехнология» для студ. направл. подг. 051401 «Биотехнология», в т. ч. иностр. студ.] / А. Н. Огурцов. – Харьков : НТУ «ХПИ», 2011. – 208 с. – На рус. яз.

ISBN 978-966-593-885-9

Учебное пособие содержит материалы по основным вопросам первого раздела курса «Биоинформатика и информационная биотехнология» в соответствии с программой подготовки студентов направления «Биотехнология».

Предназначено для студентов специальностей биотехнологического профиля всех форм обучения.

Ил. 40. Табл. 6. Библиогр.: 68 назв.

ББК 28.071.3

ISBN 978-966-593-885-9

Вступление

Предметом учебной дисциплины "Биоинформатика и информационная биотехнология" являются компьютерно-ориентированные методы решения информационных задач в области биотехнологии. Курс "Биоинформатика и информационная биотехнология" состоит из *четырёх разделов*: введение в биоинформатику, методы биоинформационного анализа, информационные принципы в биотехнологии, биоинформационные Интернет-ресурсы. Научную основу курса "Биоинформатика и информационная биотехнология" составляют молекулярная биофизика, молекулярная биология и генетика.

Методическими основами курса являются лекции, в которых излагаются основные положения каждого раздела, практические занятия и самостоятельная работа студентов, являющаяся основным способом усвоения материала в свободное от аудиторных занятий время.

Для самостоятельной работы выделяется больше половины общего объёма времени, предназначенного для изучения данной дисциплины. Самостоятельная работа проводится по всем темам, входящим в дисциплину. В процессе самостоятельной работы студент учится самостоятельно приобретать знания, которые затем используются в ходе выполнения индивидуального задания, практических занятий, при подготовке к выполнению контрольных работ и к экзамену.

Настоящее пособие подготовлено на основе адаптированных работ [1–25], послуживших также источником иллюстраций, таким образом, чтобы максимально облегчить усвоение раздела *"Введение в биоинформатику"* курса "Биоинформатика и информационная биотехнология" студентам направления подготовки 051401 "Биотехнология". Перед работой с пособием следует внимательно изучить материал пособий [18, 19], без которого невозможно понимание методов и алгоритмов, определяющих информационную составляющую биотехнологии. Словарь терминов и список условных обозначений приведены в конце пособия.

1. Предмет биоинформатики

1.1. Определение биоинформатики

Понятие "информация" проникает во все сферы деятельности человека, объединяя их в единый взаимосвязанный и взаимозависимый комплекс. Относительно недавно появился даже термин "инфосфера" – информационные структуры, системы и процессы в науке, обществе и производстве. Вместе с тем, до сих пор отсутствует единая точка зрения на *предмет* информатики, и до сих пор не вполне ясны соотношения между различными информационными дисциплинами, связанными с различными предметными областями.

Интуитивно ясно, что биоинформатика нацелена на использование информации и информационных технологий при исследовании биологических систем. В биоинформатике *биология*, *информатика и математика* сливаются в единую дисциплину. В каком-то смысле биоинформатика, изучающая применение информационных технологий для управления биологическими данными, является продолжением вычислительной биологии, изучающей применение методов количественного анализа в моделировании биологических систем.

Интенсивность исследования геномов различных организмов с каждым годом нарастает, ежегодно появляются новые базы данных, в которых хранится информация об исследованных геномах, а уже существующие базы данных непрерывно наращивают свои мощности. Следовательно, с такой же огромной скоростью растёт и объём доступной исследователям биологической информации. Без использования современных информационных технологий уже невозможно ни отыскать, ни обработать ту конкретную биологическую информацию, которая необходима в данном исследовании или в данном биотехнологическом процессе.

Триединая цель биоинформатики включает в себя

1) организацию и сохранение биологических данных;
2) разработку программных средств и создание специализированных информационных ресурсов;
3) автоматизацию анализа биологических данных, интерпретацию и использование полученных результатов.

Таким образом, ***биоинформатика*** – это *наука о хранении, извлечении, организации, анализе, интерпретации и использовании биологической информации.*

Современная биоинформатика возникла в конце семидесятых годов двадцатого века с появлением эффективных методов расшифровки нуклеотидных последовательностей ДНК. Датой выделения биоинформатики в отдельную научную область можно считать 1980 год, когда началось издание журнала Nucleic Acids Research, целиком посвящённого 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. Переход к обработке, анализу и сравнению полных геномов организмов был невозможен без использования компьютерных методов информационного анализа, в результате эти исследования вылились в самостоятельное научное направление. Геномы содержат огромное количество генов, многие из которых до настоящего времени не идентифицированы экспериментально.

Поскольку технологии чтения генетической информации невозможны без использования компьютерной техники и вычислительных

методов, то возникновение и интенсивное развитие биоинформатики происходило синхронно с возникновением и повсеместным распространением компьютерных технологий. Это является лишним подтверждением того факта, 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. Теперь нет необходимости разрабатывать программные продукты в каждой исследовательской лаборатории, поскольку большое число разнообразных баз данных и программных инструментов сегодня доступны через Интернет. Биоинформатика, пожалуй, является одной из тех областей науки, которые в очень большой степени зависимы от Интернета и успешно развиваются благодаря Интернету. Именно 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 всё более эффективных методов расшифровки биологических текстов и разработка методов биоинформатики позволяет надеяться на серьёзный прогресс в понимании строения, механизмов функционирования и регуляции живых систем.

В результате становится возможным изучение и понимание всё более сложных биологических систем, появляется возможность их системного исследования, установление эволюционных связей в живой природе, создание новых лекарственных препаратов, методов лечения и новых биотехнологий.

Биоинформатика – молодая наука. Её предметная область, её цели и объекты находятся в постоянном развитии. Поэтому сегодня ещё нет единого общепризнанного определения биоинформатики. Ниже приведены некоторые из таких определений.

- *Биоинформатика* – это технология применения компьютеров для решения информационных задач в области естественных наук; главным образом она занимается созданием обширной электронной базы данных последовательностей геномов и белков. Во вторую очередь биоинформатика развивает различные методики, например, пространственного моделирования биомолекул и биологических систем.
- *Биоинформатика* – это автоматизированное управление всеми видами биологической информации, включая гены и их продукты, целые организмы или даже экологические системы.
- *Биоинформатика* – это интеграция математических, статистических и вычислительных методов анализа биологических, биохимических и биофизических данных. Сюда входит разработка способов хранения, выборки и анализа биологических данных, например, последовательностей нуклеиновых кислот и белковых последовательностей, а также структур, функций, метаболических путей и моделей взаимодействия генов.
- *Биоинформатика* – это отрасль информатики, отвечающая за хранение и анализ биологической информации, а также за манипуляцию данными. Биоинформатика является фундаментальной инфраструктурой, на которой основаны все биологические исследования.

1.2. История становления биоинформатики

Ниже приведены исторические события, способствовавшие возникновению и развитию биоинформатики и информационной биотехнологии.

1866 г. – Грегор Мендель опубликовал результаты своих опытов над передачей наследственных "факторов" у растений гороха.
1917 г. – Карл Эреки ввел термин "биотехнология".
1928 г. – Эрвин Шрёдингер высказал предположение, что менделевский наследственный фактор имеет размеры около 1000 ангстрем.

1928 г. – Александр Флеминг обнаружил, что некоторые плесневые грибы могут остановить размножение бактерий, что впоследствии привело к открытию первого антибиотика – пенициллина.

1933 г. – Арне Тизелиус предложил метод электрофоретического разделения смеси белков в растворе.

1943 г. – Произведен пенициллин в промышленном масштабе.

1944 г. – Освальд Эвери, Колин МакЛеод и Маклин МакКарти (Oswald Avery, Colin MacLeod, Maclyn McCarty) показали, что генетический материал представляет собой ДНК.

1951 г. – Лайнус-Карл Полинг (Linus Carl Pauling) и Роберт Кори (Robert Brainard Corey) предложили модели структур, образуемых полипептидной цепью белка: α-спирали и β-листа.

1952 г. – Розалинда Франклин (Rosalind Franklin) и Морис Уилкинс (Maurice Wilkins) с помощью рентгеноструктурного анализа обнаружили регулярный характер структуры ДНК.

1953 г. – Джеймс Уотсон (James Dewey Watson) и Френсис Крик (Francis Crick) предложили модель двойной спирали ДНК.

1954 г. – Макс Перуц (Max Ferdinand Perutz) и возглавляемая им группа ученых разработали методы изоморфного замещения тяжелыми атомами, позволившие решить проблему фаз в кристаллографии белка.

1955 г. – Фредерик Сангер (Frederick Sanger) расшифровал последовательность бычьего инсулина.

1957 г. – Артур Корнберг создал первую синтетическую молекулу ДНК.

1961 г. – Учрежден журнал "Biotechnology and Bioengineering".

1965 г. – Маргарет Дейхофф вместе с сотрудниками "Национального фонда биомедицинских исследований" (NBRF), Вашингтон, впервые собрали воедино базы данных белковых последовательностей.

1966 г. – Расшифрован генетический код.

1968 г. – Вернер Арбер, Гамильтон Смит и Дэниел Нат описали принцип действия рестриктаз.

1969 г. – Объединение компьютеров Станфордского университета и Калифорнийского университета в Лос-Анджелесе привело к созданию сети ARPAnet.

1970 г. – Опубликовано подробное описание алгоритма Нидлмена-Вунша для сравнения последовательностей.

– А.Дж. Гиббс и Г.А. Макинтайр описали новый метод сравнения двух последовательностей (аминокислот или нуклеотидов) с помощью точечной матрицы.

1972 г. – Пауль Берг, применив лигазу, сконструировал первую искусственную молекулу рекомбинантной ДНК.

– Станли Коэн, Энни Чан и Герберт Бойер произвели первый организм с рекомбинантной ДНК.

1973 г. – Джозеф Сэмбрук со своей рабочей группой усовершенствовали метод электрофореза ДНК за счет применения агарозного геля.

– Герберт Бойер и Станли Коэн разработали технологию рекомбинантных ДНК, перенеся человеческий ген в плазмиду кишечной палочки (*Escherichia coli*).

– Создан "Брукхейвенский банк данных белка".

– Роберт Меткалф в своей докторской диссертации описал сеть Ethernet.

1974 г. – Винт Карф и Роберт Кан развили концепцию объединения компьютерных сетей в глобальную сеть "Интернет" и разработали протокол управления передачей TCP – (Transmission Control Protocol).

1975 г. – П.Х. О'Фаррелл изобрел метод двумерного электрофореза в полиакриламидном геле с добавлением додецилсульфата натрия.

– Эдвард Саузерн опубликовал описание разработанной им аналитической методики Саузерн-блотинг.

– Георг Кёлер и Сезар Мильштейн разработали метод производства моноклональных антител.

– Билл Гейтс и Пол Аллен основали корпорацию "Майкрософт" (Microsoft Corporation).

1976 г. – Изданы первые руководства, регламентирующие работы с рекомбинантными ДНК.

1977 г. – Фредерик Сангер и независимо Аллен Максам и Уолтер Гилберт разработали методы секвенирования ДНК

1978 г. – Фирма Genentech выпустила человеческий инсулин, полученный с помощью *Escherichia coli*.

1979 г. – В "Лос-Аламосской национальной лаборатории" (Los-Alamos National Laboratory, LANL), штат Нью-Мексико, Уолтер Гоуд с сотрудниками впервые объединили базы данных последовательностей ДНК в прототип базы данных GenBank.

1980 г. – Марк Сколник, Рей Уайт, Дэвид Ботштейн и Рональд Дейвис создали RFLP-маркерную карту генома человека.

– Впервые расшифрована полная последовательность генов организма – бактериофага "FX-174".

– Вутрих с сотрудниками опубликовал статью с подробным описанием применения метода многомерного ЯМР для определения структуры белка.

– Основана корпорация IntelliGenetics Inc. в Калифорнии. Её первым продуктом был комплект программ для анализа последовательностей ДНК и белков IntelliGenetics Suite.

– Опубликован алгоритм Смита-Уотермена для выравнивания последовательностей.

– Верховный суд США, слушая дело *Даймонд против Чакрабарти*, вынес вердикт, что микроорганизмы, полученные генно-инженерными методами, могут быть запатентованы.

– Учреждён журнал "Nucleic Acids Research"

1981 г. – Корпорация IBM выпустила на рынок первый персональный компьютер.

– Секвенирована митохондриальная ДНК человека – 16569 пар оснований.

– Д. Бенсон, Д. Липмен с сотрудниками разработали GENINFO – управляемую с помощью меню программу доступа к базе данных последовательностей.

– Поступили в продажу первые автоматические синтезаторы ДНК.

– Разрешен к применению в США первый диагностический набор моноклональных антител.

– Майзель и Ленк разработали различные схемы фильтрации и цветного отображения, которые значительно повысили удобство применения метода точечных матриц.

1982 г. – Разрешена к применению в Европе первая вакцина для животных, полученная по технологии рекомбинантных ДНК.

– В Университете штата Висконси" при Центре биотехнологий в Висконсин" открыт информационный отдел Genetics Computer Group, GCG.

1983 г. – В продаже появился лазерный компакт-диск (CD).

– Для трансформации растений применены гибридные Ti-плазмиды.

1984 г. – В сети Интернет размещена система имен доменов DNS – (Domain Name System) Джона Постела.

– Корпорация Apple Computer выпустила на рынок компьютер Macintosh.

– Секвенирован геном вируса Эпштейна-Бара: 172281 пар оснований

1985 г. – Кэри Муллис изобрел полимеразную цепную реакцию (ПЦР).

– Опубликован алгоритм FASTP.

– Роберт Синшеймер внес первое предложение о разработке проекта "Геном человека".

1986 г. – Томас Родерик ввел термин "геномика" для обозначения научной дисциплины, рассматривающей вопросы картографирования, секвенирования и анализа генов.

– Корпорация Amoco Technology Corporation приобрела Intelli-Genetics Suite.

– Отделом медицинской биохимии Женевского университета совместно с Европейской лабораторией молекулярной биологии (EMBL) была создана база данных Swiss-Prot.

1986 г. – Лерой Худ и Ллойд Смит автоматизировали процесс секвенирования ДНК.
– Шарль Дели зи созвал заседание с целью обсудить возможности определения нуклеотидной последовательности генома человека.

1987 г. – Министерство охраны окружающей среды США официально объявило о запуске проекта "Геном человека".
– Й. Кохара с сотрудниками опубликовал физическую карту генома кишечной палочки (*Escherichia coli*).

1988 г. – Дэвид Т. Бёрк с сотрудниками описал методику применения дрожжевой искусственной хромосомы (YAC).
– Пирсон и Липмен опубликовали алгоритм FASTA.
– При Национальном институте рака (США) организован Национальный центр биотехнологической информации (NCBI).

1989 г. – Национальный институт здоровья (США) учредил Национальный центр исследования генома человека (NHGRI).
– Информационный центр Genetics Computer Group стал частной компанией.
– Компания Oxford Molecular Group Ltd. (OMG), Оксфорд, выпустила программные продукты: Anaconda, Asp и Cameleon, а также программы для молекулярного моделирования, разработки лекарственных препаратов и конструирования белковых молекул.

1990 г. – Стефен Альтшуль (Stephen Frank Altschul) с группой программистов написали программу BLAST для автоматического выравнивания последовательностей ДНК.
– Майкл Левитт и Крис Ли основали компанию Molecular Applications Group в Калифорнии.
– В г. Вифезда, штат Мэриленд, учреждена компания InforMax.
– В США утвержден план испытаний генной терапии с использованием соматических клеток человека.

1991 г. – "ЦЕРН" (CERN), Женева, объявил о создании протоколов, положивших начало "Всемирной паутине" (World Wide Web).

– Крейг Вентер изобрел технологию опознавания генов с помощью EST (Expressed sequence tags – ярлыков экспрессируемых последовательностей (ЯЭПов)) – фрагментов кДНК, комплементарных матричным РНК.
– В Калифорнии создана компания Incyte Pharmaceuticals, занимающаяся развитием фармацевтической геномики.
– В Юте (США) основана компания Myriad Genetics Inc., с целью определить гены основных заболеваний и раскрыть механизмы их наследования.

1992 г. – Уильям Хазелтин открыл компанию Human Genome Systems в штате Мэриленд.
– Крейг Вентер учредил Институт геномных исследований TIGR (The Institute for Genomic Research) с целью коммерческого использования секвенирования путём идентификации генов и разработки лекарств.
– Мэл Саймон с сотрудниками Cal Tech изобрели бактериальную искусственную хромосому (BAC) – ключевой элемент в сборке гена из клонов.
– В проект "Геном человека" вошла компания Welcome Trust.
– В рамках проекта "Геном человека" завершена карта сцепления человеческого генома с высоким разрешением.
– Начало проекта по секвенированию *Caenorhabditis elegans*.

1993 г. – Френсис Коллинс принял на себя руководство проектом "Геном человека". В Великобритании был открыт "Сенгеровский центр" (The Sanger Centre). К проекту присоединились некоторые другие страны. Завершение работы над проектом запланировано на 2005 год.
– В Нью-Хейвене, штат Коннектикут, появилась корпорация CuraGen Corporation.

1994 г. – Основана корпорация Netscape Communications Corporation, выпустившая на рынок Интернет-браузер Navigator.
– Эттвуд и Бек создали базу данных белковых мотивов PRINTS.
– В штате Мэриленд образована компания Gene Logic.

1995 г. – Ученые TIGR впервые расшифровали последовательность генома свободно живущего организма *Haemophilus influenzae*.
– Патрик Браун с сотрудниками Стэндфордского университета изобрел технологию создания и применения микроматриц ДНК.

1996 г. – Определена нуклеотидная последовательность всех хромосом (секвенирован геном) эукариотического микроорганизма – пекарских дрожжей *Saccharomyces cerevisiae*.
– Получена карта человеческого генома с высоким разрешением – маркеры, разделённые фрагментами длиной в 600000 пар оснований.
– Международный консорциум по проекту "Геном человека" сформировал "Бермудские правила" публикации научных данных.
– Барух с сотрудниками сообщили о создании базы данных PROSITE.
– Компания Affymetrix выпустила первые коммерческие чипы ДНК.
– Ежегодный объём продаж первого рекомбинантного белка (эритропоэтина) превысил 1 млрд. долларов.

1997 г. – Опубликован геном *Escherichia coli.*
– Компания Oxford Molecular Group приобрела GCG.
– Клонировано млекопитающее из дифференцированной соматической клетки (овца по кличке "Долли").
– Появилась компания LION bioscience AG.

1998 г. – Опубликованы геномы нематоды *Caenorhabditis elegans* и дрожжей *Saccharomyces cerevisiae*.
– Крейг Вентер основал компанию Celera Genomics в штате Мэриленд.
– Университетский колледж Лондона учредил компанию Inphamatica – новую компанию по развитию геномики и биоинформатики.

– В Сан-Диего была образована компания Gene Formatics, с целью проводить анализ и предсказание структур и функций белков.
– Создано некоммерческое научно-исследовательское учреждение Швейцарский институт биоинформатики.
– NIH (Национальный институт здоровья) начал проект "SNP" с целью выявления изменений, происходящих в геноме человека.

1999 г – Компания Welcome Trust сформировала консорциум для развития проекта "SNP".
– Опубликована полная последовательность одной из хромосом человека.

2000 г. – Секвенированы геномы *Pseudonomas aeruginosa*, *Arabidopsis thaliana* и *Drosophila melanogaster*.
– Компания Pharmacopeia приобрела Oxford Molecular Group.

2001 г. – Журналы "Science" и "Nature" опубликовали аннотации к геному человека и результаты его анализа.

2002 г. – Установлена структура ДНК риса – первой сельскохозяйственной культуры, геном которой был расшифрован.
– Запущен геномный браузер UCSC (University of California Santa Cruz) http://www.genome.ucsc.edu/ как совместный проект European Bioinformatics Institute и Welcome Trust Sanger Institute.

2003 г. – На рынке появилось первое генно-модифицированное (ГМ) домашнее животное "GloFish" – рыба, специально выведенная для обнаружения загрязнения воды, которая светится красным светом благодаря добавлению гена биолюминесценции.
– В NCBI создана программа ассемблирования ДНК из фрагментов – http://www.ncbi.nlm.nih.gov/projects/genome/assembly/

2004 г. – Использована пересадка мультипотентных стволовых клеток из пуповинной крови для лечения травмы спинного мозга.
– Разработан метод получения красных кровяных телец из стволовых гемопоэтических клеток, и создана среда, которая имитирует условия костного мозга.

2004 г. – Создана программа ENSEMBL распознавания генов в секвенированном геноме – http://www.ensembl.org/

2005 г. – В университете Висконсин-Мэдисон разделили бластоцисты стволовых клеток человека на нервные стволовые клетки и спинные двигательные нейронные клетки.

2006 г. – Создана первая карта метилирования генома *Arabidopsis.*

1.3. Особенность биоинформационных данных

Биологию традиционно считают *описательной*, а не *аналитической* наукой. Несмотря на то, что последние успехи науки не изменили это основное направление, радикально изменилась сущность биологических данных. До последнего времени все биологические наблюдения носили в основном *случайный* характер, правда, с различным уровнем точности и некоторые были проведены 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 никогда не могут 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 и зависима от цепи исторических случайностей, чтобы сегодня можно было бы детально объяснить её свойства, исходя из фундаментальных принципов.

Вторая очевидная особенность биоинформационных данных – это их огромное количество. Сейчас банки данных нуклеотидных последовательностей содержат 16 млрд. нуклеиновых пар оснований. Если мы возьмем в качестве единицы измерения размер генома человека

(HUman Genome Equivalents, HUGE), то этот объём информации эквивалентен 5 HUGE. База данных только белковых структур содержит ~66000 записей, каждая из которых является полным описанием координат ~400 аминокислотных остатков данного белка в трёхмерном пространстве. Огромны не только размеры отдельных банков данных, но и экспоненциальные темпы их увеличения.

Такое количество и качество биологических данных стимулирует исследователей к достижению следующих целей:

- *Увидеть* картину мира живых существ *четко* и *целиком*, т. е. понять *интегрирующие* аспекты биологии организмов, рассматриваемых как согласованные комплексные системы.
- *Связать* между собой последовательность, трёхмерную 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 последовательность событий в эволюционной истории,
  - *вперед* к научно обоснованной модификации биологических систем.
- *Способствовать применению* этих знаний в медицине, сельском хозяйстве и других областях.

Молекула ДНК состоит из тысяч нуклеотидов, и поэтому определение полной последовательности нуклеотидов целой молекулы хромосомной ДНК представляет собой весьма сложную задачу. С появлением технологии клонирования генов и полимеразной цепной реакции (ПЦР) ученые получили возможность выделять отдельные фрагменты хромосомной ДНК. Эти достижения, в свою очередь, проложили путь к развитию быстрых и эффективных методов *секвенирования* ДНК.

В конце 70-х годов XX века появились два метода секвенирования, основанные, соответственно, на реакциях обрыва цепи и химического расщепления.

Эти методы с некоторыми незначительными видоизменениями заложили основу для революции секвенирования 80-х и 90-х годов и последующего рождения биоинформатики.

Благодаря своей чувствительности, специфичности и возможности автоматизации, ПЦР считается передовым методом анализа образцов геномной ДНК и построения генетических карт. Последующие усовершенствования базовой технологии ПЦР дополнительно увеличили мощность и практическую ценность этой методики.

С момента получения в 1987 году первой последовательности, секвенированной полуавтоматическим методом, практической реализации ПЦР в 1990 г. и внедрения способа флуоресцентного мечения фрагментов ДНК, производимых методом полимерного копирования по Сангеру, были осуществлены попытки крупномасштабного секвенирования, также внесшие неоценимый вклад в развитие биоинформатики. Одновременно значительное развитие получили технологии автоматизированной регистрации результатов секвенирования последовательностей.

Ещё в начале 80-х годов XX века исследователи вручную (с помощью электронных перьев) считывали последовательности ДНК с картины полос на гелиевой пленке. Затем появились устройства записи изображения, а именно камеры, которые оцифровывали оптическую информацию, полученную в ходе гель-электрофореза. В 1987 году Стивен Кравец помог разработать первое программное обеспечение для устройств автоматического считывания информации с гелиевых пленок.

В начале 90-х годов Крейг Вентер с сотрудниками изобрел новый метод определения генов. Вместо того чтобы секвенировать хромосомную ДНК с предельным разрешением в один нуклеотид, группа Вентера выделила молекулы информационной РНК, копировала их в молекулы кДНК и затем секвенировала некоторую часть молекулы кДНК, в результате чего были созданы ярлыки экспрессируемых последовательностей EST (*Expressed sequence tags*). Эти EST-последовательности могли быть использованы в качестве указателей для выделения целого гена.

Кроме того, подход с применением ярлыков EST повлек за собой организацию огромных баз данных нуклеотидных последовательностей и,

как полагают, развитие метода EST показало осуществимость проектов высокопроизводительного обнаружения новых генов и явилось ключевым толчком для развития прикладной геномики.

Хранение биоинформационных данных. К началу 1998 года в общедоступные безызбыточные базы данных было помещено уже более 300 000 белковых последовательностей, а число частично расшифрованных последовательностей в общественных и корпоративных базах данных EST оценивалось миллионами. Сегодня число пространственных структур в "Банке белковых данных" (Protein data bank, PDB) превышает 72 000 (рисунок 1) – http://www.pdb.org/

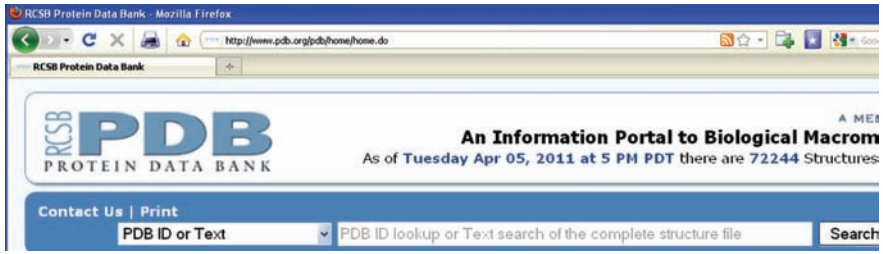


Рисунок 1 – Веб-страница Банка белковых данных PDB

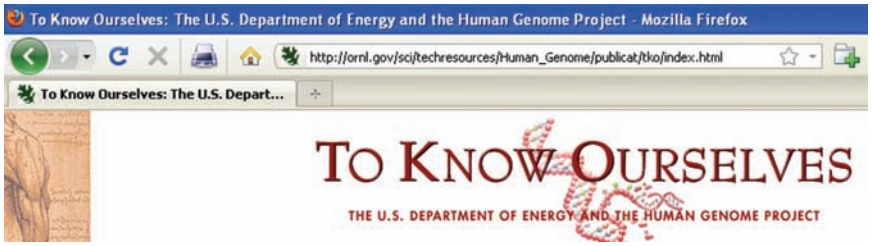
Министерство энергетики США в 80-х годах XX века запустило ряд проектов по созданию подробных генетических и физических карт генома человека (рисунок 2).

Их цель состояла в расшифровке полной последовательности нуклеотидов *генома человека* и в определении *локусов* (фиксированных положений, локализации на хромосоме) предполагаемых 30 000 генов.

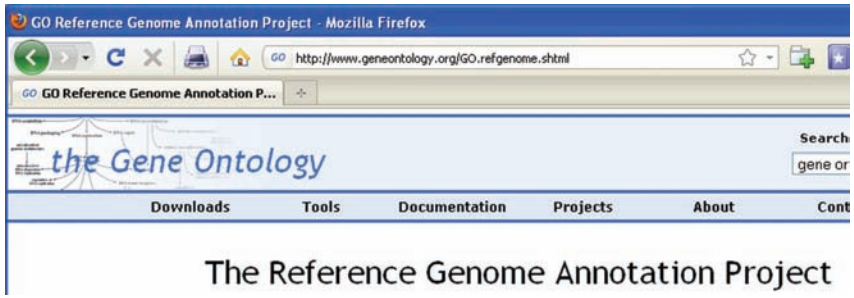
Работа столь большого размаха стимулировала развитие новых вычислительных методов анализа генетических карт и данных секвенирования последовательностей ДНК, а также потребовала разработки новых методов и лабораторного оборудования для расшифровки и анализа ДНК.

Для максимально быстрого ознакомления широкого круга исследователей с результатами расшифровки потребовалось разработать усовершенствованные средства распространения полученной информации.

а



б



в

г

Рисунок 2 – Веб-страницы геномных проектов: *а* – Геномной программы Департамента Энергии США; *б* – To Know Ourselves; *в* – Проект аннотации геномов; *г* – Национальный институт исследования генома человека

Международную научно-исследовательскую программу, явившуюся результатом этой глобальной инициативы, назвали проектом "Геном человека" (*Human Genome Project*, HGP). Более подробно о расшифровке геномов можно узнать на следующих Веб-узлах (рисунок 2):

- http://genomics.energy.gov/
- http://ornl.gov/sci/techresources/Human_Genome/publicat/tko/index.html
- http://www.geneontology.org/GO.refgenome.shtml
- http://www.genome.gov/

1.4. Цели и задачи биоинформатики

Основополагающий принцип биоинформатики состоит в том, что биополимеры, например, молекулы нуклеиновых кислот и белков, могут быть изображены в виде последовательности цифровых символов. Кроме того, для представления мономеров аминокислотных и нуклеотидных цепей необходимо лишь ограниченное число алфавитных знаков.

Подобная гибкость анализа биомолекул с помощью ограниченных алфавитов привела к успешному становлению биоинформатики. Развитие и функциональная мощь биоинформатики во многом зависят от прогресса в области разработки компьютерных аппаратных средств и программного обеспечения. Простейшие задачи, стоящие перед биоинформатикой, касаются создания и ведения баз данных биологической информации.

Предмет биоинформатики включает в себя три компонента:

1) создание баз данных, позволяющих осуществлять хранение крупных наборов биологических данных и управление ими;
2) разработка алгоритмов и методов статистического анализа для определения отношений между элементами крупных наборов данных;
3) использование этих средств для анализа и интерпретации биологических данных различного типа – в частности, последовательностей ДНК, РНК и белков, белковых структур, профилей экспрессии генов и биохимических путей.

Цели биоинформатики следующие:

1. Организовывать данные таким образом, чтобы исследователи имели доступ к текущей информации, хранящейся в базах данных, и могли вносить в нее новые записи по мере получения новых сведений.
2. Развивать программные средства и информационные ресурсы, которые помогают в управлении данными и в их анализе.
3. Применять эти средства для анализа данных и интерпретации полученных результатов таким образом, чтобы они имели биологический смысл.

Задачи биоинформатики состоят в анализе информации, закодированной в биологических последовательностях, в частности:

- обнаруживать гены в последовательностях ДНК различных организмов;
- развивать методы изучения структуры и (или) функции новых расшифрованных последовательностей и соответствующих структурных областей РНК;
- определять семейства родственных последовательностей и строить модели;
- выравнивать подобные последовательности и восстанавливать филогенетические деревья с целью выявления эволюционных связей.

Помимо перечисленных выше задач, следует упомянуть *ещё один важнейший* вопрос биоинформатики – *обнаружение мишеней* для медикаментозного воздействия и отыскание перспективных опытных соединений.

Предмет биоинформатики реализуется в следующих ***видах деятельности***.

1. Управление биологическими данными и их обработка; сюда входит их организация, отслеживание, защита, анализ и т. д.
2. Организация связи между учеными, проектами и учреждениями, вовлеченными в фундаментальные и прикладные биологические

исследования. Связь может включать в себя электронную почту, пересылку файлов, дистанционный вход в систему, телеконференции, электронные информационные табло и, наконец, учреждение сетевых информационных ресурсов.
3. Организация наборов биологической информации, документов и литературы, а также обеспечение доступа к ним, их поиска и выборки.
4. Анализ и интерпретация биологических данных с применением вычислительных методов, как-то: визуализация, математическое моделирование, а также построение алгоритмов высокопараллельной обработки сложных биологических структур.

1.5. Применение биоинформатики

Помимо обеспечения исследователей, изучающих белки и ДНК, теоретической базой и вычислительно-аналитическим аппаратом, вычислительная биология нашла применение во многих областях.

В расшифровке смыслового содержания биологических последовательностей наметились *два* различных аналитических направления:

- согласно *первому* подходу, ученые опираются на методы распознавания регулярных комбинаций, посредством которых обнаруживают подобие последовательностей и, следовательно, выявляют эволюционно связанные структуры и функции;
- согласно *второму* подходу, используют методы предсказания *ab initio* (с самого начала, из первых принципов) – для прогнозирования трёхмерных структур и, в конечном счете, выведения функции непосредственно по линейной последовательности. Прямое предсказание трёхмерной структуры белка по его линейной последовательности аминокислот – важнейшая цель биоинформатики.

Анализ гомологичности последовательностей. Одна из движущих сил биоинформатики – поиск подобий между различными биомолекулами. Помимо систематической организации данных, идентификация

белковых гомологов имеет прямое практическое применение. Теоретические модели белков обычно основаны на структурах близких гомологов, определённых опытным путём.

Всякий раз, когда ощущается недостаток биохимических или структурных данных, могут быть выполнены исследования на дрожжеподобных низших организмах, а результаты могут быть распространены на гомологичные молекулы более высоких организмов, например, человека.

Более того, данный подход упрощает проблему понимания сложных геномов – за счет непосредственного анализа простых организмов и последующего распространения тех же самых принципов на более сложные организмы. Это могло бы привести к опознаванию потенциальных мишеней для медикаментозного воздействия путём испытаний на гомологах основных микробных белков.

Разработка лекарственных препаратов. Опирающийся на биоинформатику подход к открытию лекарств даёт важное преимущество. С помощью биоинформатики могут быть описаны генотипы, сопряжённые с патофизиологическими состояниями, что в принципе позволит опознать соответствующие молекулярные мишени. Посредством программного транслятора по известной последовательности нуклеотидов может быть определена вероятная аминокислотная последовательность кодируемого белка.

В случае принятия такого подхода методы изучения последовательностей могли бы применяться для поиска гомологов у опытных организмов. На основании подобия последовательностей было бы возможно моделировать структуру конкретного белка, взяв за основу экспериментально установленные структуры. И, наконец, компьютерные алгоритмы докинга могли бы проектировать молекулы, потенциально связывающиеся с моделируемой структурой, отбирая наиболее перспективные варианты для биохимических испытаний, проверяющих биологическую активность этих молекул уже на реальном белке.

Гипотетический пример. Чтобы нагляднее понять роль вычислений в молекулярной биологии, вообразим себе в будущем пандемическую

ситуацию, вызванную появлением нового биологического вируса. Этот вирус вызывает эпидемию смертельного заболевания, как среди людей, так и среди животных. Ученые в лаборатории выделят его генетический материал – молекулу ДНК, и определят её последовательность.

Затем с помощью компьютерного скрининга этого нового генома по базам данных всего известного на тот момент генетического материала возможно будет охарактеризовать вирус и выявить его родство с ранее изученными вирусами. 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. Из последовательности ДНК вируса компьютерные программы вычислят аминокислотные последовательности одного или нескольких вирусных белков, критически важных для репликации или сборки вируса.

Из аминокислотных последовательностей другие программы вычислят структуры этих белков, следуя тому базовому принципу, что аминокислотная последовательность белка однозначно определяет его трёхмерную структуру, а, следовательно, 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 предсказания действия изменений в последовательности на структуру молекулы.

Структуры мишеней будут предсказаны с помощью метода, известного как гомологичное моделирование. 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). Последняя ситуация будет возникать все реже, по мере того, как растёт 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), определяющие функции этих белков, которые чувствительны к блокированию. Будет найдена или сделана маленькая молекула, комплементарная

такому участку (сайту) по структуре и свойствам, которая будет работать как антивирусный препарат. Альтернативный вариант – создать и синтезировать одно или несколько антител для нейтрализации вируса.

Такая последовательность событий в гипотетической ситуации основана на уже сегодня четко установленных принципах.

Многие проблемы на каждом из описанных этапов ещё не решены, и это одна из причин, по которой этот сценарий не может быть использован уже сегодня, например, для создания лекарственных препаратов против СПИДа. Другая причина в том, что вирусы знают, как себя защитить.

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

Наиболее вероятным будет параллельное совершенствование и взаимное дополнение экспериментальных и биоинформационных методов разработки лекарственных препаратов.

Моделирование. Благодаря информационным технологиям массового просмотра и сравнения данных можно получить ответ на ряд вопросов, касающихся эволюционных, биохимических и биофизических характеристик исследуемых биомолекул. Стало возможным установить:

а) специфические мотивы укладки белка, соответствующие определённым филогенетическим группам;
б) общность между различными вариантами укладки белковых глобул, наблюдаемыми у отдельных организмов;
в) долю аналогичных третичных структур, общих для родственных организмов;
г) степень родства, выведенного из тривиальных эволюционных деревьев;
д) различие метаболических путей у разных организмов.

Кроме того, на основании того факта, что особенности укладки белковой глобулы часто связаны с определёнными биохимическими функциями, можно получать данные относительно функций белка. Анализируя

информацию об экспрессии генов, одновременно со структурной и функциональной классификацией белков можно предсказать появление определённой свертки белка в геноме, что характерно для высоких уровней экспрессии. На основании анализа структурных данных можно составить карту взаимодействий всех белков того или иного организма.

Медицина. Медицинские приложения биоинформатики связаны, главным образом, с анализом экспрессии генов. Как правило, регистрируют данные об экспрессии в клетках, пораженных различными заболеваниями, и затем сравнивают эти измерения с нормальными уровнями экспрессии. Те гены, которые демонстрируют изменения в экспрессии в пораженных клетках, вероятнее всего и связаны с данным заболеванием. Это позволяет выяснить причину болезни и указывает потенциальные мишени для лекарственных препаратов.

Располагая подобной информацией, можно разрабатывать соединения, которые связываются с экспрессируемым белком. Далее могут быть проведены эксперименты на микроматрицах, чтобы оценить реакцию на фармакологическое воздействие полученного опытного соединения. Подобный подход может помочь также при разработке тестов для обнаружения или прогноза токсичности опытных лекарств на стадии клинических испытаний.

Объединение биоинформатики с экспериментальной геномикой позволит *решить* целый ряд *актуальных задач*, например, послеродовое определение генотипа для оценки восприимчивости или устойчивости индивидуума к определённым болезням и патогенам; индивидуальное предписание уникального сочетания вакцин; уменьшение затрат на лечение за счет повышения эффективности терапии и предупреждения рецидивов заболевания. Все вместе эти новшества могут привести к разработке индивидуальных пищевых рационов и выявлению заболеваний на ранних стадиях.

Кроме того, программы медикаментозного лечения могли бы *индивидуально* подбираться для конкретного пациента и его заболевания, и, таким образом, обеспечивать наиболее эффективный курс лечения с минимальными побочными эффектами.

В частности, проект "Геном человека" принесет несомненную пользу судебной медицине и фармацевтической промышленности, приведет к открытию многих "полезных" и "вредных" генов, внесет неоценимый вклад в развитие представлений об эволюции человека. Кроме того, он будет способствовать разработке методов диагностики болезней, возможных осложнений, предсказанию генетически обусловленных реакций на терапевтическое воздействие, а также будет способствовать развитию индивидуальных подходов к лечению, методов обнаружения мишеней для лекарственных препаратов и, наконец, становлению генотерапии.

Права на интеллектуальную собственность. Права на интеллектуальную собственность – это неотъемлемая часть современных деловых отношений. Под правами на интеллектуальную собственность понимают средства защиты любых нематериальных активов. Примеры интеллектуальной собственности: патент, авторское право, торговая марка и коммерческая тайна. Патент – это исключительная монополия, предоставляемая правительством изобретателю на пользование его изобретением в течение ограниченного периода времени.

Главные области биоинформатики, которые нуждаются в защите интеллектуальной собственности, следующие:

а) средства управления информацией и её анализа (например, методы моделирования, базы данных, алгоритмы, программное обеспечение и т. д.);
б) геномика и протеомика;
в) открытие (разработка) лекарственных препаратов.

Львиная доля новых разработок в биоинформатике относится к применению программного обеспечения (в том числе протоколов), предназначенного для сбора и (или) обработки биологических данных. Эти изобретения подпадают под общую категорию изобретений в области компьютерных наук и подразделяются на:

а) изобретения, реализованные на компьютерах;

б) изобретения, использующие машиночитаемые носители информации.

Все эти изобретения имеют две составляющие:

а) программное обеспечение;
б) аппаратные средства ЭВМ.

Например, основанная на критерии подобия автоматизированная система распознавания новых групп последовательностей нуклеотидов в заданном наборе нуклеотидных последовательностей может включать в себя устройство ввода, память и процессор (в качестве аппаратных компонентов системы), а также набор данных или метод использования команд, хранимых в памяти и выполняемых процессором, – как программное обеспечение системы. Патентная охрана была бы неоценима в защите методов, использующих вычислительные возможности, таких как методы выравнивания последовательностей, поиска гомологии и моделирования метаболических путей.

Геномика осуществляет выделение и описание генов, и приписывание последовательностям этих генов некоторых функций или назначений (например, экспрессии специфического белка или обозначения этого гена в качестве маркера определённой болезни). Эта работа предполагает проведение большого числа лабораторных испытаний и применения разнообразных вычислительных методов. Эти методы также могут быть защищены правами на интеллектуальную собственность.

Протеомика занимается очисткой и описанием белков, используя технологии типа двумерного гель-электрофореза, многомерной хроматографии и масс-спектроскопии. Применение этих методов к определению свойств и обнаружению связи белка, то есть маркера, со специфической болезнью, является весьма сложным и трудоемким процессом и требует значительных инвестиций.

Методы *разработки лекарственных средств* с применением автоматического моделирования, которое предполагает использование ком-

пьютеров и вычислительных алгоритмов, также могут быть отнесены к интеллектуальной собственности.

Контрольные вопросы

1. Дайте определение биоинформатики.
2. Какую дату можно считать датой выделения биоинформатики в отдельную научную область?
3. Почему до настоящего времени нет общепринятого определения биоинформатики?
4. Назовите основные исторические события, способствовавшие возникновению и развитию биоинформатики.
5. В чём состоит особенность биоинформационных данных?
6. Что такое секвенирование и какую роль играет секвенирование в биоинформатике?
7. Где хранятся биоинформационные данные?
8. Какие три компонента включает в себя предмет биоинформатики?
9. Каковы цели биоинформатики?
10. Какие задачи стоят перед биоинформатикой?
11. В каких видах деятельности реализуется предмет биоинформатики?
12. Какие два различных аналитических направления существуют в расшифровке биологических последовательностей?
13. Какую роль играет анализ гомологических последовательностей в расшифровке биологической информации?
14. Каким образом биоинформатика способствует разработке лекарственных препаратов?
15. Перечислите медицинские применения биоинформатики.
16. Какие области биоинформатики нуждаются в защите интеллектуальной собственности?
17. Какие методы геномики и протеомики нуждаются в патентной защите, и почему?

2. Понятие "информация"

Слово "информация" образовано от латинского "*informatio*" – разъяснение, изложение, ознакомление, представление. Это одно из наиболее общих понятий науки, обозначающее некоторые сведения, совокупность каких-либо данных, знаний и т. п.

В последнее время выяснилось, что информация играет в науке фундаментальную роль. Возникла потребность понять, что же это такое? Попытки связать информацию с привычными понятиями материи или энергии успехом не увенчались. Норберт Винер (1894-1964) – основоположник кибернетики и теории искусственного интеллекта – утверждал, что "информация есть информация, а не материя и не энергия", подчёркивая невещественность происхождения информации.

Попытки связать информацию с энтропией тоже оказались безуспешными, хотя они продолжаются до сих пор. Поэтому вопрос об определении понятия "информация" остаётся открытым.

2.1. Определение понятия "информация"

В гуманитарных науках популярны определения типа "информация есть сведения..." (или сообщения, знания и т. п.). Например:

- "информация есть знания, переданные кем-то другим или приобретенные путём собственного исследования или изучения";
- "информация – это сведения, содержащиеся в данном сообщении и рассматриваемые как объект передачи, хранения и обработки".

При этом фактически одно слово заменяется другим, и такие переименования по существу являются тавтологией.

Иногда информацию связывают с упорядоченностью, например:

- "информация означает порядок, коммуникация есть создание порядка из беспорядка или, по крайней мере, увеличение степени той

упорядоченности, которая существовала до получения сообщения".

Среди философов популярны определения, содержащие термин "отражение", что тоже не отличается продуктивностью:

- "информация есть отражение в сознании людей объективных причинно-следственных связей в окружающем нас реальном мире";
- "информация – это содержание процессов отражения".

Особое место в коллекции определений занимают утверждения о том, что информация – это алгоритм или инструкция, например:

- "информация есть некий алгоритм"; "совокупность приемов, правил или сведений, необходимых для построения оператора, будем называть информацией".

Одновременное существование большого числа похожих и не похожих друг на друга определений понятия "информация" означает, что общепринятого определения ещё нет. Более того, нет даже четкого понимания сути этого явления, хотя потребность в нем уже назрела.

Наиболее адекватным для биологических применений является следующее определение понятия "информация"

- *информация есть запомненный выбор одного варианта из нескольких возможных и равноправных*.

Запомненный выбор ещё называют *макроинформацией*. Если информация создаётся, но не запоминается, то её называют *микроинформацией*.

Далее под информацией мы будем понимать только запоминаемую информацию и приставку "макро" опустим.

Слова "возможных и равноправных" в определении означают, что варианты выбора принадлежат одному множеству и априорные различия между ними невелики. В идеале варианты могут быть полностью равно-

правны и равновероятны, но могут и отличаться. В этом случае слово "равноправные" означает, что априорные вероятности различных выборов – величины одного порядка.

Приведём ещё одну цитату Н. Винера: "Информация – это обозначение содержания, полученного из внешнего мира в процессе нашего приспособления к нему и приспосабливания к нему наших чувств. Процесс получения и использования информации является процессом нашего приспособления к случайностям внешней среды и нашей жизнедеятельности в этой среде. Потребности и сложности современной жизни предъявляют гораздо больше, чем когда-либо раньше, требования к этому процессу информации, и наша пресса, наши музеи, научные лаборатории, университеты, библиотеки и учебники должны удовлетворить потребности этого процесса, так как в противном случае они не выполнят своего назначения. Действенно жить – это значит жить, располагая правильной информацией. Таким образом, сообщение и управление точно так же связаны с самой сущностью человеческого существования, как и с жизнью человека в обществе."

Характерно, что Н. Винер относит понятие "информация" к категории процессов, что означает критическую важность того, каким именно способом была получена данная информация.

2.2. Количество информации

В 1948 г. сотрудник американской компании Bell Telephone Laboratories Клод Элвуд Шеннон (ему тогда было 28 лет) опубликовал в журнале "Bell System Technical Journal" фундаментальную работу "Математическая теория связи". С её появлением обычно связывают возникновение классической (статистической) теории информации. Именно к этому времени развитие технических систем коммуникации потребовало разработки оптимальных способов передачи информации по каналам связи. Решение соответствующих проблем (кодирование и декодирование сообщений, выбор помехоустойчивых кодов и т. д.) требовало прежде всего отве-

та на вопрос о *количестве информации*, которое можно передать в единицу времени, пользуясь данным набором сигналов.

И хотя в классической теории информации вопрос "Что такое информация?" даже не ставится и, вообще говоря, сама классическая теория информации практически бесполезна в вопросах биоинформатики, но для общего образования полезно познакомится с определением количества информации, введённым Шенноном на примере текстового сообщения.

Формула Шеннона. Количество информации, I_N, в сообщении, содержащем N символов, равно

$$I_N = -N\sum_i^M p_i \log_2 p_i ,$$

где M – число букв в алфавите; p_i – частота встречаемости i-й буквы в языке, на котором написано сообщение; знак "минус" перед всей правой частью формулы поставлен для того, чтобы количество информации было всегда положительным, несмотря на то, что $\log_2 p_i < 0$, поскольку $p_i < 1$. Двоичные логарифмы в формуле Шеннона выбраны для удобства.

Например, при однократном бросании монеты $M = 2$ ("орел" или "решка"), $N = 1$ и $p_i = \frac{1}{2}$. При этом получаем минимальное количеств информации ($I = 1$), которое называется "бит".

Иногда в формуле Шеннона используются натуральные логарифмы. Тогда единица информации называется "нат" и связана с битом соотношением: 1 бит = 1,44 ната.

Формула Шеннона позволила определять пропускную способность каналов связи, что послужило основанием для улучшения методов кодирования и декодирования сообщений, выбора помехоустойчивых кодов, в общем, для разработки основ теории связи.

Для примера возьмём некоторый текст, который можно рассматривать как результат выбора определённого варианта расстановки букв.

В общем случае, когда делается выбор одного варианта из n возможных (реализующихся с вероятностью p_i, $i = 1, 2, \ldots, n$), количество информации выражается формулой

$$I = -\sum_i^n p_i \log_2 p_i , \qquad i = 1, 2, \ldots, n .$$

Если все варианты равновероятны, т.е. $p_i = \frac{1}{n}$, то

$$I = \log_2 n ,$$

В частном случае сообщения из N букв из бинарного алфавита ($M = 2$) число вариантов равно: $n = 2^N$, количество информации $I = N$.

На этом примере удобно пояснить, что означает слово "равноправные" в определении информации. Представим, что в тексте имеются символы, которые в алфавите вообще не содержатся (не "буквы"). Априорная вероятность такого символа считается очень малой ($p_{n+1} \ll \frac{1}{n}$) и при суммировании не учитывается, поскольку он выпадает из рассматриваемого множества.

Отметим, что формула Шеннона отражает *количество* информации, но не *ценность* ее.

Поясним это на примере. Количество информации в сообщении, определяемое формулой Шеннона, *не зависит* от того или иного сочетания букв: можно сделать сообщение бессмысленным, переставив буквы. В этом случае ценность информации исчезнет, а количество информации останется прежним. Из этого примера следует, что *подменять* определение информации (с учетом всех её качеств) определением количества информации нельзя.

Вернемся снова к формуле Шеннона и проанализируем, например, текст "Завтра будет буря". Действительно, осмысленность или информация текста "Завтра будет буря" очевидна. Достаточно, однако, сохранив все элементы (буквы) этого сообщения, переставить их случайным обра-

зом, например, "рдеа Звубуб траяи", как оно утратит всякий смысл. Но бессмысленной информации не бывает. Согласно же формуле Шеннона оба предложения содержат одинаковое "количество информации". О какой же информации здесь идет речь? Или, вообще, можно ли говорить об информации по отношению к разрозненным элементам сообщения?..

Очевидно, отдельные элементы сообщения можно назвать "информацией Шеннона" лишь при условии, если перестать связывать информацию с осмысленностью, то есть с содержательностью. Но тогда это бессодержательное нечто вряд ли стоит называть "информацией", вкладывая в первичный термин несвойственный ему смысл. Учитывая, однако, что элементы сообщения реально используются для составления осмысленных текстов, содержащих информацию, эти элементы (буквы, сигналы, звуки) удобнее трактовать как *информационную тару*, которая может содержать информацию, а может быть и бессодержательной, *пустой*.

Очевидно, что *ёмкость тары* не зависит от того, заполнена ли она и чем она заполнена. Поэтому частотную характеристику элементов сообщения (или количество информации, связанное с i-й буквой алфавита), которое определяется как $H_i = -\log_2 p_i$ лучше называть не "количеством информации", а "ёмкостью информационной тары". Это, кстати, хорошо согласуется с формулой Шеннона, по которой "количество информации" в данном сообщении не зависит от порядка следования составляющих его букв, а только от их числа и частотных характеристик.

Очевидно, что в терминах Шеннона количество информации в интроне и экзоне одинаковой длины равно, в то время как экзон участвует в биосинтезе белка (имеет смысл) а интрон – нет.

Заметим, однако, что текст "Завтра будет буря" понятен русскому читателю, но является "китайской грамотой" для иностранца. Это говорит о том, что каждый раз, когда мы говорим о *семантике*, необходимо иметь в виду семантическое родство сообщения и воспринимающей системы.

Семантика (от др.-греч. σημαντικός – обозначающий) – раздел языкознания, изучающий смысловое значение единиц языка.

Приведем пример. Имеется текст на русском языке, содержащий N_K букв кириллицы (алфавит содержит 32 буквы). Перевод его на анг-

лийский содержит N_L букв латинского алфавита (26 букв). Русский текст – это результат выбора определённого расположения русских букв (число вариантов порядка 32^{N_K}). Английский перевод – это выбор определённого расположения латинских букв, который предопределен русским текстом (рецепция информации). Число вариантов в английском текста порядка 26^{N_L}. Количество *ценной* информации одинаково (если смысл не искажен), а количество информации различно.

Ниже, на примерах, мы увидим, что процессы генерации, рецепции и обработки ценной информации сопровождаются "переливанием" информации из одной тары в другую. Так, в процессе трансляции генетическая информация "переливается" из нуклеотидной информации в ДНК в аминокислотную информацию в белках. При этом, как правило, количество информации изменяется, но количество *ценной* информации сохраняется. Иногда "информационные тары" столь различны, что можно говорить об информациях разного типа. Этот термин мы также будем применять к информациям, имеющим одинаковый смысл и ценность, но сильно различающимся количественно, т. е. помещенным в разные тары.

Сам Шеннон хотя и не различал понятия информация и количество информации, но чувствовал, что это не одно и то же. "Очень редко, – писал Шеннон, – удаётся открыть одновременно несколько тайн природы одним и тем же ключом. Здание нашего несколько искусственно созданного благополучия слишком легко может рухнуть, как только в один прекрасный день окажется, что при помощи нескольких магических слов, таких как информация, энтропия, избыточность... нельзя решить всех нерёшенных проблем".

Информация и энтропия. Понятие "энтропия" (от греческого слова ἐντροπία, означающего "поворот", "превращение") было введено в физику в 1865 г. Рудольфом Клаузиусом как количественная мера неопределённости. Согласно второму началу термодинамики, в замкнутой системе энтропия либо остаётся неизменной (если в системе протекают обратимые процессы), либо возрастает (при неравновесных процессах), а при состоянии равновесия достигает максимума.

Статистическая физика рассматривает энтропию (обозначаемую символом S) в качестве меры вероятности пребывания системы в данном состоянии. Людвиг Больцман отмечал (1894 г.), что энтропия связана с "потерей информации", поскольку энтропия сопровождается уменьшением числа взаимоисключающих возможных состояний, которые остаются допустимыми в физической системе после того, как относящаяся к ней макроскопическая информация уже зарегистрирована.

По аналогии со статистической механикой Клод Шеннон ввел в теорию информации понятие энтропии в качестве свойства источника сообщений порождать в единицу времени то или иное число сигналов на выходе. Энтропия сообщения – это частотная характеристика сообщения, выражаемая формулой Шеннона.

Норберт Винер писал: "Как энтропия есть мера дезорганизации, так и передаваемая рядом сигналов информация является мерой организации. Действительно, передаваемую сигналом информацию возможно толковать, по существу, как отрицание её энтропии и как отрицательный логарифм её вероятности. То есть, чем более вероятно сообщение, тем меньше оно содержит информации". Мера неопределённости – это число двоичных знаков, необходимое для фиксирования (записи) произвольного сообщения от конкретного источника, либо среднее значение длины кодовой цепочки, соответствующее самому экономному способу кодирования.

Леон Бриллюэн развил так называемый *негэнтропийный* принцип информации, согласно которому информация – это энтропия с обратным знаком (*негэнтропия*). Бриллюэн предложил выражать информацию (I) и энтропию (S) в одних и тех же единицах – информационных (биты) либо физических (эрг/град). В отличие от энтропии, рассматриваемой в качестве меры неупорядоченности системы, негэнтропия – мера её упорядоченности. Применяя вероятностный подход, можно рассуждать следующим образом. Допустим, физическая система имеет несколько возможных состояний. Увеличение информации о физической системе эквивалентно фиксированию этой системы в одном определённом состоянии, что приведет к уменьшению энтропии системы, $I + S = const$. Чем больше известно о системе, тем меньше её энтропия. При утрате информации о

системе возрастает энтропия этой системы. Увеличивать информацию о системе можно, лишь увеличивая количество энтропии в среде вне системы, причем всегда $\Delta S \geq I$.

В соответствии со вторым началом термодинамики энтропия замкнутой системы не может убывать со временем. Получается, что в замкнутой системе (например, в тексте) увеличение энтропии может означать только "забывание" информации, чтобы равенство $I + S = const$ сохранялось. При этом возникновение новой информации возможно только в открытой системе, где параметры порядка становятся динамическими переменными.

2.3. Свойства информации

Для рассмотрения динамических открытых систем в настоящее время разрабатывается *динамическая теория информации*, которая тесно связана с синергетикой. Именно в динамической теории информации используется приведенное выше определение информации, как *запомненный выбор одного варианта из нескольких возможных и равноправных.*

Такое определение информации позволяет моделировать процессы генерации информации вообще, и предсказывать механизмы зарождения жизни на Земле в частности. Кроме того, оно допускает введение меры – количества информации по Шеннону.

Свойства, присущие всем видам информации, разделяются на две крупные группы, внутри каждой из которых свойства тесно связаны между собой. Для одной группы ключевым свойством является *фиксируемость* информации (рисунок 3).

Для другой группы определяющим свойством является её *действенность* (рисунок 4). Остальные свойства, входящие в эти группы, можно расценивать как раскрытие, проявление ключевых особенностей в формах, доступных для регистрации.

Фиксируемость информации. *Фиксируемость* – это свойство, благодаря которому любая информация, не будучи ни материей, ни энергией, может существовать не в свободном виде, а *только в зафиксированном*

состоянии – в виде записи на каком-либо физическом носителе. Способы записи информации на таком носителе всегда *условны*, т. е. не имеют никакого отношения к её семантике (или содержательности). Например, одно и то же предложение может быть записано и на бумаге, и на дискете.

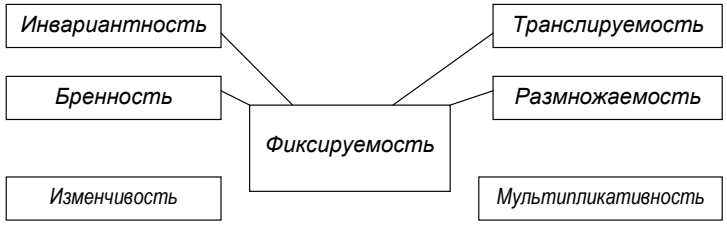


Рисунок 3 – Классификация свойств информации относительно её фиксируемости

Условность способов фиксации информации означает, что любой из таких способов, никак не связанных с семантикой, тем не менее, однозначно обусловливается *двумя* факторами, тоже не имеющими отношения к семантике, – физической *природой* носителя и спецификой *считывающего устройства* той информационной системы, к которой относится данная информация. Фиксация информации всегда представляет собой *деформацию* (в той или иной степени) носителя, среднее время релаксации которого должно быть больше среднего времени считывания, что ограничивает способы записи информации на том или ином носителе.

В нашем примере на бумагу наносятся чернила, а на дискете происходит намагничивание дорожек на магнитной плёнке.

В группу, "возглавляемую" свойством фиксации, входят также такие свойства, как *инвариантность*, *бренность*, *изменчивость*, *транслируемость*, *размножаемость* и *мультипликативность* информации.

Инвариантность информации по отношению к физической природе её носителей определяется как возможность фиксации информации (записи) на любом языке, любым алфавитом. Ни количество, ни семантика информации *не зависят* от избранной системы записи или от природы носителя. Инвариантностью определяется возможность осуществлять разные элементарные информационные акты создания, приема, передачи, хранения и использования информации. Именно свойство инвариантности лежит в основе возможности расшифровки генетического кода.

Бренность обусловлена тем, что информация всегда зафиксирована на каком-либо физическом носителе. Поэтому сохранность и само существование информации определяется судьбой её носителя. Свойство бренности позволяет говорить о *сроке жизни* информации, который зависит от состояния её носителя. Рано или поздно носитель деформируется, и информация исчезает.

Изменчивость – это свойство информации, ассоциируемое с её бренностью, другими словами, с сохранностью её носителя. Исчезновение информации может происходить не только из-за её разрушения, но и вследствие её изменения при деформации носителя. Под изменчивостью можно понимать такие преобразования, которые затрагивают количество и (или) семантику информации, но *не лишают* её смысла.

Транслируемость – это свойство, противостоящее бренности информации, это возможность передачи информации с одного носителя на другой, т. е. размножение информации.

Пусть $V_{\text{Р}}$ и $V_{\text{Г}}$ – средние скорости размножения и гибели информации. Тогда жизнеспособность информации, L, определяется отношением $\frac{V_{\text{Р}}}{V_{\text{Г}}}$. Если $L>1$, число копий записи будет возрастать. При $L<1$ данная информация обречена на вымирание, а при $L=1$ состояние нестабильно.

Очевидно, ситуация при $L>1$ отвечает проявлению свойства *размножаемости* – прямого следствия транслируемости. В свою очередь, следствием размножения является *мультипликативность*, т. е. возможность одновременного существования одной и той же информации в виде идентичных копий на одинаковых или разных носителях.

Действенность информации. Вторая группа свойств информации объединяется ключевым свойством – действенностью (рисунок 4). Это свойство выявляется следующим образом: будучи включенной в свою информационную систему, информация может быть использована для по-

строения того или иного *оператора*, который может совершать определённые целенаправленные действия. Оператор, таким образом, выступает в роли посредника, необходимого для *проявления* действенности информации. Нужно отметить, что реализацию информации в оператор *нельзя* понимать как "материализацию" информации.

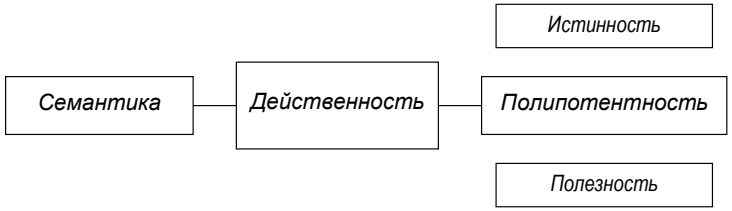


Рисунок 4 – Классификация свойств информации относительно её действенности

Семантика (или *содержательность*) информации проявляется в специфике кодируемого информацией оператора, причем каждая данная информация *однозначно* определяет оператор, для построения которого она использована. Однако природа целенаправленного действия такова, что должна *повышаться* вероятность воспроизведения информации, кодирующей такое действие. Следовательно, семантика информации всегда отражает условия, необходимые и достаточные для её *воспроизведения*. Эволюция семантики направлена в сторону *улучшения* условий воспроизведения информации.

Полипотентность информации проявляется в том, что оператор, закодированный данной информацией, может быть использован для осуществления *различных* действий (т. е. для достижения разных целей). Так, одним и тем же молотком можно вбить гвоздь, разбить стекло и проломить голову.

Это свойство не означает семантической неоднородности информации – семантика любой информации всегда однозначно отображается в операторе. Полипотентность не означает также, что на основании одной и той же информации могут быть созданы *несколько* разных операторов.

Из свойства полипотентности следуют *два* вывода.

1. Располагая некоторой информацией или созданным на её основе оператором, *невозможно* перечислить *все* ситуации и цели, для достижения которых с той или иной степенью вероятности они могут оказаться пригодными (бесконечное множество комбинаций "ситуация – цель"). Таким образом, любая информация и оператор, на ней основанный, всегда могут получить априори не предполагавшиеся применения. Такое непредсказуемое заранее использование информации может подчас оказаться даже *более эффективным и ценным*, нежели то, для которого она первоначально предназначалась.
2. Для достижения одной и той же цели в некоторой ситуации с тем или иным эффектом может быть использовано *множество разных* информаций и основанных на них операторов. Это множество всегда будет открытым, так как априори невозможно перечислить все существующие и все возможные информации, а тем более предугадать, какова будет эффективность их использования.

Полезность информации предполагает, что она кому-нибудь нужна, может быть с пользой применена для некоторых целенаправленных действий. На основании свойства полипотентности можно утверждать, что полезной может оказаться любая информация. Это делает оправданным запасание информации "впрок". Таким свойством (памятью) обладают организмы с достаточно высокой организацией. Полезность – "потенциальное" свойство, поскольку речь идет о содействии событию, которое ещё не произошло.

Истинность информации – свойство, которое выявляется в ходе реализации полезности. Критерий истинности – практика. Из свойства полипотентности информации следует относительность её истинности, т. е. зависимость от ситуации и цели. В том случае, когда целью является трансляция информации (что представляет собой достаточно общий случай), истинность оказывается условием существования информации. Получается, что жизнеспособна только истинная информация. Понятно, что

выявление истинности возможно только в том случае, если информация кому-то полезна. А это значит, что для жизнеспособности информации необходимо сочетание её истинности и полезности, т. е. "гармония объективного и субъективного аспектов информации, отражаемых этими терминами".

Ценность информации. В математической теории связи не существует вопроса о возникновении ценной информации и её эволюции. В современной динамической теории информации это одно из центральных понятий. При этом имеется в виду, что цель задана извне; вопрос о спонтанном возникновении цели внутри самой системы не ставится.

В основе понятия "ценность" лежат такие свойства информации, как действенность и полипотентность, а также способ исчисления ценности через приращение вероятности достижения той цели, для которой данная информация используется. "Бездеятельная" информация обречена на разрушение и гибель. Как уже говорилось, полипотентность информации соответствует тому, что оператор, являющийся продуктом реализации семантики информации, может быть использован для осуществления самых разных целенаправленных действий.

Целенаправленное действие специфично для живых организмов. Имеется в виду переход от "исходной ситуации" к некоторому "заданному" событию (осуществление этого события как "цели" действия) посредством оператора – механизма, применение которого в имеющихся условиях приводит к требуемому результату. Таким образом, целенаправленные действия отличаются от спонтанных изменений только в одном отношении – наличием оператора.

Мерой ценности информации является величина $C = \frac{P - p}{1 - p}$, где P – вероятность осуществления события цели в данном пространстве режимов при использовании данной информации, p – вероятность спонтанного (до получения данной информации) осуществления того же события; P и p могут изменяться от 0 до 1.

Ценность информации зависит от априорной вероятности достижения цели до получения информации, т. е. от того, какой предварительной информацией уже располагает получатель.

Предварительная осведомлённость называется *тезаурусом*. Если таковая отсутствует, то априорная вероятность во всех вариантах одинакова и равна $p = \frac{1}{n}$ (где n – число вариантов).

Отождествление "просто информации" с "ценной" и (или) "осмысленной" является ошибкой и приводит к недоразумениям.

Генерация и рецепция информации. *Генерация информации* – это выбор варианта, сделанный случайно (без подсказки извне) из многих возможных и равноправных (т. е. из принадлежащих одному множеству) вариантов. Если речь идет о возникновении новой информации, то выбор должен быть именно случайным.

Если выбор подсказан на основе уже имеющейся информации, то речь идет о восприятии, *рецепции* информации. "Запоминаемость" ассоциируется с рецепцией информации.

В работах Н. Винера и К. Шеннона (в статической математической теории информации) процессам рецепции информации практически не уделялось внимания. От рецептора (получателя) требовалась лишь способность отличать один кодовый символ от другого.

С позиций динамической теории информации (для динамических систем) рецепция информации означает *перевод* системы в одно определённое состояние независимо от того, в каком состоянии она находилась раньше. В современных технических устройствах рецепция, как правило, осуществляется с помощью электрического или светового импульса. Во всех случаях *энергия* импульса должна быть *больше барьера* между состояниями.

Переключение за счет *сторонних* сил называется *силовым*.

Другой способ переключения – *параметрический*. Он заключается в том, что на некоторое (конечное) время параметры *мультистабильной* системы изменяются настолько, что она становится *моностабильной*, т. е. одно из состояний становится неустойчивым, а затем исчезает. Система

независимо от того, в каком состоянии она находится, попадает в оставшееся устойчивым состояние.

После этого возвращаются прежние значения параметров, система *снова становится* мультистабильной, но *остаётся* в том состоянии, в которое она была переведена.

Силовое и *параметрическое* переключения представляют собой *рецепцию* информации. Различаются лишь механизмы переключения, т. е. рецепции информации.

В электронике предпочтение отдаётся силовому переключению.

В *биологических* системах преимущественно используется *параметрическое* переключение, которое может быть достигнуто неспецифическими факторами – изменением температуры, pH и др.

В случаях как генерации, так и рецепции способность генерировать или воспринимать зависит от информации, которую *уже содержит* рецептор или генератор.

Запоминание информации. Начиная с классических работ по теории информации установилась традиция связывать информацию с термодинамической величиной – энтропией. В динамической теории информации представление об информации как негэнтропии признано *неверным*, при этом существенным является необходимость разграничения понятий *макроинформация* и *микроинформация*.

Согласно определению информации, информация есть запомненный выбор, т.е. макроинформация. На физическом языке "запомнить", т.е. зафиксировать информацию, означает привести систему в определённое устойчивое состояние. Таких состояний должно быть не менее двух. Каждое из них должно быть устойчивым, иначе система может самопроизвольно выйти из того или иного состояния, что равносильно исчезновению информации.

Простейшая запоминающая система содержит всего *два* устойчивых состояния и называется *триггер*. Этот элемент играет важную роль во всех информационных системах.

Свойством запоминания могут обладать только макроскопические системы, состоящие из многих атомов. Невозможно что-либо запомнить,

располагая одним атомом, поскольку атом может находиться лишь в одном (устойчивом) состоянии, то же относится и к простым молекулам.

Наименьшая по своим размерам самая простая система, которая может запомнить только один вариант из двух возможных, – это молекула, способная находиться в *двух* различных изомерных состояниях, – при условии, что спонтанный переход из одной формы в другую происходит так редко, что его вероятностью практически можно пренебречь.

Примером таких молекул могут служить *оптические изомеры*, обладающие "правой" и "левой" хиральностью – они различаются по способности содержащих их растворов вращать вправо или влево плоскость поляризации света, пропускаемого через растворы.

К таким оптическим изомерам относятся сахара и аминокислоты, содержащие 10–20 атомов. Молекулярными триггерами могут служить макромолекулы (в частности, белковые молекулы), способные существовать в нескольких (по крайней мере, двух) конформационных состояниях.

Биологические системы *высокого* иерархического уровня (клетка, мозг, организм, популяция) тоже, разумеется, могут быть запоминающими. При этом механизм запоминания *не всегда* сводится к генетическому (т. е. макромолекулярному). Например, клетка (в частности, нервная), способная функционировать в двух и более устойчивых состояниях, уже является запоминающим устройством.

Важную роль играет и *время запоминания*. В устойчивых динамических системах оно, с формальной точки зрения, бесконечно. Триггерное переключение одного состояния на другое возможно лишь за счет стороннего сигнала, что равносильно рецепции информации. В реальности возможно спонтанное переключение за счет случайных флуктуации.

Итак, *макроинформация* может содержаться только в макрообъектах. Граница между макро- и микрообъектами проходит на уровне макромолекул, размеры которых имеют порядок нанометров.

Что касается *микроинформации*, то она не обязательно ассоциируется с микрочастицами. Любая *незапоминаемая* информация – это микроинформация.

В реальной жизни речь всегда идет о *макроинформации*, которая в частности подразумевается, когда мы говорим об информации в живых системах. Любое изменение макроинформации, увеличение или уменьшение, сопровождается ростом энтропии, что естественно, поскольку эти процессы необратимы. Количественной связи между изменениями макроинформации и физической энтропии не существует.

2.4. Генетическая информация

Теперь перейдем к генетической информации, носителями которой являются молекулы ДНК. Слова "ДНК", "гены", "наследственная информация" стали настолько привычными, что нередко воспринимаются как синонимы. В действительности это далеко не так.

Гигантская по длине молекула ДНК состоит из четырёх типов нуклеотидов, которые могут быть соединены в любой последовательности. Эти молекулы обладают свойством, которое Герман Мёллер (Hermann Joseph Muller) назвал *аутокатализом*. Если в раствор, содержащий такие молекулы, внести в должном количестве все четыре нуклеотида (основания), то при соблюдении некоторых дополнительных условий эти молекулы начнут *пристраивать* основания вдоль своей цепи точно в той же последовательности, как и в них самих, а затем отделять от себя готовые копии. Процесс этот не зависит от того, какова последовательность оснований, составляющих исходные молекулы ДНК. Это может быть случайная последовательность, или строго чередующаяся, или любая иная – копии будут всегда похожи на оригинал, если не произойдет мутации, т. е. случайной замены, вставки или выпадения одного или нескольких оснований.

Если ДНК состоит из случайной последовательности оснований, это далеко не ген, поскольку никакой наследственной информации она не содержит, хотя и может самовоспроизводиться. Информация возникает на отрезках молекулы ДНК лишь тогда, когда благодаря мутированию (или по иным причинам) там сложится такая последовательность оснований, которая сможет *повлиять* на химические процессы, протекающие в её

окружении. Только тогда, выступая в роли "катализатора", ген сможет ускорить одни или притормозить другие процессы, изменяя тем самым свое химическое окружение. Постепенно все большие *преимущества* будут получать такие структуры ДНК, которые в непосредственном своем окружении могут увеличивать концентрацию нуклеотидов и других веществ, необходимых для их размножения. Лишь когда этот процесс завершится и в "первичной" молекуле ДНК возникнут отрезки, каждый из которых *стимулирует* образование необходимых для удвоения ДНК соединений или *угнетает* синтез соединений, препятствующих их удвоению, можно считать, что в молекуле ДНК *возникли гены*, и что сама эта молекула стала *носителем* генетической информации.

Генетическая информация, следовательно, содержится в наборе генов, контролирующих синтез соединений, которые *обеспечивают удвоение* молекул ДНК в некоторых данных условиях. Появление генов тесно связано с возникновением аппарата трансляции, а также с формированием оболочек или мембран, отделяющих от внешней среды участок, где находятся молекулы ДНК. Это уже можно рассматривать как возникновение живых объектов, которые могут расти, размножаться и приспосабливаться к новым условиям благодаря генам, возникающим и изменяющимся в результате мутаций; они умирают, когда разрушаются содержащиеся в них гены или когда они не в состоянии приспособиться к внешним условиям. Изменяясь, гены влияют и на другие структуры организма, обеспечивая тем самым "заселение" все новых мест обитания, появление многоклеточных растений, грибов и животных, т. е. эволюцию жизни на Земле. Как писал Г. Мёллер, в основе жизни лежит ген.

Таким образом, совокупность генов, или генетическая информация, регулирующая целенаправленную деятельность любой живой клетки, определяется не самими основаниями ДНК, а *последовательностью их расположения*.

Различие между генетической информацией и молекулой ДНК позволяет также ввести понятие генетической информации и выяснить отличие таких её носителей от информации как таковой. Поэтому-то мы и говорим, что генетическая информация записана в ДНК определённой

последовательностью оснований. Именно эта информация, т. е. запись последовательности тех событий, которые *должны произойти*, чтобы вновь возникающие клетки могли вырасти, а затем вновь поделиться и т. д., – самый важный компонент живой клетки.

То, о чем писал Мёллер около 70 лет назад, можно сформулировать следующим образом: *живое – это совокупность объектов, содержащих информационные структуры, обладающие свойствами аутокатализа и гетерокатализа, обеспечивающие размножение этих объектов в разнообразных условиях внешней среды*. Жизнь – это возникновение все новых содержащих информацию объектов, материальные компоненты которых обеспечивают её воспроизведение во все более разнообразных и сложных ситуациях. Очевидно, что чем сложнее эти ситуации, тем больше нужно информации, чтобы в соответствии с ней построить живой объект, способный в этих ситуациях существовать.

В мире неживой Природы нет примеров информационных систем, в которых носители информации отличались бы качественно от остальных элементов системы.

Мы привыкли к словосочетанию "генетическая информация", забыли даже, что ввел его в научный обиход физик Эрвин Шредингер в середине 40-х годов. В своей книге "Что такое жизнь с точки зрения физика?" он опирался на работу Н.В. Тимофеева-Ресовского, К.Г. Циммера и М. Дельбрюка "О природе генных мутаций и структуре гена", увидевшую свет в Германии в 1935 г. Это произошло вскоре после того, как Г. Мёллер, ученик Т. Моргана, впервые показал, что гены не только воспроизводят себя и изменяются (мутируют), но что можно повлиять на частоту их мутирования, например, повышением температуры или действием ионизирующих излучений.

В 1928 г. Мёллер в статье "Ген как основа жизни" показал, что именно гены (образования неизвестной тогда природы), способные к ауто- и гетерокатализу, положили начало феномену жизни на нашей планете. "Ясно, что, став на эту точку зрения, мы избегаем логических трудностей, связанных с происхождением современной протоплазмы, с её взаимодействием частей, действующих совместно в направлении продолжения роста и точного воспроизведения целого. Система эта образовалась, так же как и сложная макроскопическая форма высших растений и животных, ... постепенно, шаг за шагом, каждый из которых проверялся по мере того, как в первичных аутокаталитических генах мутация следовала за мутацией. В этом процессе преимущественно выживали, размножались и вновь мутировали лишь те гены, побочные продукты которых оказывались наиболее полезными для дальнейшего воспроизведения... Согласно этому взгляду, который, по-видимому, наилучшим образом выдерживает проверку исчерпывающим анализом, по крайней мере значительная часть протоплазмы явилась вначале лишь побочным продуктом активности генного вещества; её функция... заключается лишь в питании генов; первичные же, свойственные всякой жизни, тайны скрыты глубже, в самом генном веществе... Мутабильного типа структуры в генном веществе несомненно претерпели в процессе эволюции глубокие изменения и подверглись усложнениям, а под их влиянием, конечно, эволюционировала и протоплазма, но другие структуры – те черты строения гена, которые ответственны за его первичное свойство аутокатализа – должны быть ещё и сейчас такими же, какими они были в незапамятные времена, когда зеленая тина ещё не окаймляла берегов морей".

Всего через 20 с небольшим лет после этой публикации было установлено, что гены представляют собой отдельные участки молекулы ДНК, размножающиеся путём комплементарного пристраивания друг к другу четырёх видов нуклеотидов; гены мутируют, когда происходят ошибки в этом процессе; они управляют синтезом разного рода белков, составляющих протоплазму, переключаясь, время от времени, с аутокатализа (построения собственных копий) на гетерокатализ (построение инородных молекул) путём синтеза РНК и, с её помощью, молекул белка.

Сейчас все это хорошо известные процессы. Можно ли проводить аналогии между свойствами живых клеток и, например, кристаллов? Рост и размножение кристаллов основаны на присоединении к исходной "затравке" всё новых, точно таких же молекул из раствора. Вероятность этого равновесного процесса зависит от температуры и концентрации раствора. Размножение вирусной частицы также зависит от условий

окружающей среды. Но вирусы (подобно живым организмам) – это открытые системы, и они с большей эффективностью используют окружающую среду для выживания и размножения. Это касается, например, поиска клетки-хозяина и размножения в ней. Прикрепившись к поверхности живой клетки, вирус с помощью специального белкового устройства впрыскивает в нее свою молекулу ДНК или РНК, содержащую его гены. Гены вируса не только воспроизводят себя, используя синтезируемые зараженной клеткой молекулы, но также заставляют эту клетку создавать новые, не свойственные ей белковые молекулы, которые, окружая готовые генетические структуры новых вирусных частиц, создают белковую оболочку вируса, приспособленную для осуществления следующего цикла – заражения других клеток и размножения в них.

Все теории происхождения жизни вращаются вокруг попыток ответить на вопрос: как возникла ДНК и та информация, которая записана в ней?

Молекулярная эволюция. Гиперциклы Эйгена. В 1971 г. Манфред Эйген (Manfred Eigen) сформулировал последовательную концепцию предбиологической молекулярной эволюции. Эйген распространил идеи дарвиновского отбора на популяции макромолекул в первичном бульоне. Далее он показал, что кооперирование молекул в "гиперциклы" приводит к компартментализации в виде отдельных клеточных единиц. Гиперцикл – это средство объединения самовоспроизводящихся единиц в новую устойчивую систему, способную к эволюции. Он построен из автокатализаторов, которые сочленены посредством циклического катализа, т. е. посредством ещё одного автокатализа, наложенного на систему.

Теория гиперциклов является абиогенетической теорией происхождения жизни, а также её эволюции. Гиперциклы, которые сами по себе ещё чистая химия, уже обладают некоторыми признаками живого: круговорот веществ и энергии, воспроизведение информации с её наследованием, приспособляемость к изменяющимся условиям. Гиперциклы подвержены дарвиновскому естественному отбору, но не на уровне видов, а на уровне молекул, т.е. это гипотеза о молекулярной эволюции, приведшей к

созданию первой живой клетки, использующей генетический код для матричного синтеза белка.

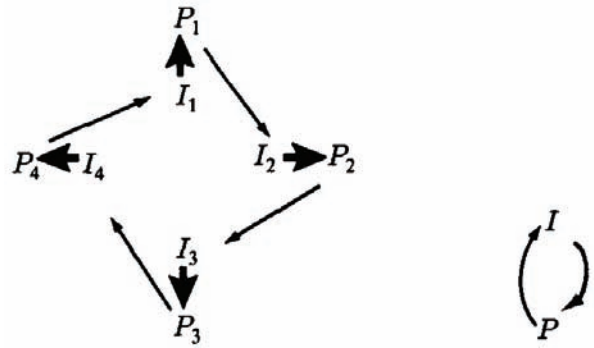
Дарвиновский отбор, являющийся предпосылкой для возникновения гиперциклов, на молекулярном уровне может иметь место в системах, обладающих следующими свойствами:

1. *Метаболизм* – Система должна быть далека от равновесия. Образование и разложение молекулярных видов должны быть независимы. Отбор должен действовать только на промежуточные состояния, которые образуются из высокоэнергетических предшественников и разрушаются в низкоэнергетические отходы. Система должна использовать освободившуюся энергию и вещества.
2. *Самовоспроизведение* – Система должна быть способна инструктировать свой собственный синтез;
3. *Мутабилъностъ* – Система должна быть способна мутировать. Мутабилъностъ всегда сопутствует самовоспроизведению.

Ошибки копирования – это основной источник новой информации. Образование и усовершенствование эйгеновских гиперциклов в ходе эволюции привели к созданию аппарата трансляции. Образование вслед за этим клеточной мембраны завершило предбиологический период эволюции.

Гиперцикл соответствует циклу биохимических процессов, в которых белки, P_i, катализируют образование полинуклеотидов, I_i, а последние кодируют биосинтез белков ($i = 1, 2, \ldots, n$). Схема гиперцикла по Эйгену приведена на рисунке 5(а). Тонкие стрелки соответствуют катализу, жирные – "кодированию".

Простейший гиперцикл содержит всего один белок-репликазу (полимеразу) и один нуклеотид ($i = 1$); схема его представлена на рисунке 5(б). Современный биосинтез белка является гиперциклом, причем достаточно сложным (рисунок 5(а)). Он содержит белок полимеразу, мРНК, набор адаптеров, набор тРНК и рибосому, т. е. количество белков и нуклеиновых кислот в нём достаточно велико. Оценим количество информации, содержащейся в таком гиперцикле.



а *б*

Рисунок 5 – Схемы гиперцикла: *а* – сложного; *б* – простейшего

В белке, состоящем из n аминокислот, полное количество информации равно: $I = \log_2 20^n$. При $n = 200$ $I = 860$ бит. Величина I соответствует количеству ценной информации в случае, когда в белке все остатки фиксированы (т. е. находятся на данном месте), как, например, в гистонах. В большинстве функциональных белков не все остатки должны быть фиксированы. Так, замена многих остатков на другие, но аналогичные, например одинаковой гидрофобности, не ведет к потере функции. В связи с этим количество ценной информации, обеспечивающей функцию белка-фермента, в общем случае меньше. Так, например, количество ценной информации в белке бактериородопсин составляет 130 бит. Того же порядка должна быть ценная информация в полинуклеотидах.

Для грубой оценки примем, что количество ценной информации в среднем белке равно 100 бит.

Количество ценной информации в системе, состоящей из m разных белков, соответственно, в m раз больше. В современном гиперцикле биосинтеза белков задействовано более ста полимеров. Поэтому полное количество ценной информации всей системы составляет приблизительно 100·100 = 10 000 бит.

Вероятность спонтанного и одномоментного возникновения всей системы равна $W \approx 2^{-10000} \approx 10^{-3300}$. Эта величина абсурдно малая.

Дело в том, что любые физические величины (длина, масса, интервал времени, число частиц) в нашем мире *не являются* бесконечно большими или бесконечно малыми.

Например, считается, что наша Вселенная появилась около 14 млрд. лет назад, то есть со времени Большого взрыва прошло "всего" порядка $4{,}4 \cdot 10^{17}$ секунд. Даже если за масштаб времени взять период тепловых колебания атомов в кристаллической решетке (порядка 10^{-12} секунд), то за время существования Вселенной произошло "только" $\sim 10^{30}$ колебаний. Кстати возраст Земли (равно как и Солнечной системы) оценивается в 4,5 млрд. лет ($1{,}4 \cdot 10^{17}$ с). Жизнь на Земле зародилась ещё в архее – примерно 3,5 млрд. лет назад (10^{17} с).

Ещё один пример, масса наблюдаемой части Вселенной оценивается в $8 \cdot 10^{52}$ кг ($\sim 10^{50}$ тонн), что соответствует 12,8 10^{77} масс атомов углерода (или $\sim 10^{79}$ масс атомов водорода).

Считается, что все "разумные" значения физических величин выражаются числами от 10^{-100} до 10^{+100}. В связи с этим американским математиком Эдвардом Каснером (*Edward Kasner*) в 1938 году было введено новое понятие – "гугол" (*googol*) – равное 10^{+100}, такое, что никакая физическая величина не может иметь значение превышающее гугол.

А само слово гугол, как название для числа со ста нулями, придумал племянник Эдварда Каснера, девятилетний Милтон Сиротта (*Milton Sirotta*), во время прогулки с дядей и обсуждения больших чисел.

Соответственно, "обратным гуглом" называют число 10^{-100}. Обратный гугол, хотя формально является конечной величиной, реально должен рассматриваться как бесконечно малая величина. В частности, вопрос: как ведёт себя функция внутри интервала порядка обратный гугол, лишён смысла. Функцию на таком интервале следует заменить числом (средним по интервалу), поскольку более детальное её поведение *принципиально* не наблюдаемо.

Кстати, название Интернет-поисковика Google было придумано на основе слова "гугол".

Таким образом, самопроизвольное возникновение аппарата биосинтеза в его современном виде абсолютно невозможно. Однако современ-

ный вид гиперцикла биосинтеза появился в результате около 3 млрд. лет эволюции, в ходе которой исходный простейший гиперцикл совершенствовался и усложнялся. При этом, если на первых этапах происходило химическое копирование молекул, то на последнем этапе эволюции аппарата биосинтеза произошёл выбор единого кода копирования – того, что мы сегодня наблюдаем как единый генетический код на Земле. Выбор единого кода имел место уже после образования (и конкуренции) нескольких различных популяций гиперциклов с различными вариантами кода. Выбранный таким способом вариант постепенно вытеснил все остальные варианты генетического кодирования.

Контрольные вопросы

1. Сформулируйте определение понятия "информация", которое наиболее адекватно для биологических применений.
2. Почему не существует единого для всех наук определения понятия "информация"?
3. Запишите формулу Шеннона и объясните смысл входящих в неё параметров.
4. Проиллюстрируйте отличие между понятиями "количество информации" и "ценность информации", используя какое--либо предложение.
5. Что такое негэнтропия и как она связана с информацией?
6. Что такое фиксируемость информации?
7. Что такое инвариантность информации?
8. Что такое бренность информации?
9. Что такое изменчивость информации?
10. Что такое транслируемость информации?
11. Что такое размножаемость информации?
12. Что такое мультипликативность информации?
13. Что такое действенность информации?
14. Что такое оператор, порождаемый информацией?
15. Что такое семантика информации?
16. Что такое полипотентность информации?

17. Что такое полезность информации?
18. Что такое истинность информации?
19. Что такое ценность информации? В чём она проявляется? Как определяется мера ценности информации?
20. Какой процесс называется генерация информации?
21. Что такое рецепция информации? Какие выделяют два способа рецепции?
22. Каким образом происходит запоминание информации?
23. Чем различаются макроинформация и микроинформация?
24. Что такое гиперцикл Эйгена? Как он устроен?
25. Что такое гугол и обратный гугол? Где они используются?
26. Оцените количество информации, содержащееся в гиперцикле биосинтеза белка?

3. Основания биоинформатики

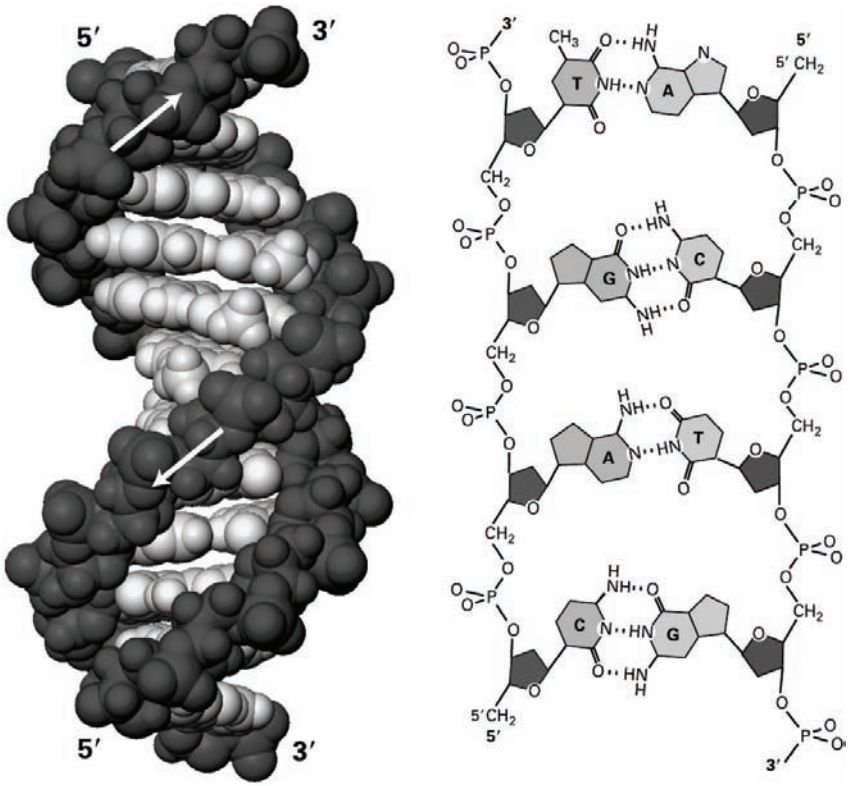
3.1. Основные положения молекулярной биологии

В информационном архиве каждого организма (геноме) содержится детальный план будущего развития и функционирования этого индивидуума, представленный ДНК или (у некоторых вирусов) РНК. Молекулы ДНК – длинные, линейные, цепочечные молекулы, несущие сообщения в четырёхбуквенном алфавите. Даже у микроорганизмов сообщение длинное, обычно состоит из миллиона букв.

Для определённости условимся обозначать нуклеотиды строчными английскими буквами:

a – аденин, g – гуанин, c – цитозин, t – тимин, u – урацил.

В структуре ДНК полностью оговорены механизмы репликации и переноса информации с гена на белок. Двойная спираль и её внутренний принцип комплементарности, необходимый для точной репликации, хорошо известны (рисунок 6).



а *б*

Рисунок 6 – Схема строения ДНК: *а* – ван-дер-ваальсовая модель, *б* – схема химических связей в ДНК

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 реализуется 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 – это длинные линейные цепочечные полимеры, состоящие из мономеров – аминокислот.

Двадцать природных (протеиногенных) аминокислот по полярности бокового радикала можно разделить на неполярные, полярные и заряженные. Мы будем использовать однобуквенные обозначения аминокислот прописными латинскими буквами следующим образом:

Неполярные аминокислоты:

G – глицин (Gly)	A – аланин (Ala)	P – пролин (Pro)
V – валин (Val)	I – изолейцин (Ile)	L – лейцин (Leu)
F – фенилаланин (Phe)	M – метионин (Met)	

Полярные аминокислоты:

S – серин (Ser)	C – цистеин (Cys)	T – треонин (Thr)
N – аспарагин (Asn)	Q – глутамин (Gln)	Y – тирозин (Tyr)
W – триптофан (Trp)		

Заряженные аминокислоты:

D – аспарагиновая кислота (Asp)	K – лизин (Lys)
E – глутаминовая кислота (Glu)	A – аргинин (Arg).

Генетический код – это шифр: триплеты букв из последовательности ДНК обозначают аминокислоты (таблица 1). 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 нуклеотидов в ДНК. Однако в большинстве организмов не вся ДНК кодирует РНК или белки. Некоторые участки последовательности ДНК существуют для управления процессами транскрипции и репликации, а бо́льшая часть генома, похоже, является "ненужной" (вероятнее всего, нам просто пока ничего не известно о её функции).

Таблица 1 – Стандартный генетический код

Первый нуклеотид	Второй нуклеотид				Третий нуклеотид
	u	c	a	g	
u	Phe	Ser	Tyr	Cys	u
	Phe	Ser	Tyr	Cys	c
	Leu	Ser	STOP	STOP	a
	Leu	Ser	STOP	Trp	g
c	Leu	Pro	His	Arg	u
	Leu	Pro	His	Arg	c
	Leu	Pro	Gln	Arg	a
	Leu	Pro	Gln	Arg	g
a	Ile	Thr	Asn	Ser	u
	Ile	Thr	Asn	Ser	c
	Ile	Thr	Lys	Arg	a
	Met (START)	Thr	Lys	Arg	g
g	Val	Ala	Asp	Gly	u
	Val	Ala	Asp	Gly	c
	Val	Ala	Glu	Gly	a
	Val	Ala	Glu	Gly	g

Молекулы ДНК, содержащие стандартные четыре буквы (a, c, g, t), сходны по химическому строению, а сама пространственная структура молекулы ДНК в первом приближении однородна.

Белкам же, наоборот, свойственно большое разнообразие трёхмерных конформаций. Эти конформации необходимы белкам для выполнения их разнообразных структурных и функциональных ролей.

Последовательность аминокислот в белке – *первичная структура* белка – определяет его трёхмерную структуру.

Для каждой природной аминокислотной последовательности существует уникальное стабильное нативное состояние – *третичная структура*, – 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", денатурирует, образуя беспорядочную биологически неактивную структуру. Именно поэтому в нашем организме существуют механизмы для поддержания относительно постоянных внутренних условий.

При восстановлении же нормальных условий пептидные молекулы вновь приобретают свою функциональную третичную структуру, которая неотличима 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 трёхмерному миру, в котором мы все живем.

Однако существует следующий *парадокс*. С одной стороны, трансляцию последовательностей ДНК в последовательности аминокислот очень легко описать логически – она определяется генетическим кодом. А сворачивание полипептидной цепи в точно определённую трёхмерную структуру очень трудно описать логически. С другой стороны, для осуществления трансляции необходимы исключительно сложный механизм работы рибосомы, транспортные РНК (тРНК) и связанные с ними молекулы. А сворачивание белков происходит самопроизвольно без посторонней помощи.

Функции белков зависят от приобретения ими нативной третичной структуры. Например, нативная структура фермента может иметь на своей поверхности впадину (активный центр), которая связывает одну малую молекулу субстрата и помещает её рядом с аминокислотными остатками каталитического центра.

Таким образом, мы имеем следующие информационно-управляемые зависимости:

- 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, из поля зрения выпадают такие вопросы, как специализация тканей во время развития или, в более обобщенном смысле, влияние условий окружающей среды на генетические события.

В некоторых тривиальных 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 фундаментальные вопросы о потоке информации и регуляции этого потока внутри организма сейчас начинают активно изучаться методами биоинформатики.

3.2. Информационно-компьютерные компоненты биоинформатики

Сегодня информационно-компьютерные компоненты являются неотъемлемой частью биотехнологии. Компьютеры необходимы для управления биологическими данными, объём и сложность которых непрерывно растут. Появление международной сети Интернет произвело революцию в мире связи. Создание World Wide Web (WWW, "Всемирная паутина") способствовало успешному внедрению и развитию Интернета. Интернет, будучи глобальной сверхмагистралью, даёт возможность пользователям свободно перемещаться в пределах WWW – крупнейшего собрания разнородных информационных ресурсов.

Компьютер – это электронная вычислительная машина, применяемая для хранения и обработки информации в режиме двоичного счета. Появление биоинформатики было бы невозможным без достижений в области конструирования аппаратных средств и разработки программного обеспечения. Для хранения информации необходимы носители с высокой

скоростью работы и большой ёмкостью. Для осуществления выборки и анализа информации нужны специальные программы.

Аппаратными средствами компьютера являются физические устройства: процессор, дисководы и дисплей.

Программное обеспечение – это собирательный термин, обозначающий совокупность различных программ, предназначенных для выполнения на компьютерах.

Программное обеспечение подразделяют на две категории: *системное* и *прикладное*.

Системное программное обеспечение включает в себя операционную систему компьютера и совокупность любых других программ, необходимых для запуска приложений, тогда как *прикладное* программное обеспечение устанавливается пользователем для выполнения специальных задач.

Компьютерные программы пишут на самых разных *языках программирования*: в машинных кодах, на ассемблерах или же языках высокого уровня. Программы, написанные на ассемблере или языке высокого уровня, должны быть преобразованы в машинный код путём ассемблирования и компиляции.

В операционной системе Windows файлы в машинном коде называются *исполняемыми файлами*, а соответствующие файлы в системе Unix – *исполняемыми образами*. Такие файлы непосредственно выполняются процессором компьютера.

Сценарии – это файлы, выполняемые какой-либо программой. Их пишут на таких *языках подготовки сценариев*, как, например, Microsoft Visual Basic, Java Script и PERL.

Существует множество различных *языков программирования*, *подготовки сценариев* и *разметки*, нашедших свое применение в биоинформатике.

HTML (*HyperText Markup Language* – Язык разметки гипертекста) – предназначен для задания внешнего вида гипертекстового документа, включая определение позиций гиперссылок. Следует отметить, что HTML не является языком программирования.

Java Script – это популярный язык подготовки сценариев, который расширяет функциональные возможности гипертекстового документа, позволяя включать в веб-страницы такие элементы, как всплывающие окна, анимации, а также объекты, изменяющие внешний вид при наведении на них указателя мыши.

Java представляет собой универсальный и машинонезависимый язык программирования, предназначенный для создания приложений, выполнимых на различных аппаратных платформах. Исходный код Java – "C++". Java отличается от Java Script. Апплеты Java встраивают в гипертекстовые документы.

XML (*Extensible Markup Language* – Расширяемый язык разметки РЯР) – позволяет описывать файлы по типу содержащихся в них данных.

Phyton (Питон) – это полный объектно-ориентированный интерпретируемый, переносимый язык сверхвысокого уровня, язык подготовки сценариев, написанный Гёйдо ван Россумом в 1998 году. Программирование на Phyton позволяет получать быстро и качественно необходимые программные модули. Интерпретатор Питона может быть перенесён на любую платформу, будь то Unix, Windows, Linux, RiscOS, MAC, Sun. При написании кода на Phyton не нужно заботиться о конечной платформе, кроме тех случаев, когда используются специфические модули для данной системы. Таким образом, Phyton представляет конкурента для Java, обеспечивая лёгкую переносимость, одновременно сочетая в себе средства доступа к ресурсам операционной системы. Phyton содержит средства быстрого и легкого формирования графического интерфейса пользователя, библиотеку применяемых в структурной биологии функций и обширную библиотеку численных методов. Phyton может быть загружен со своей домашней страницы: http://python.org.

PERL (*Practical Extraction and Reporting Language* – Практический язык извлечения данных и формирования отчетов, "ПЕРЛ") – универсальный язык сценариев, который широко используется в анализе данных секвенирования. PERL был изобретен Лэрри Уоллом на основе языков "Sed", "Awk", оболочки Unix и "C". PERL позволяет выполнять превосходное сопоставление регулярных комбинаций знаков, имеет гибкий синтаксис,

или грамматику, и требует сравнительно небольшое число кодов для программирования различных операций. Он хорош для обработки строк, то есть основных действий, производимых при анализе последовательностей и управлении базами данных. Этот язык контролирует и оптимизирует распределение памяти ЭВМ, а также имеет хорошую совместимость с вычислительными системами, работающими на Unix. Он доступен в сети для свободного копирования, компиляции и распечатки. "ПЕРЛ" может быть загружен со своей домашней страницы: http://www.perl.org.

Языки PERL и Phyton наиболее пригодны при создании приложений для биоинформатики – во многом благодаря своей эффективности и способности удовлетворять разнообразным функциональным требованиям данной области.

BSML (*Bioinformatic Sequence Markup Language* – Язык разметки последовательностей в биоинформатике, БСМЛ) графически описывает генетические последовательности и методы хранения и передачи закодированной информации о структуре последовательностей, а также сопутствующей графической информации.

BIOML (*Biopolymer Markup Language* – Язык разметки биополимеров, БИОМЛ) обеспечивает описание типа данных для аннотирования информации о последовательности молекулярного биополимера и данных о его структуре.

Операционная система (ОС) – это основная программа, которая управляет всеми периферийными устройствами и контролирует работу других (прикладных) программ.

BIOS (*Basic Input-Output System* – базовая система ввода-вывода, БИОС) – операционная система низкого уровня, которая частично или полностью реализована аппаратным путём (то есть записана в ПЗУ).

БИОС управляет действиями компьютера, например, принятия решений о подключении тех или иных устройств при включении компьютера, чтения и записи дисков, возвращения ответов на ввод, отображения на мониторе отчетов системы и диагностики служебных устройств. Затем управление переходит к операционной системе высокого уровня и на дисплее компьютера появляется типичный графический интерфейс поль-

зователя. Файлы, которые содержат команды для операционной системы, в Windows называют командными файлами, а в Unix – основными сценариями.

Операционная система **Windows**, принадлежащая корпорации Microsoft, – наиболее привычная операционная система для домашних и офисных персональных компьютеров.

Большая часть *корпоративных* рабочих станций и серверов работает под различными версиями операционной системы **Unix**. Операционные системы GNU и Linux соответствуют стандарту Unix.

Операционная система обеспечивает доступ к имеющимся в компьютере файлам и программам.

Unix – это мощная операционная система для работы в режиме коллективного обслуживания пользователей. Первое программное обеспечение для работы World Wide Web было разработано именно на базе ОС Unix. Операционная система Unix изобилует различными командами и функциональными возможностями – от сетевых программ до текстовых редакторов и от электронной почты до программ чтения новостей. Кроме того, она обеспечивает свободный доступ к предназначенным для загрузки из сети программам, написанным для систем Unix. В настоящее время ОС Unix существует в различных формах и реализациях.

Операционная система **Linux** считается некоммерческой версией Unix для персональных компьютеров, поскольку она может быть бесплатно загружена из сети и установлена на компьютер. Под управлением ОС Linux персональные компьютеры оказались весьма универсальными и удобными рабочими станциями. Некоторые важные пакеты программ для вычислительной биологии рассчитаны на работу в ОС Linux.

IBION ("ИБИОН") – новая машинонезависимая и функционально законченная система для биоинформатики. Это крупнейший сервер, приспособленный для нужд биоинформатики: он содержит в себе вебсервер Apache (Апач), реляционную базу данных postgreSQL, статистический язык "R" и работает на аппаратных средствах фирмы Intel (Интел) с предварительно установленными ОС Linux и полным комплектом программ и баз данных для биоинформатики.

Обычно программное обеспечение поставляется на дискетах или компакт-дисках. Мы говорим, что файл закачивается (загружается, *Download*), когда он копируется с удаленного источника на местный компьютер, и что он скачивается (*Upload*), когда копируется с жесткого диска компьютера и передаётся к удаленному источнику.

Загрузка (*Download*) из Интернета возможна тремя путями:

1) непосредственно из гипертекстового документа;
2) с FTP-сервера;
3) по электронной почте.

3.3. Интернет-компоненты биоинформатики

Интернет – это глобальная сеть компьютеров и местных компьютерных сетей, связывающая многочисленные правительственные, учебные и коммерческие учреждения. Она позволяет компьютерам общаться на своих электронных языках. Биологическая информация хранится на многих компьютерах, рассеянных по всему миру, и самый легкий путь доступа к этой информации – объединение всех этих компьютеров в единую сеть.

Компьютеры могут быть соединены друг с другом разными способами, наиболее часто – оптоволоконными или коаксиальными кабелями и линиями беспроводной (*wireless*) связи, что позволяет осуществлять обмен данными между удаленными пользователями.

Для эффективной работы созданной системы объединенных сетей был разработан *единый протокол связи* TCP/IP – *Transmission Control Protocol* / *Internet Protocol* (Протокол управления передачей (данных) / Интернет-протокол).

TCP определяет правила разбиения данных на пакеты и последующей сборки переданных по каналу связи пакетов.

IP управляет адресацией и выбором маршрута передачи информационных пакетов по сети.

Подключенные к сети компьютеры рассматриваются как узлы и поддерживают взаимную связь посредством передачи пакетов данных.

Для осуществления передачи *данные* сначала *разбиваются* на маленькие посылаемые независимо друг от друга *пакеты* (единицы информации), которые потом объединяются при достижении своего адресата. Но пакеты не обязательно пересылаются непосредственно от одной машины к другой; они могут пройти через несколько компьютеров, стоящих на пути к конечному получателю. На случай если какой-либо из промежуточных узлов выбранного маршрута не работает, в сетевых протоколах предусмотрена функция поиска альтернативного пути, что возможно благодаря взаимному пересечению различных маршрутов.

Интернет предоставляет средства распространения программного обеспечения и позволяет исследователям проводить сложный анализ на удаленных серверах.

До конца 1980-х гг. существовало *три основных способа* доступа к базам данных через Интернет:

1) серверы электронной почты;
2) FTP;
3) сервер TELNET (ТЕЛНЕТ).

Сервер электронной почты – это средство передачи текстовых сообщений с одного компьютера на другой. FTP – средство (протокол) пересылки компьютерных файлов (например, программ) между удаленными машинами. TELNET – сетевой протокол, который позволяет оператору подключаться к удаленным компьютерам и работать на них, как будто они имеют физический доступ к этим машинам.

Серверы электронной почты позволяли ученым обмениваться информацией путём отправки запроса в электронном письме по адресу почтового сервера. Рано или поздно запрос обрабатывался сервером, и результат отсылался обратно в почтовый ящик отправителя. Однако такая система имела свои недостатки – запросы обрабатывались плохо с ошибками и необходимо было неопределённо долго ждать ответ.

FTP позволял исследователю закачать полную базу данных и производить поиск на своем компьютере. Этот способ доступа к базам данных

также имеет свой изъян – исследователь должен периодически загружать все используемые им базы данных после каждого их обновления.

TELNET даёт пользователю возможность подключаться к удаленному компьютеру и получать доступ к его программным и аппаратным ресурсам. Этот метод полезен для эпизодических запросов. К его неудобствам можно отнести сложное управление опознаванием пользователей и перегрузка вычислительных возможностей удаленного компьютера.

После того как машины были соединены друг с другом посредством сети, возникла необходимость найти однозначный способ обозначения отдельных компьютеров таким образом, чтобы сообщения и файлы могли быть отправляемы строго своему адресату.

С целью облегчения связи между узлами, каждому компьютеру в сети Интернет присвоен *уникальный опознавательный номер* (его IP-адрес, *Internet Protocol address*). IP-адрес уникален и обозначает только одну машину. Его записывают арабскими цифрами, разделёнными точками.

Например, компьютер Национального центра биотехнологической информации (*National Center for Biotechnology Information*, NCBI) при Национальной медицинской библиотеке (*National Library of Medicine*, NLM) при Национальном институте здоровья (*National Inatitute of Health*, NIH) при правительстве США имеет следующий IP-адрес: 130.14.29.110.

Эти числа обозначают конкретную машину, узел, в котором расположена эта машина, а также домен (и субдомен), которому этот узел принадлежит. Эти числа помогают компьютерам определять направления передачи данных.

Помимо этого была создана альтернативная *иерархическая система имен доменов*, устанавливающая соответствие между числовыми IP-адресами и текстовыми именами, и благодаря которой адреса Интернета можно записывать в более понятной форме. Например, запись "ncbi.nlm.nih.gov" равносильна представленным выше числам и означает: узел "Национального центра биотехнологической информации" (ncbi) при "Национальной медицинской библиотеке" (nlm) при "Национальном институте здоровья" (nih) при правительстве США (gov).

Определить соответствие между числовыми IP-адресами и текстовыми именами, а также географическое местоположение IP-адреса или узла Интернета можно на сайте: http://smart-ip.net/tools/geoip.

"Всемирная паутина" значительно повысила возможности доступа по перекрестным ссылкам, обеспечив эффективную интеграцию баз данных, рассредоточенных в сети Интернет, и таким образом устранив потребность загрузки и ведения на местных компьютерах многочисленных копий баз данных. Благодаря этому исследователь может легко просматривать записи баз данных с помощью активных гипертекстовых перекрестных ссылок с возможностью возвращения к последней просмотренной записи.

ExPASy – http://www.expasy.org/ – первый веб-сервер молекулярной биологии (Expert Protein Analysis System – Экспертная система анализа белков, "Экспази") был создан в 1993 году совместно Клиникой Женевского университета и самим Женевским университетом.

Веб-страницами называют документы, которые появляются в окне программы-обозревателя (браузера), когда мы путешествуем по "Всемирной паутине". Каждый отображаемый браузером документ сети называют веб-страницей, а совокупность веб-страниц данного сервера в собирательном значении называют *веб-узлом*.

По своему содержанию веб-страницы подобны обычным текстовым документам, за исключением лишь того, что они намного более гибки, поскольку могут содержать ссылки на любые другие страницы и файлы, размещенные в пределах сети.

Веб-узел – это собрание взаимосвязанных веб-страниц, находящихся на одном компьютере. Каждому веб-узлу в сети Интернет присвоен уникальный адрес. Наиболее замечательная особенность веб-страниц – наличие ссылок. Ссылка на веб-странице (гиперссылка) позволяет пользователю перейти к другой странице, расположенной в том же веб-узле, или даже к какой-либо странице на другом веб-узле, расположенном в любой точке мира.

Весьма ценное качество "Всемирной паутины" – простой доступ к статическим страницам с подсвеченным текстом, по которому можно

щелкать мышью и таким образом просматривать связанные между собой страницы с рассредоточенной по ним информацией.

Объектная сеть предназначена для поддержки высокофункциональных диалоговых систем. Это многозвенная архитектура, которая содержит два объекта и уровень связи. Один объект может представлять интерфейс пользователя, а другой – обеспечивать необходимые вычисления. Для передачи данных между этими двумя объектами необходимо описать сообщения, которые они могли бы принимать.

Обмен сообщениями между двумя или более объектами осуществляется посредством специального кода ORB (*Object Request Broker* – Брокер объектных запросов, БОЗ), установленного на каждой машине и способного интерпретировать описания пересылаемых сообщений и переводить их на собственный язык каждого объекта. С помощью объектной сети система может быть разбита на самостоятельные компоненты, написанные на разных языках и работающие на разных аппаратных системах.

CORBA (*Common Object Request Broker Architecture* – Общая архитектура брокеров объектных запросов, ОАБОЗ) обеспечивает стандарты, унифицирующие эту связь. CORBA включает в себя язык для описания структуры сообщений, IDL (*Interface Definition Language* – Язык описания интерфейсов, ИДЛ), а также архитектуру для программ-посредников, или БОЗов. БОЗы обеспечивают "прозрачную" связь между удаленными объектами и формируют магистраль (разводку объектной сети).

Интернет-браузеры или программы-обозреватели. Весь потенциал Интернета был полностью осознан только с появлением программ-обозревателей (браузеров), которые впервые обеспечили свободный доступ к информации, расположенной на разных веб-узлах.

*Бр**а**узерами* (от англ. *browser*) или *обозревателями* называют приложения-клиенты, посылающие запросы серверам, используя набор стандартных протоколов и соглашений. Типичный браузер сети содержит минимальный набор программных средств, необходимых для осуществления поиска, извлечения, отображения и пересылки информации по сети Интернет.

Первая точка контакта между обозревателем и сервером – домашняя страница. После загрузки этой начальной страницы обозреватель раскрывает интерфейс, удобный для выборки документов, доступа к файлам, поиска в базах данных и т. д. Наиболее популярными программами-обозревателями стали: Internet Explorer, Lynx, Mosaic, Mozilla Firefox и Netscape Navigator.

С помощью браузера пользователи могут перемещаться по содержимому окна или между интернет-окнами, щелкая по специальным словам, кнопкам или картинкам. Эти активизируемые щелчком мыши объекты известны под общим названием *гиперссылки*.

Гиперссылки при наведении на них указателя мыши обычно выделяются некоторым способом – контрастным цветом, подчеркиванием, рамкой и т. д. Щелчок по выделенной ссылке вызывает необходимый документ независимо от его местоположения: на том же самом сервере или на сервере в другой части света.

Каждому гипертекстовому документу присвоен уникальный адрес, называемый URL (*Uniform Resource Locator* – Унифицированный указатель (информационного) ресурса). Строка URL имеет следующий стандартизованный формат:

```
http://собственно адрес.
```

Здесь http – аббревиатура протокола связи, используемого серверами сети – протокола передачи гипертекстовых файлов (*HyperText Transfer Protocol*, HTTP). HTTP – протокол, используемый для обмена информацией в пределах "Всемирной паутины". Собственно адрес указывает местоположение гипертекстового документа в сети Интернет.

Гипертекстовые документы пишут на стандартном *языке разметки*, известном как HTML – язык разметки гипертекста (*Hypertext Markup Language*, HTML). Код HTML строго текстоориентированный, и любая сопутствующая графическая или звуковая информация этого документа существует в виде отдельных файлов в общем формате. Команды разметки позволяют автору веб-страницы выделять текст жирным шрифтом (команда <B>), вставлять горизонтальные линейки разметки (<HR>), изо-

бражения (<IMG>) и т. д.; каждый из этих режимов выключается соответствующим знаком </> (например </B>).

XML (*eXtensible Markup Language*, Расширяемый язык разметки, РЯР) – это другая технология, поддерживающая создание функционального хранилища генетической информации. XML, подобно HTML, может быть использован для создания веб-страниц. XML помечает данные способом, понятным любому другому приложению. Эта технология обеспечивает общий язык представления данных в стандартном формате. Она позволяет описывать файлы по типу содержащихся в них данных.

XML – текстовый формат, предназначенный для хранения структурированных данных (взамен существующих файлов баз данных), для обмена информацией между программами, а также для создания на его основе более специализированных языков разметки (например, XHTML), иногда называемых словарями.

XML более гибкий и надёжный по сравнению с HTML. Он обеспечивает метод описания смысла, или семантики содержимого документа. Одно из его преимуществ заключается в возможности управления не только способом отображения данных на веб-странице, но также и способом обработки этих данных различными программами или DBMS (*Database Management System* – Система управления базами данных, СУБД).

XML является упрощённым подмножеством языка SGML (*Standard Generalized Markup Language* – Стандартный обобщённый язык разметки) – метаязыка, на котором можно определять язык разметки для документов. SGML – наследник разработанного в 1969 году в IBM языка GML (*Generalized Markup Language*).

3.4. Биоинформационные данные, сети и базы

Компьютеры хранят информацию о последовательностях в виде строк – простых рядов последовательных знаков. Каждый знак выражен двоичным кодом и представлен наименьшей единицей информации, называемой *байтом*. Каждый байт состоит из 8 *битов*, и каждый бит может

принимать значение 0 либо 1, что даёт 255 различных комбинаций битов, то есть возможность кодирования одним байтом 255-ти знаков.

Последовательность ДНК обычно хранится и обрабатывается в компьютере в виде ряда 8-битовых слов в упомянутом двоичном формате.

Белковая последовательность представлена как ряд 8-битовых слов, состоящих из буквенных обозначений аминокислот (см. п.3.1) в двоичной форме.

Обычно информацию о последовательности ДНК или белка записывают в текстовый файл в стандартном формате ASCII или в формате программы FASTA (FAST Allignment – быстрое выравнивание, "ФАСТА").

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.
- Используются *однобуквенные коды* для нуклеотидов и аминокислот, заданные Международным Объединением Биохимии и Международным Объединением Чистой и Прикладной Химии (IUB/IUPAC).

http://www.chem.qmw.ac.uk/iupac/misc/naabb.html
http://www.chem.qmw.ac.uk/iupac/AminoAcid/

- Используют обозначения Sec и U как трёхбуквенный 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 для фермента глутатион пероксидаза быка:

```
>gi|121664|sp|P00435|GSHC_B0VTN GLUTATHIONE PEROXIDASE
MCAAQRSAAALAAAAPRTVYAFSARPLAGGEPFNLSSLRGKVLLIENVASLUGTTVRDYTQ
MNDLQRRLGPRGLWLGFPCNQFGHQENAKNEEILNCLKYVRPGGGFEPNFMLFEKCEVNGE
KAHPLFAFLREVLPTPSDDATALMTDPKFITWSPVCRNDVSWNFEKFLVGPDGVPVRRYSR
RFLTIDIEPDIETLLSQGASA
```

Строка заголовка имеет следующие поля:

> – обязательный символ в столбце 1;

gi|121664 – это GI-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 создаёт 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** свидетельствует, что источником информации является Swiss-Prot, и что номером доступа к записи Swiss-Prot является P00435.

GSHC_B0VTN GLUTATHIONE PEROXIDASE это идентификатор Swiss-Prot для последовательности и видов, (GSHC_BOVIN), за которым следует имя молекулы.

EMBnet – European Molecular Biology net. Чтобы связать европейские лаборатории молекулярной биологии, применявшие в своих исследованиях методы биоинформатики и вычислительной биологии, в 1988 году была организована сеть. Эта сеть, получившая название EMBnet (*European Molecular Biology net*), была разработана с целью предоставления информационных и образовательных услуг сотрудникам лаборато-

рий, расположенных в различных государствах Европы, через специально выделенные узлы, работающие на местных языках.

Впоследствии организация этой сети избавила отдельные учреждения от необходимости хранить периодически обновляемые копии ряда биологических баз данных, устанавливать программы поиска, покупать дорогостоящие пакеты коммерческих программ и т. д.

Сегодня EMBnet обслуживает 34 узла. Из них 20 узлов – специально выделенные Национальные узлы. Соответствующие нации обязаны поддерживать базы данных, предоставлять программное обеспечение и сетевые услуги (анализ последовательностей, моделирование белков, создание генетических карт и т. д.), обеспечивать поддержку и обучение пользователей, а также проводить научные исследования и внедрять новые разработки. Восемь узлов EMBnet имеют специальное назначение. Это учебные, производственные или научно-исследовательские центры, которые предназначены для работы со специальными знаниями в определённых узких областях биоинформатики. В основном они ответственны за обслуживание баз данных и разработку программного обеспечения для нужд биологии.

Остальные шесть узлов были интегрированы в EMBnet как Присоединенные узлы. Это центры вычислительной биологии в неевропейских странах, которые предоставляют своим пользователям те же виды услуг, что и типичный Национальный узел. Почти все эти узлы предлагают отвечающий современному уровню доступ к базам данных и программам анализа последовательностей, наряду с разнообразными средствами молекулярного моделирования, анализа геномов, картографирования генов и т. д.

Система выборки последовательностей SRS (*Sequence Retrieval System*) является сетевым браузером баз данных молекулярной биологии. Она была разработана с целью предоставления пользователям EMBnet дополнительных сервисных услуг. Интернет-адрес SRS следующий: http://srs.ebi.ac.uk/ (рисунок 7). SRS позволяет вносить любую одноуровневую базу данных в предметный указатель любой другой базы данных.

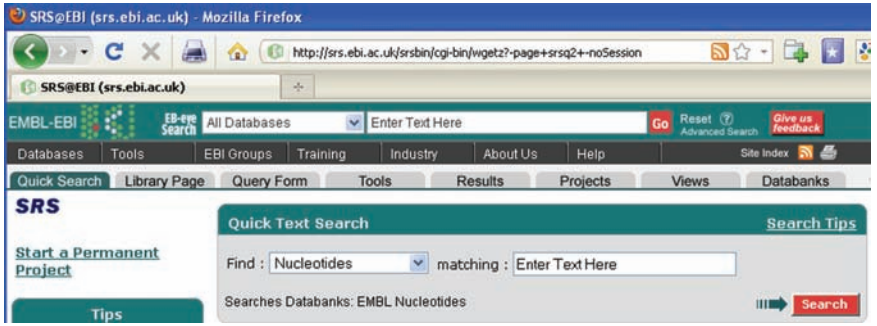


Рисунок 7 – Веб-страница SRS@EBI

Преимущество этой системы состоит в том, что производные указатели могут быть быстро найдены, что позволяет операторам выбирать, связывать ссылками и получать доступ к записям во всех ресурсах, объединенных данной системой. По своему желанию пользователь SRS может легко переопределять список подключенных баз данных.

Система выборки последовательностей связывает базы данных нуклеиновых кислот, ярлыков EST (*Expressed Sequence Tags*), белковых последовательностей, образцов белковых сверток, структур белка, а также специализированные библиографические базы данных. Таким образом, SRS представляет собой очень мощную систему, дающую пользователям возможность формулировать запросы в базы данных различных типов через единый унифицированный интерфейс, без необходимости волноваться о внутренней структуре данных, языках запросов и т. п.

SRS – интегрированная система информационного поиска во многих разнородных базах данных последовательностей и передачи выбранных последовательностей аналитическими средствами например программ сравнения и выравнивания последовательностей. В общей сложности SRS может производить поиск более чем в 140 базах данных последовательностей белков и нуклеотидов, метаболических путей, пространственных структур и функций белков, геномов, описаний болезней и фенотипов. Сюда же входят небольшие базы данных, такие как базы данных структурных мотивов белков Prosite и Blocks, базы данных факторов транскрипции и специализированные базы данных некоторых патогенов.

Помимо собственно доступа к огромному числу баз данных, SRS обеспечивает тесные связи (посредством перекрестных ссылок) между базами данных и легкость в запуске приложений. Поиск в отдельной базе данных может быть расширен до поиска в полной сети, то есть все записи, имеющие отношение к некоторому белку, могут быть легко найдены во всех содержащих их базах данных. Программы поиска подобия и построения выравниваний могут быть запускаемы непосредственно, причем без сохранения результатов запроса в промежуточном файле.

NCBI (*National Center for Biotechnology Information* – Национальный центр биотехнологической информации, НЦБИ) был основан в 1988 году в США как подразделение "Национальной медицинской библиотеки" (National Library of Medicine) и расположен в университетском городке "Национального института здоровья" (NIH), Bethesda (Бетесда), штат Мэриленд (http://www.ncbi.nlm.nih.gov/).

Задача NCBI – разработка новых информационных технологий для изучения молекулярных и генетических процессов, протекающих в здоровом и больном организме. К специальным целям относятся – создание автоматизированных систем хранения и анализа биологической информации, развитие передовых технологий машинной обработки информации, облегчение доступа пользователей к базам данных и программному обеспечению, а также координация усилий по сбору биотехнологической информации по всему миру.

Помимо этого, NCBI обслуживает GenBank – базу данных последовательностей ДНК (http://www.ncbi.nlm.nih.gov/genbank/), созданную при NIH. Группы *аннотаторов* создают записи о структуре расшифрованных последовательностей – на основании как информации из научной литературы, так и информации, представляемой самими исследователями, – и осуществляют *обмен* ими с такими *международными базами данных* нуклеотидов, как EMBL (*European Molecular Biology Laboratory*) и DDBJ (*DNA DataBank of Japan*).

Entrez. Подобно SRS для сети EMBnet, в NCBI был разработан браузер Entrez (http://www.ncbi.nlm.nih.gov/sites/gquery) с целью обеспе-

чения выборки данных молекулярной биологии (а также организации ссылок на библиографические источники) из баз данных, объединенных в NCBI: (рисунок 8).

Рисунок 8 – Веб-страница Entrez

Entrez позволяет связывать друг с другом похожие записи из разных баз данных, вне зависимости от того, есть ли между ними перекрестные ссылки. Entrez обеспечивает доступ к базам данных последовательностей ДНК (GenBank, EMBL и DDBJ), белковых последовательностей (Swiss-Prot (http://www.expasy.org/sprot/), PIR (http://pir.georgetown.edu/),

PRF (http://www.genome.jp/dbget-bin/www_bfind?prf), SeqDB, PDB, последовательностей белка, полученных трансляцией последовательностей ДНК), к базам данных картографирования генома и хромосом, трёхмерных белковых структур из PDB и также к библиографической базе данных PubMed. Подобная связь между различными базами данных – сильная сторона данной системы. Entrez можно назвать отправным пунктом для выборки последовательностей и структур из ресурсов NCBI.

Entrez – сетевая информационно-поисковая система. Она интегрирует информацию, содержащуюся во всех базах данных NCBI. Это общий внешний интерфейс для всех баз данных, поддерживаемых NCBI, и притом чрезвычайно удобный. В общей сложности Entrez имеет связь с 11 базами данных. NCBI разработал модель отношений разнородных данных, описывающих последовательности. Благодаря этому стало возможно бурное развитие программного обеспечения и интеграции баз данных, находящихся в ведении популярной информационно-поисковой системы Entrez; на этой же модели построена база данных GenBank.

К преимуществам модели следует отнести возможность легкого перехода между описанием последовательностей ДНК и кодируемых ими белков, генетическими картами хромосом и пространственными структурами соответствующих белков, а также списком опубликованной литературы, содержащей относящуюся к этим объектам информацию.

Модель данных NCBI работает непосредственно с последовательностью ДНК и последовательностью белка. Процесс трансляции представлен в виде связи между этими двумя последовательностями, а не взаимными аннотациями друг на друга. Аннотации, содержащие описание белка (например, продукты распада пептида), представлены в виде характеристик, аннотированных непосредственно на последовательности белка. Благодаря этому принципу стало очень удобно анализировать последовательности белка, полученные путём трансляции, и характеристики кодирующих последовательностей ДНК с помощью программы BLAST или любого другого средства выборки последовательностей (и притом без потери обратной связи с исходным геном). Набор, состоящий из последовательности ДНК и продуктов её трансляции, называют набором Nuc-prost.

Разработанная в NCBI модель данных описывает тип последовательности как "сегментированная последовательность". GenBank, EMBL и DDBJ представляют восстановленные сборки сегментированных последовательностей в виде *непрерывно покрытых областей* (НПО, или *контигов*). Entrez показывает такую сборку как линию, соединяющую все составляющие её последовательности.

Контиг (от англ. *contiguous* – смежный, прилегающий) – это (1) набор клонированных фрагментов ДНК, непрерывно перекрывающих в известном порядке часть генома или весь геном; (2) вид физической карты, в которой маркерами являются клонированные фрагменты.

Зеркала и Интранет. Зеркалами называются дублирующие серверы, предоставляющие услуги и информацию почему-либо недоступного основного сервера. Чтобы получить доступ к необходимому веб-узлу, нужно набрать его URL в адресной строке браузера.

Многие учебные заведения имеют "*Интранет*", то есть корпоративную локальную сеть, к которой можно подключаться только с компьютеров данного учреждения. Именно разветвленная сеть (веб) делает "Всемирную паутину" столь мощной. Для начального ознакомления рекомендуются следующие основные шлюзовые веб-узлы:

NCBI – http://www.ncbi.nlm.nih.gov/
EMBL-EBI – http://www.ebi.ac.uk/
ExPASy Proteomics Server – http://www.expasy.ch/
EMBL-Heidelberg – http://www.embl.de/

Помимо перечисленных веб-узлов, есть большое число специальных узлов, так или иначе относящихся к биологии. В поиске этих ресурсов могут быть полезны универсальные поисковые машины:

Google – http://www.google.com/
Yahoo – http://www.yahoo.com/
Alta Vista – http://www.altavista.com/
Hotbot – http://www.hotbot.com/

Контрольные вопросы

1. Запишите однобуквенные обозначения нуклеотидов и аминокислот.
2. Что такое генетический код?
3. В чём состоит парадоксальное различие протекания процессов трансляции и фолдинга белка?
4. Анализом каких трёх информационно-управляемых процессов преимущественно занимается биоинформатика?
5. Что такое World Wide Web?
6. Чем отличается системное и прикладное программное обеспечение? Приведите примеры.
7. Чем отличается Download от Upload? Какими тремя способами возможно загружать информацию из Интернета?
8. В чём сходство и различие IP-адреса компьютера и его текстовым именем в иерархической системе доменных имён?
9. Чем отличаются веб-страница и веб-узел?
10. Что такое объектная сеть?
11. Что такое интернет-браузер? Какие браузеры вы знаете?
12. Что такое гиперссылка?
13. Что такое URL и каков его формат?
14. Как записывается последовательность белка в формате FASTA?
15. Что такое GI-номер?
16. Что такое EMBnet и какой браузер используется в этой сети?
17. Что такое SRS (*Sequence Retrieval System*) и в какой сети она используется?
18. Что такое NCBI и какой браузер используется для поиска в сети баз данных NCBI?
19. Что такое Entrez и в какой сети он используется?
20. Что такое набор NUC-prost?
21. Что такое контиг?

22. Что такое интернет-зеркала?
23. Что такое Интранет?
24. Какие универсальные поисковые машины вы знаете?

4. Примеры сравнения данных

4.1. Биологическая классификация и номенклатура

Биологическая номенклатура основана на идее, что живые организмы подразделяются на виды – группы схожих организмов с одинаковым геномом.

Шведский натуралист Линней классифицировал организмы согласно иерархии: царство, тип, класс, порядок, семейство, род и вид (таблица 2).

Таблица 2 – Классификация человека и плодовой мушки

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

Современные таксономисты (классификаторы) (от англ. *taxonomist* – систематик) 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?

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 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 даёт 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 структура генома может полностью измениться.

Рибосомные РНК (рРНК) являются необходимым компонентом всех организмов.

Основываясь на анализе 16S и 18S рибосомных РНК, Карл Вёзе (Carl Richard Woese) разделил все живые организмы на *три империи*: бактерии, археи и эукариоты (рисунок 9).

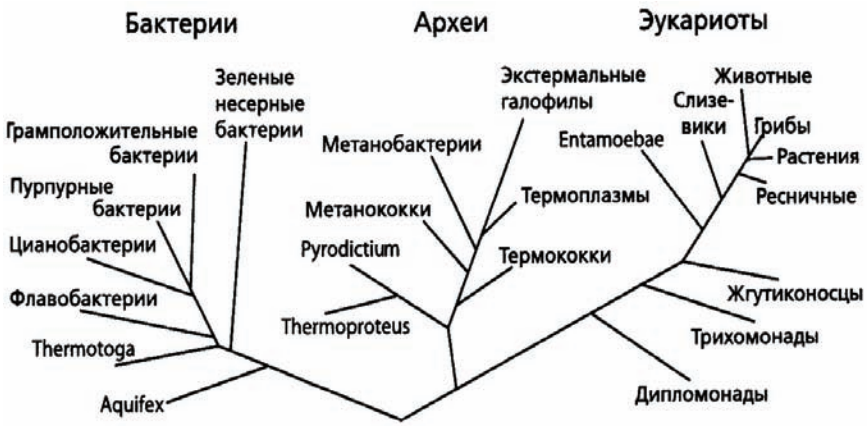


Рисунок 9 – Схема классификации живых организмов

Бактерии и археи – прокариоты, их клетки не содержат оформленного ядра. Типичные представители бактерий – микроорганизмы, являющиеся причиной многих заболеваний, а также *Escherichia coli* – главный модельный организм молекулярной биологии. В империю архей входят термофилы, галлофилы, серовосстанавливающие и метанообразующие археобактерии.

Человек относимся к эукариотам – организмам, клетки которых содержат ядро. К эукариотам относятся дрожжи, амебы, инфузории, все многоклеточные организмы и многие другие организмы.

Наиболее изучены геномы бактерий, так как это клинически важно. При этом выяснилось, что их геном относительно прост. Тем не менее, мы можем 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 9 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). Филогенетическое дерево metazoa представлено на рисунке 10.

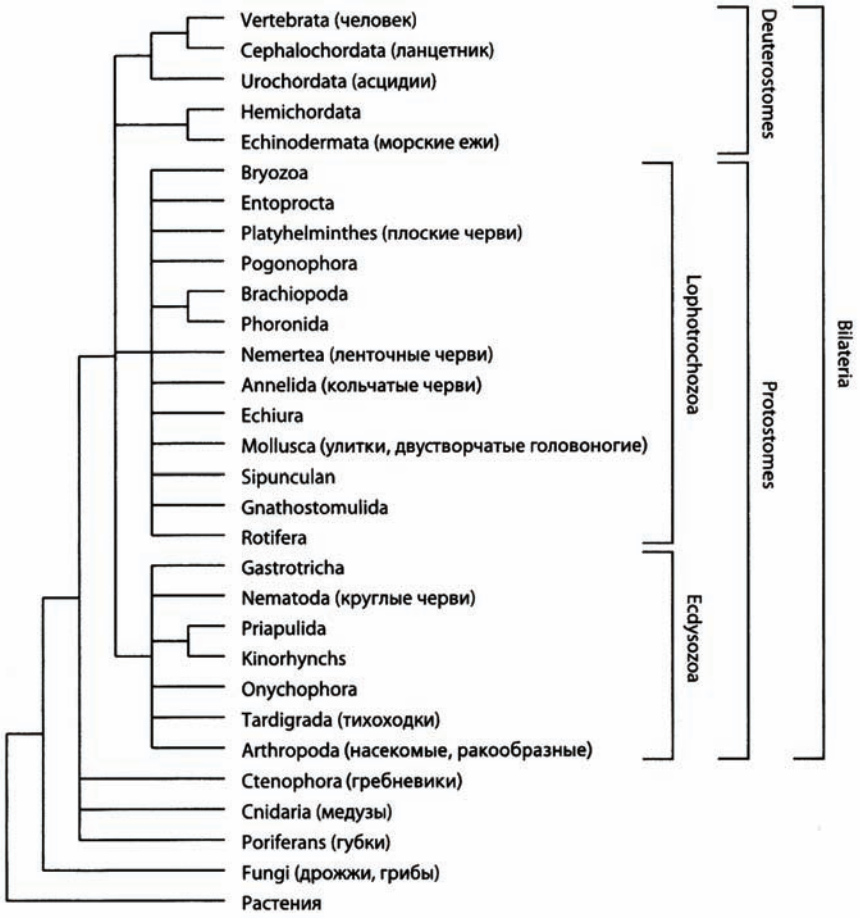


Рисунок 10 – Филогенетическое дерево metazoa (многоклеточных животных)

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 основании анализа 18S рРНК (из малой рибосомной субъединицы) и так называемых генных последовательностей HOX (большинство из которых являются гомеозисными генами (*HOmeotic compleX*) – генами, мутации которых приводят к превращению одних частей организма в другие). Морфологически Ecdysozoa имеет линяющие кутикулы – жесткий внешний слой органической материи. Lophotrochozoa имеет мягкое тело.

Человек и наши ближайшие родичи являются вторичноротыми (рисунок 11). Хордовые, включая позвоночных, и иглокожие все являются вторичноротыми.

Рисунок 11 – Филогенетическое дерево дейтростом. Хордовые, включая позвоночных, и иглокожие все являются вторичноротыми

4.2. Биологические последовательности

Информация баз данных *используется* в виде *выборок*. Выборкой называется множество последовательностей, выбранных для исследования из базы данных с помощью определённой процедуры. Накопление информации о последовательностях без обеспечения путей её извлечения делает эту информацию абсолютно бесполезной.

Однако, гораздо ценнее получить от системы *больше* знаний, чем было в нее вложено. Такой принцип получения *нового знания* может привести к биологическим открытиям. Исследователи могут совершать такого рода открытия либо обнаруживая новые отношения между различными элементами информационного содержания (которые ранее не могли быть выявлены в силу того, что эти элементы вносились в базу данных по отдельности), либо проводя над этими элементами вычисления, открывающие новое прочтение записей.

В модели данных NCBI акцент поставлен на поощрении такого рода открытий; это означает, что данные должны быть *описаны* таким образом, чтобы было возможно легко устанавливать отношения и проводить вычисления.

NCBI использует *четыре основных элемента данных*:

1) библиографические ссылки;
2) последовательности ДНК;
3) последовательности белков;
4) пространственные структуры.

В 1992 году NCBI начал присваивать *регистрационные* номера GI (*Geninfo Identifiers*) всем последовательностям, внесенным в Entrez, включая сюда же последовательности нуклеотидов из DDBJ, EMBL, GenBank, последовательности белков из транслированных ярлыков EST (*Expressed Sequence Tags*), белковые последовательности из Swiss-Prot, PIR, PRF, PDB, баз данных патентов и др.

GI присваивается в дополнение к номеру доступа, полученному в исходной базе данных. GI – простой числовой номер, иногда называемый

"номер GI". Это уникальный номер, обозначающий только одну определённую последовательность; он постоянный и по нему может быть произведен поиск и выборка последовательности.

Bioseq, или *биологическая последовательность*, является центральным элементом в модели данных NCBI. Он состоит из отдельной непрерывной молекулы нуклеиновой кислоты или белка и таким образом определяет линейную целочисленную систему координат для этой последовательности.

Seq-annot – автономный пакет аннотаций к целой последовательности Bioseq или информации, относящейся к конкретным позициям на определённых участках Bioseq.

Выравнивания последовательностей описывают отношения биологических последовательностей путём указания взаимно соответствующих частей этих последовательностей. Такое соответствие может отразить эволюционную консервативность, структурное подобие, функциональное подобие или случайное событие.

Подробнее выравнивание последовательностей будет рассмотрено в разделе "Методы биоинформационного анализа", сейчас же мы введём только основные понятия, связанные с выравниванием последовательностей.

Итак, будем считать, что выравнивание последовательностей – это установление соответствия остаток-остаток.

Мы можем искать:

- *Глобальное совпадение* – это выравнивание всей последовательности относительно другой последовательности.

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

Здесь символом "|" обозначены *соответствия*, "пробелы" обозначают *несоответствия*, "-" обозначает те *вставки* (инсерции, от англ. *insertion*) и *удаления* (делеции, от англ. *deletion*), которые

необходимо сделать в обеих последовательностях, чтобы достичь максимального количества соответствий.

- *Локальное совпадение* – это 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* – это поиск совпадения короткой последовательности в одном или более отрезках длинной последовательности. В этом случае допускается несовпадение одного символа. Можно также потребовать полного совпадения, либо допустить большее число несовпадений или даже пропусков.

Для примера, найдём мотивы «match» совпадения для строк «for the watch to babble and to talk is most tolerable» и «Any thing that's mended is but patched: virtue that transgresses is but patched with sin; and sin that amends is but patched with virtue»:

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

или:

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


- *Множественное выравнивание* – это взаимное выравнивание многих последовательностей. Например, выровняем пять строк:

```
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, шестая строка показывает символы, сохраненные во всех последовательностях выравнивания.

Рассмотрим теперь несколько реальных примеров извлечения последовательностей из банка данных и сопоставления биологических последовательностей для их анализа.

Пример 1. Получим аминокислотную последовательность панкреатической рибонуклеазы лошади.

Используем сервер UniProt (*Universal Protein Resource*) – для доступа к базам данных EMBL, GenBank, DDBJ и др.: http://www.uniprot.org/.

Введем идентификатор Swiss-Prot для панкреатической рибонуклеазы лошади RNP_HORSE в поле поиска (рисунок 12) и нажмём "Search".

Рисунок 12 – Веб-страница UniProt

Результаты поиска представлены на рисунке 13.

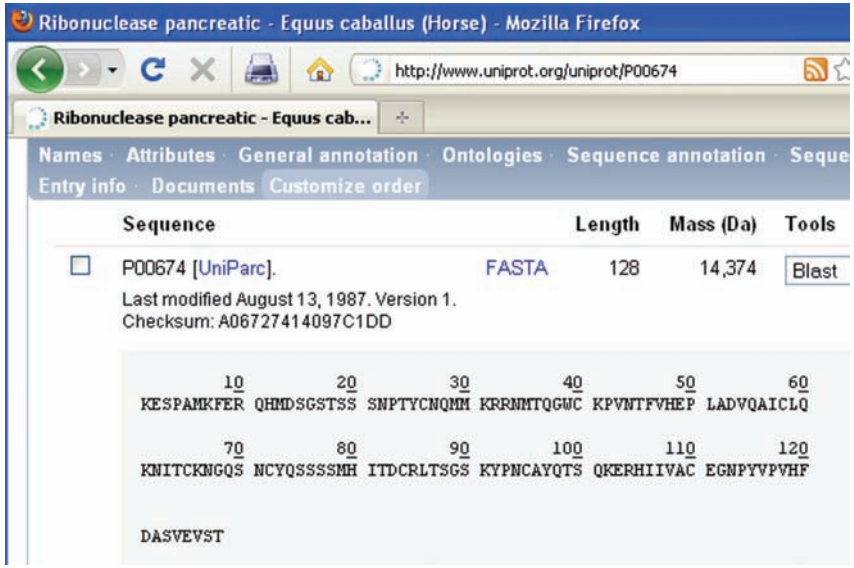


Рисунок 13 – Результаты поиска аминокислотной последовательности панкреатической рибонуклеазы лошади

Затем в этом же окне (рисунок 13) нажмём кнопку "FASTA", чтобы получить аминокислотную последовательность панкреатической рибонуклеазы лошади в FASTA-формате.

Результат показан на рисунке 14.

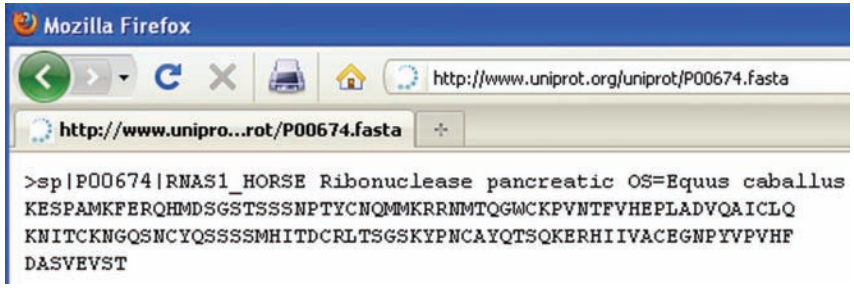


Рисунок 14 – Аминокислотная последовательность панкреатической рибонуклеазы лошади в FASTA-формате

Этот результат можно скопировать в буфер компьютера и вставить в другие программы.

Пример 2. Получим таким же методом и затем выровняем панкреатические эндонуклеазы лошади (*Equus caballus*), малого полосатика (*Balaenoptera autorostrata*) и большого рыжего кенгуру (*Macropus rufus*).

Вводим по очереди идентификаторы Swiss-Prot для панкреатических эндонуклеаз лошади RNAS1_HORSE, малого полосатика RNAS1_BALAC и большого рыжего кенгуру RNAS1_MACRU в окно поиска программы UniProt (рисунок 12).

Нажимая кнопку "FASTA" в окне результатов (рисунок 13), получаем последовательности в формате FASTA.

Затем копируем их в один текстовый (ASCII) файл.

FASTA-описание эндонуклеаз лошади (*Equus caballus*), малого полосатика (*Balaenoptera autorostrata*) и большого рыжего кенгуру (*Macropus rufus*) имеет вид:

```
>sp|P00674|RNAS1_HORSE Ribonuclease pancreatic OS=Equus caballus
GN=RNASE1 PE=1 SV=1
KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEPLADVQAICLQ
KNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTSQKERHIIVACEGNPYVPVHF
DASVEVST

>sp|P00673|RNAS1_BALAC Ribonuclease pancreatic OS=Balaenoptera
acutorostrata GN=RNASE1 PE=1 SV=1
RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHESLEDVKAVCSQ
KNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTSQKEKHIIVACEGNPYVPVHF
DNSV

>sp|P00686|RNAS1_MACRU Ribonuclease pancreatic OS=Macropus rufus
GN=RNASE1 PE=1 SV=1
ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEPKSVVDAVCHQE
NVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETSNLNKQIIVACEGQYVPVHFDA
YV
```

Построим для этих последовательностей множественное выравнивание с помощью программы ClustalW2 (рисунок 15)

http://www.ebi.ac.uk/Tools/clustalw2/index.html

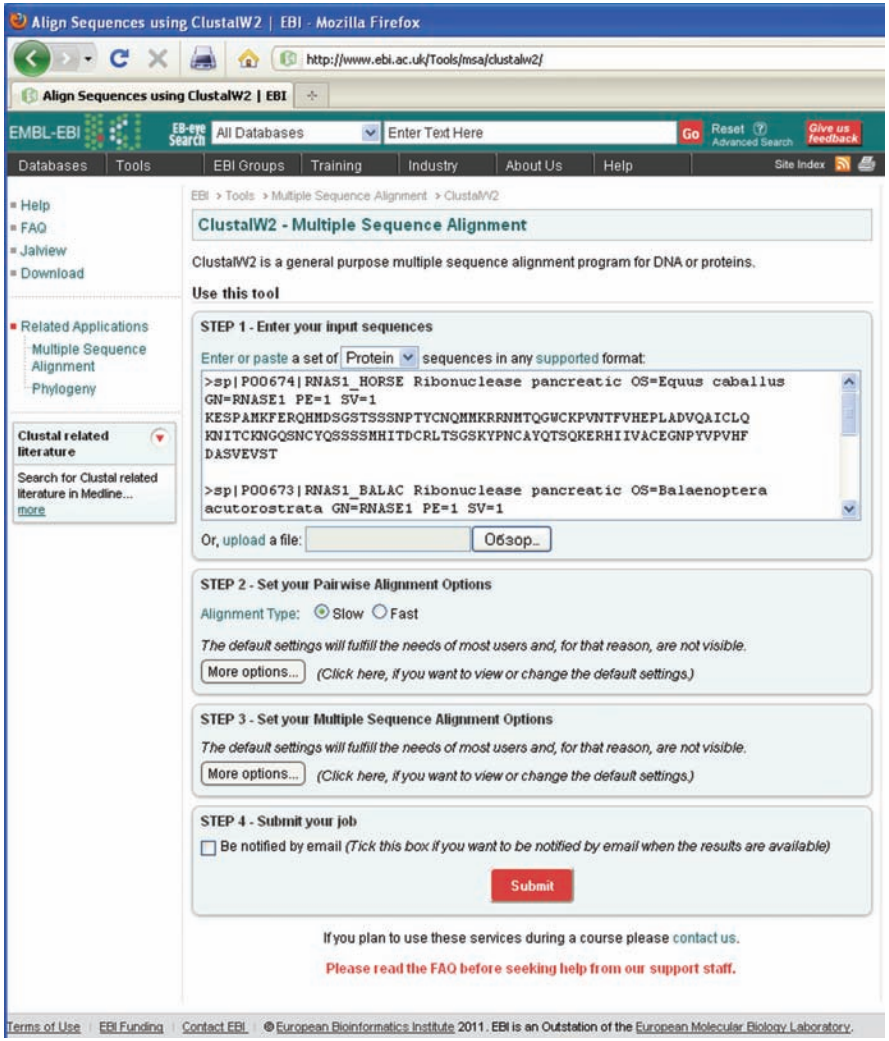


Рисунок 15 – Окно программы ClustalW2 с загруженными параметрами задачи

Результат множественного выравнивания последовательностей RNAS1_HORSE, RNAS1_BALAC и RNAS1_MACRU представлен на рисунке 16.

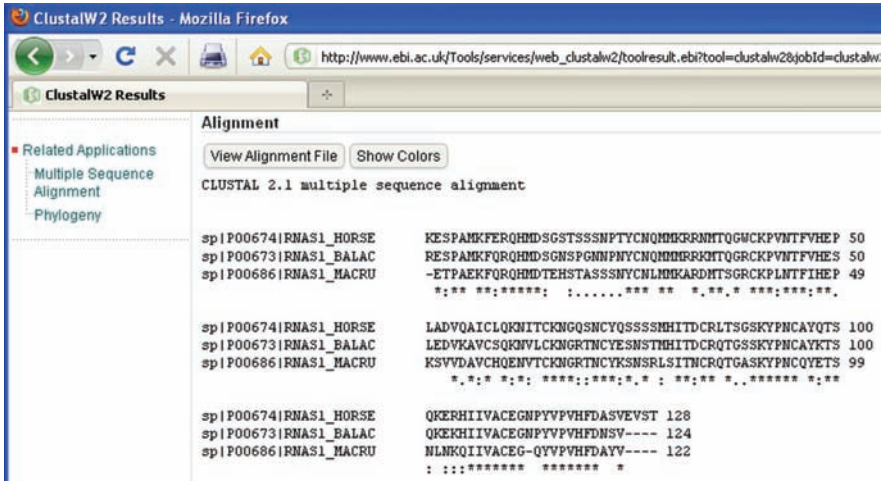


Рисунок 16 – Результат множественного выравнивания аминокислотных последовательностей панкреатических эндонуклеаз лошади, малого полосатика и большого рыжего кенгуру

В этой таблице в нижней строке под последовательностями символами обозначают:

"*" – неизменная (одинаковая во всех последовательностях) аминокислота;

":" – очень сходные по физико-химическим параметрам аминокислоты;

"." – просто сходные по физико-химическим параметрам аминокислоты;

" " – "пробел" означает отсутствие сходства.

Символами "-" в аминокислотных последовательностях показаны вставки, автоматически добавленные программой для оптимального выравнивания. Большие фрагменты последовательностей идентичны. Есть большое число замещений, но только одно внутреннее удаление.

Если нажать кнопку "Show Colors" в окне программы (рисунок 16), то результаты выравнивания будут представлены в цвете.

Проведём теперь попарное выравнивание последовательностей. Результаты попарного выравнивания представлены на рисунке 17:

а) лошадь и малый полосатик;

б) малый полосатик и большой коричневый кенгуру;

в) лошадь и большой коричневый кенгуру.

```
sp|P00674|RNAS1_HORSE      KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEP 50
sp|P00673|RNAS1_BALAC      RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHES 50
                           :*******:*******.:...**.****** **:**** **********.

sp|P00674|RNAS1_HORSE      LADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTS 100
sp|P00673|RNAS1_BALAC      LEDVKAVCSQKNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTS 100
                           * **:*:* ***: ****::***:*.*:******* *..********:**

sp|P00674|RNAS1_HORSE      QKERHIIVACEGNPYVPVHFDASVEVST 128
sp|P00673|RNAS1_BALAC      QKEKHIIVACEGNPYVPVHFDNSV---- 124
                           ***:****************** **
```

а

```
sp|P00673|RNAS1_BALAC      RESPAMKFQRQHMDSGNSPGNNPNYCNQMMMRRKMTQGRCKPVNTFVHES 50
sp|P00686|RNAS1_MACRU      -ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEP 49
                            *:** ********: :*.....**** **  *.**.*****:***:**.

sp|P00673|RNAS1_BALAC      LEDVKAVCSQKNVLCKNGRTNCYESNSTMHITDCRQTGSSKYPNCAYKTS 100
sp|P00686|RNAS1_MACRU      KSVVDAVCHQENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETS 99
                           . *.*** *:** **********:*** : **:*****:****** *:**

sp|P00673|RNAS1_BALAC      QKEKHIIVACEGNPYVPVHFDNSV 124
sp|P00686|RNAS1_MACRU      NLNKQIIVACEG-QYVPVHFDAYV 122
                           : :*:*******  *******  *
```

б

```
sp|P00674|RNAS1_HORSE      KESPAMKFERQHMDSGSTSSSNPTYCNQMMKRRNMTQGWCKPVNTFVHEP 50
sp|P00686|RNAS1_MACRU      -ETPAEKFQRQHMDTEHSTASSSNYCNLMMKARDMTSGRCKPLNTFIHEP 49
                            *:** **:*****:  :::*...*** *** *:**.* ***:***:***

sp|P00674|RNAS1_HORSE      LADVQAICLQKNITCKNGQSNCYQSSSSMHITDCRLTSGSKYPNCAYQTS 100
sp|P00686|RNAS1_MACRU      KSVVDAVCHQENVTCKNGRTNCYKSNSRLSITNCRQTGASKYPNCQYETS 99
                            : *:*:* *:*:*****::***:*.* : **:** *..****** *:**

sp|P00674|RNAS1_HORSE      QKERHIIVACEGNPYVPVHFDASVEVST 128
sp|P00686|RNAS1_MACRU      NLNKQIIVACEG-QYVPVHFDAYV---- 122
                           : :::*******  ******** *
```

в

Рисунок 17 – Результаты попарного выравнивания аминокислотных последовательностей панкреатических эндонуклеаз лошади, малого полосатика и большого рыжего кенгуру

При попарном сравнивании последовательностей, число идентичных остатков между парами в этом выравнивании представлено в таблице 3.

Таблица 3 – Число идентичных остатков в последовательностях панкреатических эндонуклеаз

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. Это согласуется с тем фактом, что лошадь и кит являются плацентарными млекопитающими, а кенгуру – сумчатое.

Таким образом, даже простейший анализ структуры последовательностей с помощью выравнивания демонстрирует важность такой процедуры для оценки эволюционной близости и филогенетических взаимодействий организмов.

Пример 3. Два ныне живущих рода слонов представлены африканским слоном (*Loxodonta Africana*) и индийским слоном (*Elephas maximus*). Сравним аминокислотные последовательности митохондриального цитохрома *b* этих слонов и ископаемого сибирского шерстистого мамонта (*Mammuthus primigenius*).

Ищем аминокислотные последовательности в UniProt (рисунок 18).

Найденные идентификаторы в стандарте Swiss-Prot: CYB_LOXAF, CYB_ELEMA, CYB_MAMPR – используем для поиска последовательностей и получения последовательностей в FASTA-формате.

FASTA-описание этих цитохромов:

```
>sp|P24958|CYB_LOXAF Cytochrome b OS=Loxodonta africana GN=MT-CYB
PE=3 SV=2
MTHIRKSHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM
TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL
LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA
LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL
LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI
```

```
LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEYPYIIIGQMASILYFS
IILAFLPIAGVIENYLIK

>sp|O47885|CYB_ELEMA Cytochrome b OS=Elephas maximus GN=MT-CYB
PE=3 SV=1
MTHTRKFHPLFKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM
TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL
LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA
FHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL
LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSILI
LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEHPYIIIGQMASILYFS
IILAFLPIAGMIENYLIK

>sp|P92658|CYB_MAMPR Cytochrome b OS=Mammuthus primigenius GN=MT-
CYB PE=3 SV=3
MTHIRKSHPLLKILNKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM
TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL
LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTDLVEWIWGGFSVDKATLNRFFA
LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILFLL
LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI
LGIMPLLHTSKHRSMMLRPLSQVLFWTLATDLLMLTWIGSQPVEYPYIIIGQMASILYFS
IILAFLPIAGMIENYLIK
```

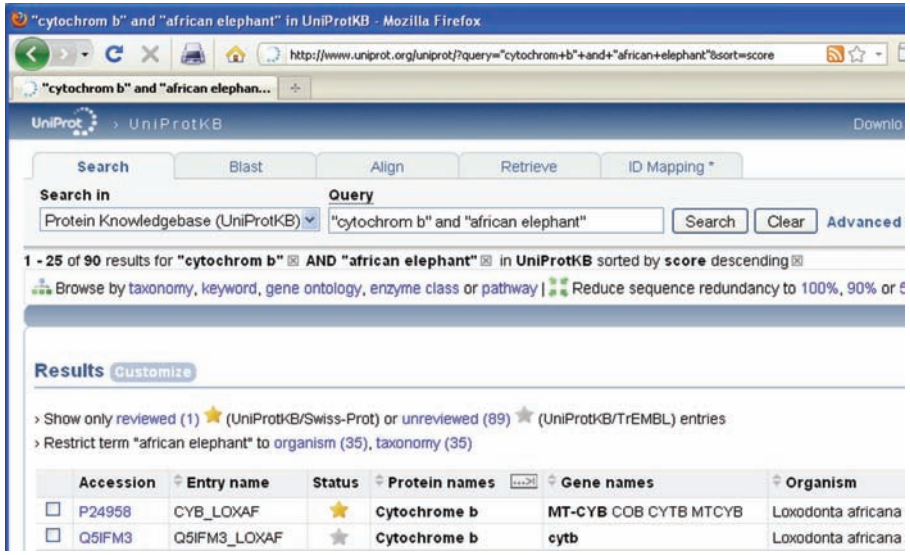


Рисунок 18 – Результаты поиска идентификатора цитохрома *b* африканского слоноа Loxodonta Africana в стандарте Swiss-Prot

Копируем эти три последовательности в окно программы ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html) и проводим выравнивание. Получаем:

```
sp|P24958|CYB_LOXAF      MTHIRKSHPLLKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60
sp|P92658|CYB_MAMPR      MTHIRKSHPLLKILNKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60
sp|O47885|CYB_ELEMA      MTHTRKFHPLFKIINKSFIDLPTPSNISTWWNFGSLLGACLITQILTGLFLAMHYTPDTM 60
                         *** ** ***:**:************************************************

sp|P24958|CYB_LOXAF      TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
sp|P92658|CYB_MAMPR      TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
sp|O47885|CYB_ELEMA      TAFSSMSHICRDVNYGWIIRQLHSNGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL 120
                         ************************************************************

sp|P24958|CYB_LOXAF      LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA 180
sp|P92658|CYB_MAMPR      LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTDLVEWIWGGFSVDKATLNRFFA 180
sp|O47885|CYB_ELEMA      LITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFA 180
                         ************************************:***********************

sp|P24958|CYB_LOXAF      LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL 240
sp|P92658|CYB_MAMPR      LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILFLL 240
sp|O47885|CYB_ELEMA      FHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLL 240
                         :*********:*****************************************************:**

sp|P24958|CYB_LOXAF      LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI 300
sp|P92658|CYB_MAMPR      LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSILI 300
sp|O47885|CYB_ELEMA      LLALLSPDMLGDPDNYMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSILI 300
                         ************************************************************:*****

sp|P24958|CYB_LOXAF      LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEYPYIIIGQMASILYFS 360
sp|P92658|CYB_MAMPR      LGIMPLLHTSKHRSMMLRPLSQVLFWTLATDLLMLTWIGSQPVEYPYIIIGQMASILYFS 360
sp|O47885|CYB_ELEMA      LGLMPLLHTSKHRSMMLRPLSQVLFWTLTMDLLTLTWIGSQPVEHPYIIIGQMASILYFS 360
                         **:*************************: *** **********:***************

sp|P24958|CYB_LOXAF      IILAFLPIAGVIENYLIK 378
sp|P92658|CYB_MAMPR      IILAFLPIAGMIENYLIK 378
sp|O47885|CYB_ELEMA      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. Возникает вопрос, являются ли такие различия существенными?

Обсудим этот пример подробнее. Мы считаем, что африканский и индийский слоны и мамонты должны быть близкими родственниками – для этого достаточно простого взгляда.

Вопрос первый – можем ли мы сказать только из этих последовательностей, что они принадлежат близким видам?

Вопрос второй – представляют ли эти малые различия *эволюционные* отклонения, возникшие из отбора, или же они есть просто *случайный* шум или *случайное* отклонение?

Необходимо иметь чувствительный статистический критерий для определения значимости совпадений и различий.

Для пояснения данных вопросов, введём *два понятия*: ***подобие*** (или сходство, *similarity*) и ***гомология*** (*homology*).

Сходство – это наличие или измерение сходства и различия, *независимо* от источника сходства.

Гомология означает, что последовательности и организмы, в которых они обнаружены, являются *потомками общего предка*, при этом предполагается, что подобные характеристики имели и предки.

О *подобии* последовательностей (или макроскопических биологических характеристик) можно судить, проведя их выравнивание, 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 путём 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?

- Возможно, 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 результате оказалось, что 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 трёх самостоятельными предков, а общее избирательное воздействие вынудило их стать похожими. (Помним, что выводы мы делаем только на основании 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.
- Существует и четвертая гипотеза о том, что все общие предки слонов и мамонтов имели сильно различающиеся цитохромы *b*, но жившие слоны и мамонты размножили общий ген путём переноса из неродственных организмов с помощью вирусов.

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

Ответить на эти вопросы только на основании данных выравнивания последовательностей невозможно.

Специалисты проводят аккуратную калибровку сходства и различия последовательностей по многим белкам из многих видов, для которых таксономическое положение было уже установлено ранее классическими методами.

В примере с панкреатическими рибонуклеазами рассуждения от сходства к гомологии оправданы.

Вопрос о том, ближе мамонты к африканским или индийским слонам, ещё не разрешен, даже используя все имеющиеся анатомические доказательства и сходство последовательностей.

В настоящее время метод анализа сходства последовательностей полностью признан и считается, что это наиболее надёжный метод 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 и эффективные инструменты для решения конкретных задач, а также многочисленные инструменты для анализа.

Но никогда машинный анализ не заменит содержательное научное обсуждение профессионалами.

4.3. Поиск схожих последовательностей в базах данных

Прежде чем проводить анализ биологических последовательностей, необходимо эти последовательности отыскать в базах данных.

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 с одной стороны *чувствителен* (который определяет даже дальнее родство), а, с другой стороны, *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 последовательностей, которые на самом деле существуют, или же он упускает большую их часть? А также, сколько из выданных этим методом результатов являются неправильными?

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, не обнаруженных последовательностей) и 200 *ложных положительных* результатов (обнаруженные последовательности в действительности не являются искомыми). 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 программа BLAST (Basic Linear Alignment Sequence Tool), которую можно использовать с сайта NCBI http://www.ncbi.nlm.nih.gov/ (рисунок 19).

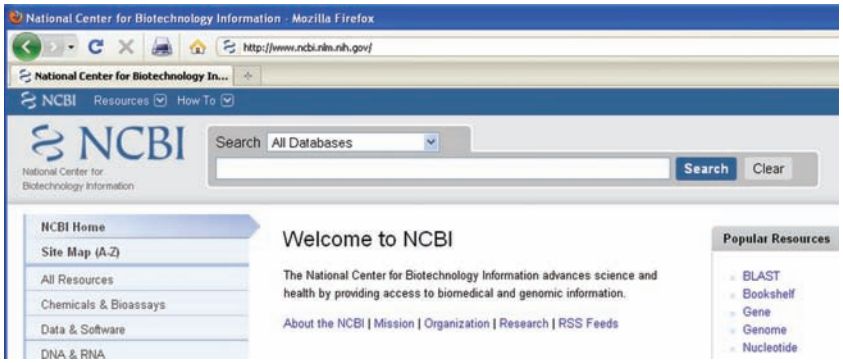


Рисунок 19 – Веб-страница NCBI; ссылка на программу BLAST показана в нижнем правом углу рисунка

Переход по ссылке "BLAST" http://blast.ncbi.nlm.nih.gov/Blast.cgi показан на рисунке 20. На рисунке отображена та часть страницы, которая относится только к основным (Basic) подпрограммам. Она включает:

- *nucleotide blast* – поиск данной последовательности нуклеотидов в базах данных нуклеиновых кислот используя алгоритмы blastn, megablast, dmegablast (discontiguous megablast);
- *protein blast* – поиск данной аминокислотной последовательности в базах данных белков используя алгоритмы blastp, psi-blast, phi-blast;

- *blastx* – переводит изучаемую нуклеотидную последовательность в кодируемые аминокислоты, а затем сравнивает её с имеющейся базой данных аминокислотных последовательностей белков;
- tblastn – изучаемая аминокислотная последовательность сравнивается с транслированными последовательностями базы данных секвенированных нуклеиновых кислот;
- tblastx – сравнивает транскрибированную нуклеотидную последовательность с транслированными последовательностями базы данных секвенированных нуклеиновых кислот.

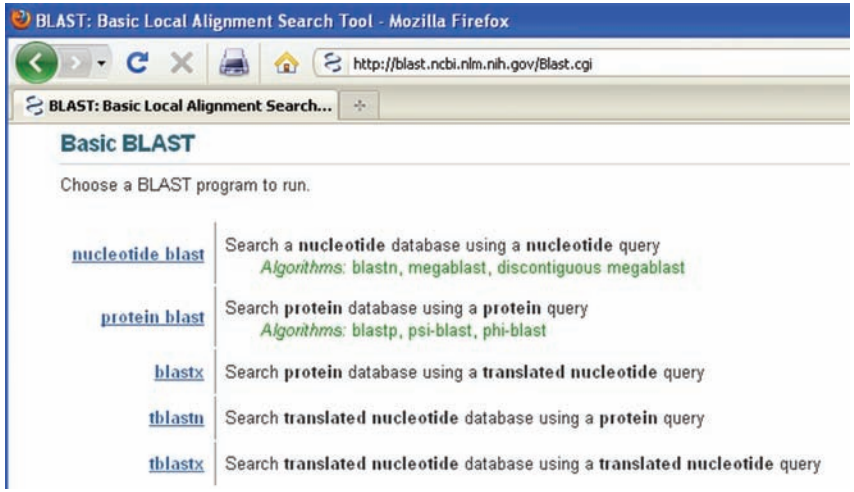


Рисунок 20 – Веб-страница программы BLAST

Здесь:

megablast – быстрое сравнение с целью поиска высоко сходных последовательностей;

dmegablast – быстрое сравнение с целью поиска дивергировавших последовательностей, обладающих незначительным сходством;

blastn – медленное сравнение с целью поиска всех сходных нуклеотидных последовательностей;

blastp – медленное сравнение с целью поиска всех сходных белковых (*protein*) последовательностей;

psi-blast – Position-Specific Iterated BLAST – сравнение с целью поиска последовательностей, обладающих незначительным сходством;

phi-blast – Pattern Hit Initiated BLAST – поиск белков, содержащих определённый пользователем паттерн.

Паттерн – (от англ. *pattern* – образец, шаблон, модель) – это либо фрагмент последовательности, либо (реже) некий стандартный набор процедур, применяемый к разным объектам.

Пример 4. Гомологи PAX-6 гена человека.

PAX-6 гены контролируют развитие глаза в широком наборе видов.

Глаза человека, мухи и осьминога сильно различаются по строению. Ранее, принимая во внимание то конкурентное преимущество, которое даёт зрение, считалось, 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, контролирующим сложный каскад 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*-ген (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.

Ген PAX-6 человека кодирует белок, имеющий Swiss-Prot-идентификатор – P26367.

Значение этого идентификатора можно получить, если в окно поиска программы UniProt (рисунок 12) ввести "PAX-6" и нажать "Search".

Из окна программы "BLAST" (рисунок 20) запускаем "protein blast" и вводим идентификатор sp|P26367 в окно "Enter Query Sequence" (рисунок 21).

Выбираем алгоритм "PSI-BLAST" в окне "Program Selection" и запускаем поиск, нажав кнопку "BLAST" в нижнем левом углу окна (рисунок 21).

Результат поиска представляет собой огромное (в длину) окно, большую часть которого занимает список записей схожих с последовательностью, заданной для поиска, сортированный в порядке убывания статистической значимости.

Начало этого списка показано на рисунке 22.

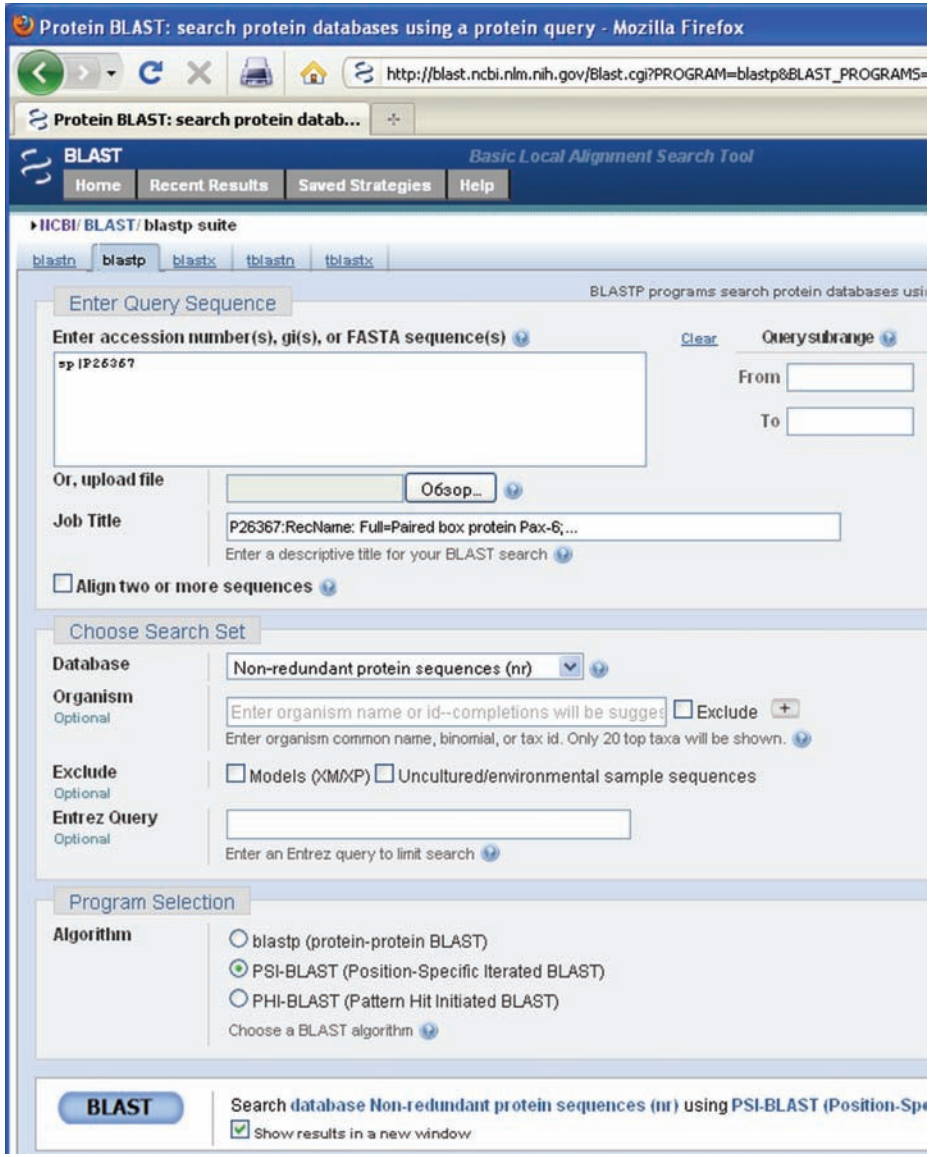


Рисунок 21 – Окно ввода программы BLAST

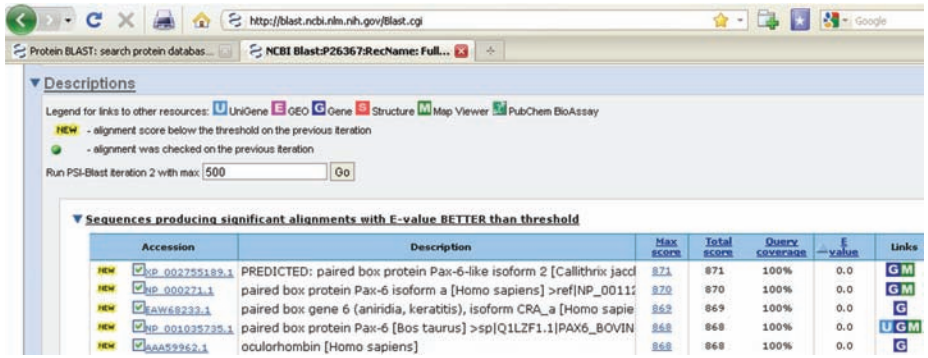


Рисунок 22 – Окно ввода программы BLAST

Каждая строка содержит одно совпадение с каким-либо геном. Рассмотрим например четвёртую от начала строку, в которой приведены результаты для:

```
paired box protein Pax-6 [Bos taurus] >sp|Q1LZF1.1|PAX6_BOVIN
RecName: Full=Paired box protein Pax-6; AltName:
Full=Oculorhombin >gb|AAI16039.1| Paired box 6 [Bos taurus]
>gb|DAA21851.1| paired box protein Pax-6 [Bos taurus]
```

В первом столбце "Accesion" располагается идентификатор гена (NP_001035735.1). Это гомолог Paired box protein Pax-6 [Bos taurus]. Базы данных U, G и M обозначены в последнем столбце, в данном случае это базы данных UniGene, Gene Structure и Map Viewer.

Число 868 – это количество очков, присвоенное обнаруженному совпадению.

Значимость данного совпадения (*E-value*) измерена как $E = 0.0$.

E-value (*expectation value*) определяется вероятностью того, что данная степень сходства может быть случайной:

E-value – это ожидаемое количество последовательностей, которые совпадут также или лучше чем данная, если поиск будет производиться базе данных такого же размера, но со случайными последовательностями.

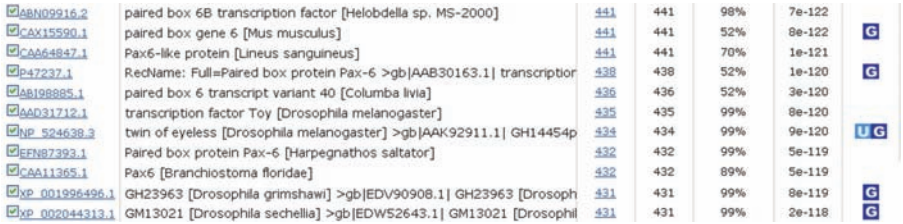
$E = 0.0$ означает полное соответствие.

Результат попарного выравнивания генов PAX-6 человека и PAX-6 буйвола (который находится в этом же окне внизу после списка совпадений) демонстрирует их абсолютное подобие (рисунок 23).

```
>ref|NP_001035735.1| UGM paired box protein Pax-6 [Bos taurus]
 sp|Q1LZF1.1|PAX6_BOVIN G RecName: Full=Paired box protein Pax-6; AltName: Ful
 gb|AAI16039.1| G Paired box 6 [Bos taurus]
 gb|DAA21851.1| G paired box protein Pax-6 [Bos taurus]
Length=422

 GENE ID: 286857 PAX6 | paired box 6 [Bos taurus] (10 or fewer PubMed links)

 Score =  868 bits (2244),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 421/422 (99%), Positives = 421/422 (99%), Gaps = 0/422 (0%)

Query  1    MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY  60
            MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY
Sbjct  1    MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY  60

Query  61   YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV  120
            YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV
Sbjct  61   YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV  120

Query  121  SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ  180
            SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ
Sbjct  121  SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ  180

Query  181  EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR  240
            EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR
Sbjct  181  EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR  240

Query  241  ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP  300
            ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP
Sbjct  241  ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP  300

Query  301  QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT  360
            QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT
Sbjct  301  QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT  360

Query  361  SPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR  420
            SPSVNGRSYDTYTPPHMQTHMNSQPMGTSG TSTGLISPGVSVPVQVPGSEPDMSQYWPR
Sbjct  361  SPSVNGRSYDTYTPPHMQTHMNSQPMGTSGATSTGLISPGVSVPVQVPGSEPDMSQYWPR  420

Query  421  LQ  422
            LQ
Sbjct  421  LQ  422
```

Рисунок 23 – Парное выравнивание генов PAX-6 человека и буйвола в окне вывода программы BLAST

Чем больше значение *E*, тем больше отклонений при сравнении последовательностей. Так, например, для гомолога twin eyeless Drosophila (NP_524638.3) значение $E = 8 \cdot 10^{-120}$ (рисунок 24(а)).



ABN09916.2	paired box 6B transcription factor [Helobdella sp. MS-2000]	441	441	98%	7e-122	
CAX15590.1	paired box gene 6 [Mus musculus]	441	441	52%	8e-122	G
CAA64847.1	Pax6-like protein [Lineus sanguineus]	441	441	70%	1e-121	
P47237.1	RecName: Full=Paired box protein Pax-6 >gb|AAB30163.1| transcriptior	438	438	52%	1e-120	G
ABI98885.1	paired box 6 transcript variant 40 [Columba livia]	436	436	52%	3e-120	
AAD31712.1	transcription factor Toy [Drosophila melanogaster]	435	435	99%	8e-120	
NP_524638.3	twin of eyeless [Drosophila melanogaster] >gb|AAK92911.1| GH14454p	434	434	99%	9e-120	UG
EFN87393.1	Paired box protein Pax-6 [Harpegnathos saltator]	432	432	99%	5e-119	
CAA11365.1	Pax6 [Branchiostoma floridae]	432	432	89%	5e-119	
XP_001996496.1	GH23963 [Drosophila grimshawi] >gb|EDV90908.1| GH23963 [Drosoph	431	431	99%	8e-119	G
XP_002044313.1	GM13021 [Drosophila sechellia] >gb|EDW52643.1| GM13021 [Drosophil	431	431	99%	2e-118	G

а

```
>ref|NP_524638.3| UG twin of eyeless [Drosophila melanogaster]
 gb|AAK92911.1|  G GH14454p [Drosophila melanogaster]
 gb|AAF59395.4|  G twin of eyeless [Drosophila melanogaster]
 gb|ACL83705.1|   toy-PA [synthetic construct]
 gb|ACL88816.1|   toy-PA [synthetic construct]
Length=543

 GENE ID: 43833 toy | twin of eyeless [Drosophila melanogaster]
(Over 10 PubMed links)

 Score =  434 bits (1117),  Expect = 9e-120, Method: Compositional matrix adjust.
 Identities = 271/521 (52%), Positives = 309/521 (59%), Gaps = 110/521 (21%)

Query  5    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  64
            HSG+NQLGGV+VNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct  30   HSGINQLGGVYVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  89

Query  65   SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN  124
            SI+PRAIGGSKPRVAT  VV KIA YKRECPSIFAWEIRDRLLSE VC +DNIPSVSSIN
Sbjct  90   SIKPRAIGGSKPRVATTPVVQKIADYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSIN  149

Query  125  RVLRNLASEKQQMGA---DGMYDKLRMLNGQTGSWGTRPGWYPGTS--------------  167
            RVLRNLAS+K+Q      + +Y+KLRM NGQTG W     WYP  +
Sbjct  150  RVLRNLASQKEQQAQQQNESVYEKLRMFNGQTGGW----AWYPSNTTTAHLTLPPAASVV  205

Query  168  -----VPGQPTQDGCQQQE----GGGENTNSISS--------NGEDSDEAQMRLQLKRKL  210
                 + GQ  +D  Q++E         +TNS  S         N    +++QMRL+LKRKL
Sbjct  206  TSPANLSGQADRDDVQKRELQFSVEVSHTNSHDSTSDGNSEHNSSGDEDSQMRLRLKRKL  265

Query  211  QRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEK  270
            QRNRTSF+ EQI++LEKEFERTHYPDVFARERLA KI LPEARIQVWFSNRRAKWRREEK
Sbjct  266  QRNRTSFSNEQIDSLEKEFERTHYPDVFARERLADKIGLPEARIQVWFSNRRAKWRREEK  325

Query  271  LRNQRRQA-----------SNTPSHIPISSSFSTS-------VYQPIPQPTTPVSSFTSG  312
            +R QRR A           +N PS    SSS +TS       V   I       S+  S
Sbjct  326  MRTQRRSADTVDGSGRTSTANNPSGTTASSSVATSNNSTPGIVNSAINVAERTSSALVSN  385

Query  313  S----------MLGRTDTALT--------------------NTYSALPPMPSFTMANNL-  341
            S          + G  +T  T                    NT  +  P P  TMA N
Sbjct  386  SLPEASNGPTVLGGEANTTHTSSESPPLQPAAPRLPLNSGFNTMYSSIPQPIATMAENYN  445

Query  342  ----PMQPPVPSQTSSYSCMLPT-----SPSVNGRSYDTYT--------PPHMQTHMNSQ  384
                 M P    Q  +Y  M        SP V+    +T          PP    + NS
Sbjct  446  SSLGSMTPSCLQQRDAYPYMFHDPLSLGSPYVSAHHRNTACNPSAAHQQPPQHGVYTNSS  505

Query  385  PMGTSGTTSTGLISPGVSVPVQVPG---SEPDMSQYWPRLQ  422
            PM +S   +TG+IS GVSVPVQ+     S+   S YWPRLQ
Sbjct  506  PMPSS---NTGVISAGVSVPVQISTQNVSDLTGSNYWPRLQ  543
```

б

Рисунок 24 – Схожие гены PAX-6 человека и twin eyeless дрозофилы: *а* – результат поиска; *б* – парное выравнивание в окне вывода программы BLAST

Парное выравнивание генов PAX-6 человека и *twin eyeless* дрозофилы показывает уже значительные различия в последовательностях (рисунок 24).

Контрольные вопросы

1. Перечислите уровни иерархии биологической номенклатуры на примере человека и плодовой мушки.
2. Какие органы называются гомологичными?
3. Чем отличается дивергентная и конвергентная эволюции?
4. На какие три империи разделил все организмы Карл Вёзе, основываясь на анализе рибосомных РНК?
5. Что называется выборкой из базы данных?
6. Какие четыре типа элементов данных использует NCBI для биологических последовательностей?
7. Что такое выравнивание последовательностей?
8. Что такое глобальное совпадение при выравнивании последовательностей?
9. Что такое локальное совпадение при выравнивании последовательностей?
10. В чём заключается поиск мотивов совпадения при выравнивании последовательностей?
11. Что такое множественное выравнивание?
12. Какими символами в окне результатов программы ClustalW2 обозначаются: одинаковая аминокислота; сходные аминокислоты; вставки; отсутствие сходства в последовательностях?
13. Чем различаются подобие и гомология последовательностей?
14. Для чего предназначена программа BLAST?
15. Что такое nucleotide blast?
16. Что такое protein blast?
17. Что такое паттерн?
18. Что такое E-value (E-значение) последовательностей?

5. Белковая информация

5.1. Структуры белков

Функциональные свойства белков определяются их *третичной* структурой – особенностями пространственной укладки полипептидной цепи белковой молекулы. При переходе от одномерных аминокислотных последовательностей к пространственным трёхмерным молекулярным структурам должны быть *адекватные изменения* и в биологической информации, описывающей такие структуры.

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* (гемоглобин) и *информационные* белки ; *регуляторные* белки, включая гормоны и рецепторные белки; белки, *контролирующие* генетическую *транскрипцию*; белки, *участвующие* в *узнавании*, включая клеточную адгезию, антитела и другие белки иммунной системы.

Белки – достаточно крупные молекулы. В большинстве случаев лишь малая часть их структуры – функциональный центр – явно выполняет специфическую функцию. Остальная часть белковой глобулы играет роль некой *инфраструктуры*, с достаточно произвольной и, в некоторых случаях, даже рыхлой, разупорядоченной структурой, служащей только для того, чтобы *точно* сформировать каталитические и сорбционные участки – *точно* расположить в пространстве компоненты этих участков.

Белки эволюционируют благодаря изменениям, вызванным мутациями в генах, которые их кодируют, что приводит к изменениям в аминокислотной последовательности. Первый принцип эволюции состоит в том, что изменения в ДНК изменяют и структуру, и функции белков, что сказывается на репродуктивной способности индивидуума, в результате чего и становится возможным естественный отбор.

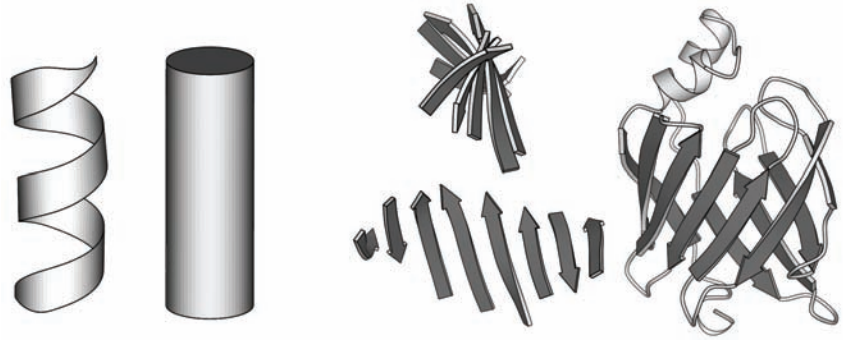
На данный момент в PDB записано более 66 000 структур белков (рисунок 1). Большая часть этой информации была получена с помощью

методов рентгеновской кристаллографии и ядерного магнитного резонанса (ЯМР, NMR). Знание пространственных особенностей укладки аминокислотной цепи позволило определить специфические функции индивидуальных белков, например, объяснить химическую каталитическую активность ферментов.

Первичной структурой белка называется последовательность расположения аминокислотных остатков в полипептидной цепи.

Полипептидная цепь определяет повороты в пространстве; направление цепи, определяющее модель изгиба.

Вторичной структурой белка называется упорядоченное строение полипептидных цепей, обусловленное внутрибелковыми водородными связями между группами C=O и N–H разных аминокислот. Наиболее устойчивыми вторичными структурами, обеспечивающими максимальное число внутрибелковых водородных связей, являются α-*спирали* и β-*структуры* (рисунок 25).



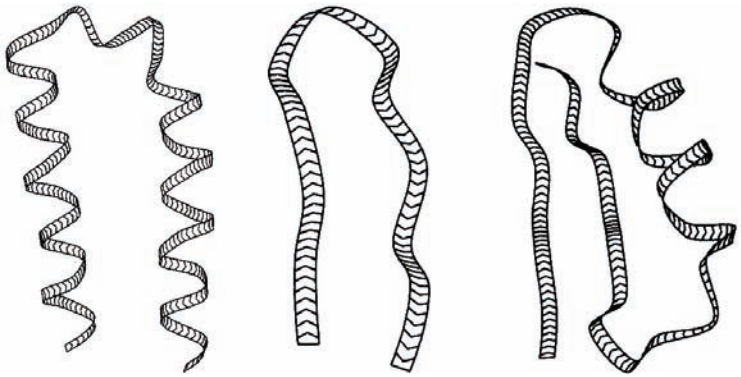
а *б*

Рисунок 25 – Схематическое изображение вторичных структур:
а – α-спирали, *б* – β-структуры

Выделяют *надвторичные* (*супервторичные*) *структуры* (*элементарные комплексы*) – термодинамически или кинетически стабильные комплексы α-спиралей и β-структур (рисунок 26).

Третичной структурой называют пространственную организацию всех α-спиралей и β-структур белка, распределение в пространстве *всех атомов* белковой молекулы.

Четвертичной структурой белка называется агрегация двух или большего числа полипептидных цепей, имеющих третичную структуру, в олигомерную функционально значимую композицию.



а *б* *в*

Рисунок 26 – Супервторичные структуры: *а* – шпилька α-спирали; *б* – β-шпилька; *в* – β-α-β-мотив. Шевроны указывают направление первичной цепи

В некоторых случаях, субъединицы могут эволюционно объединяться в единую аминокислотную цепь. В этом случае четвертичная структура переходит в третичную. Например, пять различных 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 олигомерную четвертичную структуру белка различными способами; например, глобины формируют тетрамеры в гемоглобине млекопитающих, в то время как в моллюске *Scapharca inaeguivalvis* эти же глобины образуют димеры.

Помимо четырёх основных уровней структурной организации, приведенных выше, выделяют следующие дополнительные уровни.

- *Супервторичные структуры*. В белках нередко повторяются взаимодействия между β-структурами и α-спиралями; супервторичные структуры включают шпильки α-спиралей, β-шпильки, β-α-β-мотивы (рисунок 26).
- *Домены*. Многие белки включают несколько компактных единиц в одной цепи, которые могут существовать независимо стабильно. Они называются доменами. В иерархии структур домены располагаются между супервторичными структурами и третичными структурами.
- *Модульные белки*. Модульные белки являются многодоменными беками, 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 трёх 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. Модульным белкам посвящен сайт http://www.bork.embl-heidelberg.de/Modules/ (рисунок 27). Там же приведены схемы модульных белков (например, рисунок 28) и их номенклатура.

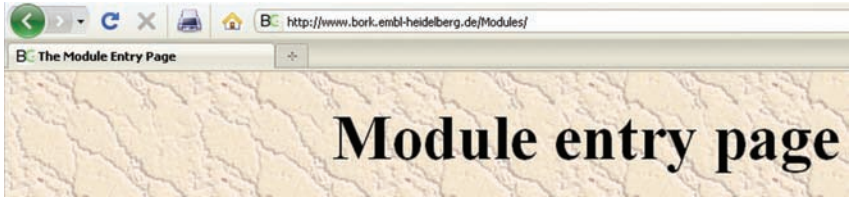


Рисунок 27 – Веб-сайт модульных белков

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

В этих достаточно широких категориях белки имеют большое разнообразие способов укладки (таблица 4).

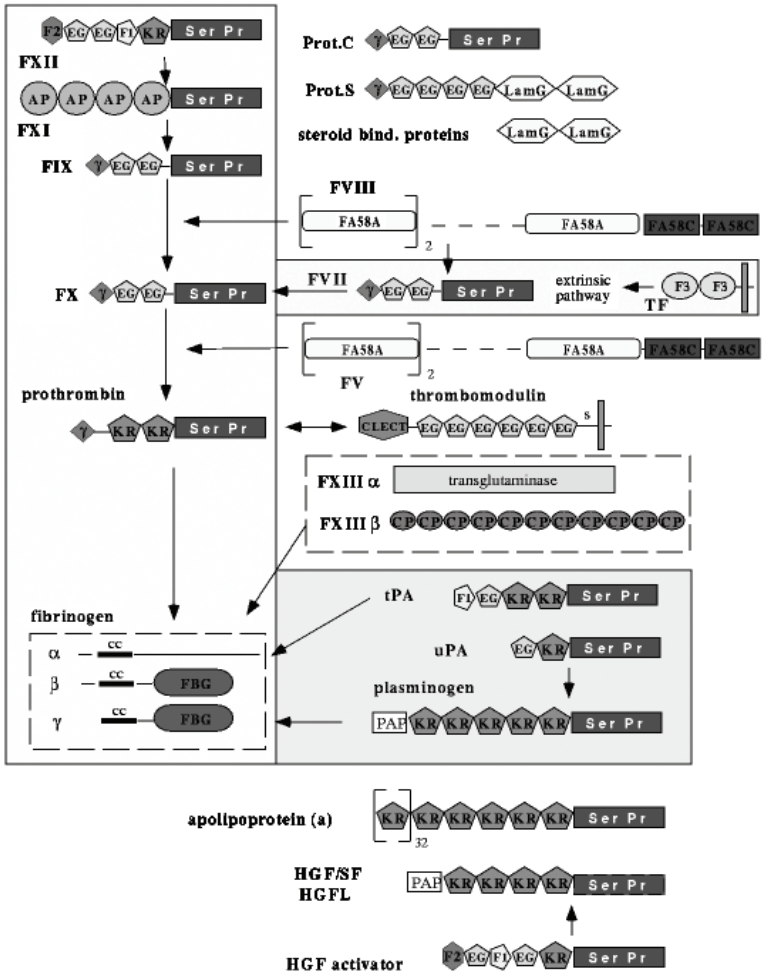


Рисунок 28 – Схемы модульных белков, обеспечивающих сворачивание крови

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.

Таблица 4 – Классы белков

Класс	Характеристика
α-спираль	вторичная структура почти исключительно содержит α-спирали
β-структура	вторичная структура почти исключительно содержит β-листы
α + β	α-спирали и β-листы находятся в разных частях молекулы; отсутствуют β-α-β-супервторичные 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
Бесструктурный	мало или нет элементов вторичной структуры

Классификация белковых структур занимает одно из центральных мест в биоинформатике, по крайней мере как мост между последовательностью и функцией. В последующих разделах мы вернемся к этой теме и опишем основные результаты и подходящие Интернет-ресурсы.

5.2. Предсказание структур белков и белковая инженерия

Аминокислотная последовательность (первичная структура) белка определяет его пространственную структуру. Если поместить белок в подходящие условия, например, такие, которые есть в цитозоле клетки, то он восстанавливает свое нативное активное состояние – происходит самопроизвольный фолдинг белка. Некоторые белки для правильного сворачивания нуждаются в помощи специальных белков – шаперонов. Но это просто ускоряет процесс, а не направляет его.

Если аминокислотная последовательность содержит достаточно информации для определения её собственной пространственной структуры, то должна существовать возможность создать алгоритм предсказания

пространственной структуры по последовательности. Однако это очень трудно. Поэтому для решения фундаментальной проблемы – предсказания структуры белков по его последовательности – исследователи ставят более простые задачи:

- *Предсказание вторичной структуры*. Какие сегменты последовательности образуют α-спирали или тяжи β-листов?
- *Распознавание фолда*. 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?
- *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%.

На рисунке 29 приведено наложение трёхмерных структур двух родственных белков: лизоцим из белка куриного яйца LYSC_CHIK и α-лактальбумин павиана LALBA_PAPCY.

Выравнивание последовательностей этих двух белков с помощью программы ClustalW2 показало, что их последовательности достаточно близкие (37% идентичных остатков в этих двух последовательностях), и, следовательно, их трёхмерные структуры очень похожи. Каждый белок мог бы послужить в качестве хорошей модели для другого настолько значительно, насколько схожи пространственные трассировки их главных (пептидных) цепей.


```
sp|P00698|LYSC_CHICK    MRSLLILVLCFLPLAALGKVFGRCELAAAMKRHGLDNYRGYSLGNWVCAA 50
sp|P12065|LALBA_PAPCY   ------------------KQFTKCELSQNLY--DIDGYGRIALPELICTM 30
                                          * * :***:  :   .:*.*   :* : :*:

sp|P00698|LYSC_CHICK    KFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPCSA 100
sp|P12065|LALBA_PAPCY   FHTSGYDTQAIVENNE-STEYGLFQISNALWCKSSQSPQSRNICDITCDK 79
                         . *.::***  .*.: **:**::**..  **:..::* ***:*:*.*.

sp|P00698|LYSC_CHICK    LLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL 147
sp|P12065|LALBA_PAPCY   FLDDDITDDIMCAKKIL-DIKGIDYWIAHKALCT-EKLEQWLCEKE- 123
                        :*..*** .: *****: * :*:: *:* :  *.  .:: *:   .
```

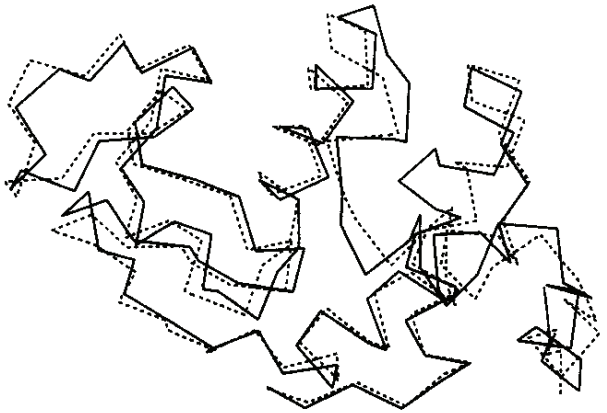


Рисунок 29 – Трассировка полипептидного остова лизоцима (из белка куриного яйца) и α-лактальбумина (павиана)

До появления генной инженерии молекулярные биологи были схожи с астрономами – они могли наблюдать исследуемые молекулярные объекты, но не могли модифицировать их. Теперь это не так. В лаборатории стало возможно *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, например, при разработке каталитических антител (абзимов). Мы можем даже пытаться создавать *новые белки*.

Большинство правил о белковой структуре было выведено благодаря наблюдениям за природными, нативными 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. Поэтому белковая инженерия сегодня обособляется в новое научное направление.

5.3. Биоинформатика в медицине

Изучение последовательности человеческого генома, равно как и геномов других организмов, может помочь в улучшении здоровья человечества. Несмотря на некоторые энергичные возражения, как правило, исходящие от неграмотных либо заинтересованных в блокировании таких работ людей, имеются следующие медицинские приложения.

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 надёжную диагностику в следующих случаях:

а) когда пациент ощущает симптомы;
б) для внутриутробной диагностики потенциальных аномалий таких, как, например, кистозный фиброз;
в) для генетической консультации пар, собирающихся завести детей;
г) заранее предупредить появления симптомов, как в тестах на наследуемые поздно-приобретаемые заболевания, такие как болезнь Хантингтона.

Болезнь Хантингтона является наследственным нейродегенеративным расстройством, которым, например, 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. Болезнь Хантингтона возможно была причиной некоторых обвинений в колдовстве. Ген не утратился у населения, поскольку болезнь проявляется в возрасте 30-50 лет, что значительно позже начала репродуктивного периода.

Прежде члены семей, затронутых болезнью Хантингтона, в молодом возрасте боялись иметь детей, они не знали, унаследовали ли эту болезнь. Открытие в 1993 г. гена, мутация в котором приводит к болезни Хантингтона, сделало возможным идентифицировать носителей заболевания. Ген содержит многократные повторы тринуклеотида CAG, кодирующие полиглутаминовые блоки в соответствующем белке. Болезнь Хантингтона – это одно из семейных нейродегенеративных расстройств, при которых наблюдаются тринуклеотидные повторы.

Чем больше блок CAG-повторов, тем ранее начнётся заболевание, и тем более серьезными будут 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.

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, может изменить уровень экспрессии или распределения по тканям. Такие аномалии могут быть обнаружены измерением белковой активности. Анализ модели белковой экспрессии является также важным путём к поиску лечения.

Генетика реакции на терапию – индивидуально специфическое лечение. Поскольку разные люди отличаются по особенностям метаболизма одного и того же лекарственного препарата, то разным пациентам в одинаковых условиях могут потребоваться *разные* дозировки. Анализ последовательности генома данного пациента позволяет индивидуально выбирать лекарства и дозировки, оптимальные для него.

Эту быстро развивающуюся область называют *фармакогеномика* (*pharmacogenomics*). Врачи теперь смогут избежать *риска* при назначении терапии для процедур, которые опасны с точки зрения побочных эффектов, часто даже фатальных, и в любом случае дорогих. Сегодня на лечение пациентов от неблагоприятных реакций на предписанные лекарства (побочные действия лекарств) в здравоохранении расходуются миллиарды долларов.

Например, лекарство 6-*меркаптопурин* является очень токсичным, хотя используется при лечении лейкемии у детей. Небольшая группа пациентов, у которых высока вероятность летального исхода из-за отсутствия фермента тиопуринметилтрансферазы, нуждается во время терапии во введении в организм дополнительных лекарственных препаратов.

В результате тестирования на этот фермент пациенты были отнесены к повышенной группе риска.

С другой стороны, теперь появилась возможность использовать такие лекарственные препараты, которые безопасны и эффективны для пациентов из повышенной группы риска, хотя ранее эти препараты были отвергнуты до или во время клинических испытаний из-за медленного действия и тяжелых побочных явлений у некоторых пациентов.

Идентификация мишеней для лекарственных веществ. *Мишень* – это белок, функциональностью которого можно управлять с помощью лекарственного препарата, для того, 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*. И только приблизительно 7% оказывает влияние на неизвестные мишени.

Растущая устойчивость бактерий к антибиотикам привела к кризису в предупреждении инфекционных заболеваний. Скорее всего, наши потомки будут вспоминать вторую половину XX века, как тот небольшой отрезок времени, когда ещё можно было контролировать бактериальные инфекции, чего не удавалось ни до, ни после.

Опираясь на анализ геномов, можно *модифицировать* существующие препараты и снизить остроту необходимости поиска новых лекарств.

Анализ генома может пригодиться при поиске мишеней для лекарственных препаратов. 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 микроорганизмов.

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.

Генная терапия. Если ген пропущен или имеет дефект, то было бы желательно *заменять* его нормальным геном или хотя бы *повысить* уровень его экспрессии – увеличивать концентрацию его продукта. И наоборот, если ген гиперактивен, бы желательно уметь *выключать* его.

Простое введение белков помогает при многих заболеваниях, из которых, наверное, наиболее известными примерами являются введение инсулина больным сахарным диабетом и введение фактора VIII при общей формы гемофилии.

Пересадка генов была успешно проделана в опытах над животными: 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 называется *антисмысловая терапия*. Идея заключается во введении ДНК или РНК, которые особым образом связываются с определённым участком гена. Присоединение к эндогенной ДНК может препятствовать транскрипции; присоединение к мРНК может препятствовать трансляции. У антисмысловой терапии есть определённые успехи в лечении таких заболеваний, как, например, цитомегаловирусный колит и болезнь Крона.

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

Контрольные вопросы

1. Что такое первичная структура белка и каковы её функции?
2. Что такое вторичная структура белка и каковы её функции?
3. Что такое супервторичная структура и каковы её функции?
4. Что такое третичная структура белка и каковы её функции?
5. Что такое четвертичная структура белка и каковы её функции?
6. Какие белки называются модульными?
7. Какую информацию получают при предсказании вторичной структуры белка?

8. Какую информацию получают при распознавании фолда ?
9. Какую информацию получают при моделировании белка по гомологии?
10. В каких случаях идентификация специфических последовательностей гена, связанного с болезнями, позволяет получить быструю и надёжную диагностику?
11. За счёт чего анализ генома данного пациента позволяет проводить индивидуально специфическое лечение?
12. Какие белки называются мишенями?
13. Каким образом анализ геномов позволяет проводить поиск мишеней?
14. Как дифференциальная геномика и сравнение профилей экспрессии белков позволяют находить мишени для лекарственных препаратов?
15. Что такое генная терапия?
16. Что такое генная заместительная терапия?
17. Что такое антисмысловая терапия?

6. Секвенирование и анализ биологических последовательностей

Достижения в области биологии и химии позволили значительно повысить скорость информационной расшифровки последовательностей генов и белков. С появлением *технологии рекомбинантных ДНК* появилась возможность относительно просто встраивать последовательности чужеродной ДНК во многие биологические системы. Кроме того, благодаря этой технологии было освоено быстрое массовое производство специфичных последовательностей ДНК – необходимых компонентов лабораторного анализа биологических последовательностей.

Технология синтеза олигонуклеотидов дала возможность исследователям конструировать необходимые короткие фрагменты ДНК из последовательностей нуклеотидов.

Во-первых, эти олигонуклеотиды могут быть использованы для зондирования обширных библиотек кДНК с целью извлечения генов, содержащих эту последовательность.

Во-вторых, эти фрагменты ДНК могут быть использованы в полимеразных цепных реакциях (ПЦР) для амплификации или модификации известных последовательностей ДНК.

Анализ биологических последовательностей проводится в случаях, когда необходимо:

а) распознать последовательности, которые кодируют белки, определяющие весь клеточный метаболизм;
б) обнаружить последовательности, которые регулируют экспрессию генов или иные клеточные процессы.

6.1. Геномика

Предметом геномики является развитие и применение методов молекулярной картографии и секвенирования, а также методов описания, расшифровки и анализа целых геномов организмов и полных наборов генных продуктов.

Под *геномом* организма понимают суммарную ДНК гаплоидного набора хромосом и каждого из внехромосомных генетических элементов, содержащуюся в отдельной клетке зародышевой линии многоклеточного организма. Анализ полных геномов даёт информацию о глобальной организации, экспрессии, регулировании и эволюции наследственных материалов (рисунок 30).

Разделяют *структурную*, *функциональную* и *сравнительную* геномику.

Структурная геномика занимается составлением *генетических* и *физических* карт, а также *расшифровкой* полных геномов.

Генетические карты служат исходным материалом для построения физических карт и карт последовательностей с более высоким разрешением и, кроме того, указывают молекулярные точки входа при клонировании генов.

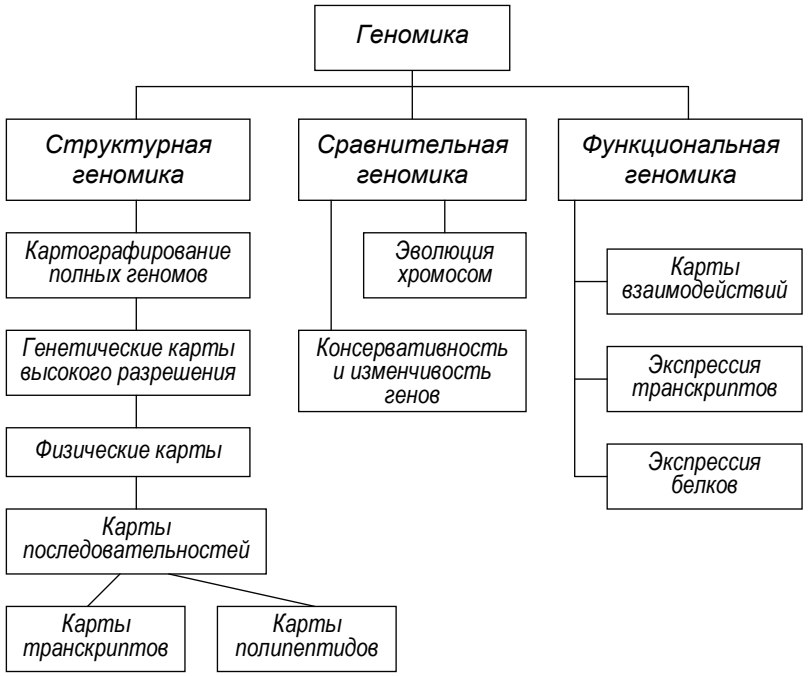


Рисунок 30 – Анализ генома: иерархическое представление

Физические карты дают представление о том, *как именно* клоны из библиотек геномных клонов распределены в целом геноме. Они обеспечивают информацию для позиционного клонирования. Последовательности ДНК генома необходимы при описании функций всех генов, включая экспрессию и регуляцию генов.

Функциональная геномика занимается общим изучением структуры, картин экспрессии, взаимодействий и регуляции молекул РНК и белков, кодируемых геномом. Это всесторонний функциональный анализ генов и не содержащих гены последовательностей, проводимый на уровне целых геномов.

Сравнительная геномика рассматривает методы сравнения полных геномов различных биологических видов с целью определения функций

каждого гена, а также об эволюционных связях организмов-носителей этих геномов.

Расшифровка полной *геномной последовательности* ДНК какого-либо организма даёт возможность распознать все гены этого организма и таким образом определить его генотип. Для выполнения обработки, анализа и описания огромного числа генов и больших количеств ДНК были изобретены специальные экспериментальные методы.

Поскольку обычные методы секвенирования могут быть применимы только к коротким отрезкам ДНК (100-1000 пар оснований), более длинные последовательности можно разделить на фрагменты, а затем собрать заново, чтобы получить полный *сиквенс* большого отрезка ДНК.

Сиквенс (от англ. *sequence* – последовательность) – это последовательность нуклеотидов в фрагменте ДНК. Для получения полного сиквенса используются два основных метода:

1) метод *прогулки по хромосоме* (*chromosome walking*), который даёт шаг за шагом *сиквенс* большого отрезка ДНК;
2) метод *дробовика* (*shotgun sequencing*), который намного быстрее, но и сложнее, так как используются случайные фрагменты ДНК, которые затем необходимо собрать вместе (с помощью специальных компьютерных программ).

Метод дробовика (*Shotgun sequencing* или шотган-секвенирование-клонирование) – метод, используемый для секвенирования длинных цепей ДНК (см. п. 6.6). Суть метода состоит в получении случайной массированной выборки клонированных фрагментов ДНК – *контигов* (*contig*, от англ. *contiguous* – смежный, прилегающий) – данного организма (то есть "дробление" генома). Затем эти контиги секвенируются обычными методами, использующими обрыв цепи (см. п. 6.4). Полученные перекрывающиеся случайные фрагменты ДНК затем *собирают* с помощью специальных программ в одну целую большую последовательность. Однако некоторую трудность при сборке могут представлять ДНК-повторы.

Анализ геномных последовательностей показывает, что каждый организм располагает определённым набором генов, необходимых для

протекания *основных* метаболических процессов (таких как размножение, гликолиз, синтез АТФ, обслуживание генетических механизмов), и также набором генов, продукты которых определяют *специфическую функцию* данного организма. Поэтому расшифровка полного генома даёт те базовые знания, на основании которых можно анализировать экспрессию генов и синтез белков, но сама по себе такая расшифровка недостаточна для определения полного набора белков организма.

Размер генома, то есть количество генетической информации на клетку, и последовательность нуклеотидов в ДНК – практически всегда постоянны для всех особей одного вида, но сильно различаются у разных видов.

В таблице 5 представлены размеры геномов некоторых организмов. Не вся ДНК кодирует белки. Кроме того, некоторые гены представлены многочисленными копиями. Поэтому число генов в геноме не может быть оценено только из размера генома.

Таблица 5 – Размер геномов

Организм	Число пар оснований	Число генов	Комментарий
Вирус ϕX-174	5386	10	вирус, инфицирующий *E. coli*
Человеческая митохондрия	16569	37	субклеточная органелла
Вирус Эпштейна-Барра (EBV)	172282	80	вызывает мононуклеоз
Mycoplasma pneumoniae	816394	680	возбудитель эпидемии циклической пневмонии
Rickettsia prowazekii	1 111 523	878	бактерия, возбудитель эпидемического тифа
Treponema pallidum	1 138 011	1039	бактерия, вызывает сифилис
Borrelia burgdorferi	1 471 725	1738	бактерия, вызывает болезнь Лайма
Aquifex aeolicus	1 551 335	1749	бактерия из горячих источников

Организм	Число пар оснований	Число генов	Комментарий
Thermoplasma acidophilum	1 564 905	1509	архея, не имеет клеточной стенки
Campylobacter jejuni	1 641 481	1708	частая причина пищевых отравлений
Helicobacter pylori	1667 867	1589	основная причина язвы желудка
Methanococcus jannaschii	1 664 970	1783	архея, термофил
Hemophilus influenzae	1 830 138	1738	бактерия, причина инфекций среднего уха
Thermotoga maritima	1 860 725	1879	морская бактерия
Archaeoglobus fulgidus	2 178 400	2437	архея
Deinococcus radiodurans	3 284 156	3187	радиационно-устойчивая бактерия
Synechocystis	3 573 470	4003	цианобактерия, сине-зеленая водоросль
Vibrio cholerae	4 033 460	3890	возбудитель холеры
Mycobacterium tuberculosis	4 411 529	4275	возбудитель туберкулеза
Bacillus subtilis	4 214 814	4779	грамположительная почвенная бактерия
Escherichia coli	4 639 221	4406	кишечная палочка
Pseudomonas aeruginosa	6 264 403	5570	прокариот
Saccharomyces cerevisiae	$12{,}1\cdot10^6$	5885	дрожжи
Caenorhabditis elegans	$95{,}5\cdot10^6$	19099	Червь
Arabidopsis thaliana	$1{,}17\cdot10^8$	25498	цветковое растение (покрытосемянное)
Drosophila melanogaster	$1{,}8\cdot10^8$	13601	плодовая мушка
Fugu rubripes	$3{,}9\cdot10^8$	30000	рыба-собака (*fugu fish*)
Человек	$3{,}2\cdot10^9$	34000	

6.2. Протеомика

Протеомика занимается каталогизацией и анализом белков с целью установления:

1) в каком состоянии (и на каком этапе жизненного цикла) клетки данный белок экспрессируется;
2) в каком количестве данный белок синтезируется;
3) с какими другими белками данный белок может взаимодействовать.

Термин "протеомика" относится ко всем белкам, экспрессируемым геномом. Это систематический анализ профилей белков, синтезирующихся в различных тканях организма.

Слово "протеом" означает белки, производимые данным биологическим видом в определённое время.

Протеом изменяется с течением времени и определяется как *совокупность белков отдельного образца или экземпляра* (*ткань*, *организм*, *клеточная культура*) *в определённый момент времени*. Протеомика отражает биологическую активность генома в динамике.

Протеомику *подразделяют* на

- *выразительную* протеомику (изучение глобальных изменений в экспрессии (выражении) белков),
- *цитокартографическую* протеомику (систематическое изучение взаимодействий между белками путём выделения белковых комплексов).

В последнее время наблюдается все возрастающий интерес к протеомике. Это связано с тем, что информация, полученная при расшифровке последовательности ДНК, отображает *статическую* генетическую информацию клетки, а жизнь клетки есть *динамический* процесс.

Состав белков, экспрессируемых организмом, изменяется во время роста, болезни и смерти клеток и тканей.

Протеомика систематизирует и описывает белки, сравнивает изменения в уровнях их экспрессии в здоровых и больных тканях, изучает их взаимодействия и определяет их функциональную роль.

Протеомика начинается с выделения функционально видоизмененного белка и кончается определением гена, ответственного за экспрессию этого белка.

Цели протеомики следующие:

1) определить все белки протеома;
2) расшифровать последовательность каждого белка и внести полученные данные в базы данных;
3) провести общий анализ уровней экспрессии белков в клетках разных типов и на разных стадиях их развития.

Протеомику разделяют на *структурную* и *функциональную* протеомику.

Структурная протеомика, или экспрессия белков, измеряет число и типы белков, присутствующих в здоровых и больных клетках.

Этот подход полезен при определении структуры клеточных белков. Некоторые из этих белков могут оказаться мишенями для новых лекарственных препаратов.

Функциональная протеомика занимается изучением биологических функций белков.

Исследование протеома можно разбить на *три* основные стадии.

1. Разделение смеси белков с помощью двумерного (2D) электрофореза в полиакриламидном геле.
2. Определение отдельных очищенных от геля белков посредством масс-спектрометрии или секвенирования с N-конца.
3. Хранение, обработка и сравнение полученных данных с помощью методов биоинформатики.

В постгеномную эру *значение* протеомики состоит в том, что именно методами протеомики исследуются экспрессия и функции генов, открытых методами геномики.

Дифференциальная демонстративная протеомика, предметом которой является сравнение уровней экспрессии белка, используется в лечении широкого спектра заболеваний.

Часто бывает довольно трудно предсказать функцию белка на основании его гомологии с другими белками или даже по его трёхмерной структуре, поэтому именно *функциональный анализ* методами протеомики определяет компоненты сложных белковых комплексов.

Протеомика играет важную роль также в *открытии* и *разработке* лекарственных препаратов – за счет описания болезненного процесса путём непосредственного отыскания наборов белков (путей или групп), совокупность которых (последовательное или одновременное действие которых) вызывает болезнь.

Протеомику можно рассматривать как основанный на принципе массового разделения *метод молекулярной биологии*, регистрирующий общее распределение белков в клетках, с целью определить и охарактеризовать отдельные целевые белки и, в конечном счете, объяснить их взаимодействия и функциональные роли.

Такой прямой анализ на белковом уровне необходим, поскольку результаты исследования генов, проводимого методами геномики, *не позволяют* предсказывать структуру белков и динамику их экспрессии. А именно на уровне белков происходит основная доля регулятивных процессов, на этом же уровне в основном протекают болезнетворные процессы и здесь же где может быть найдена большая часть мишеней для медикаментозного воздействия.

6.3. Картографирование генома

До появления технологии анализа геномов генетический базис знаний об организме обычно включал в себя *хромосомные карты* сравнительно низкого разрешения и *физические карты генов*, производящих известные мутантные фенотипы. Начиная с карт сцепления генетических признаков, составление молекулярных карт целого генома в общем случае

проходит через несколько этапов последовательного увеличения разрешения (рисунок 31).

Генетическая карта – это изображение относительных расстояний между генами, оцениваемых на основании измеренных частот рекомбинации этих генов.

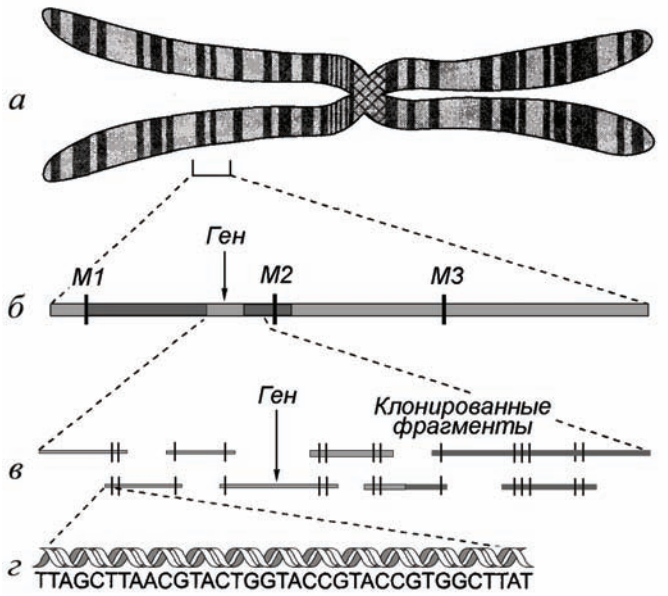


Рисунок 31 – Иерархия картографирования полного генома: *а* – цитогенетическая карта; *б* – генетическая карта высокого разрешения; *в* – физическая карта; *г* – последовательность нуклеотидов ДНК

Составление генетических карт – это процесс установления принадлежности генов к определённым хромосомам и приписывания им генетических расстояний относительно других (уже известных) генов.

Генетические карты геномов строят по данным генетических скрещиваний или, в случае человека, анализа родословной. Генетические скрещивания позволяют установить местоположения генетических маркеров на хромосомах и определить генетическое расстояние между ними.

Генетический маркер – это участок ДНК с известной локализацией. Им может служить *аллель* с известной локализацией, определяющая какой-либо признак; отличительный *морфологический признак* какой-либо хромосомы, например, перетяжка (морфологический маркер); *полиморфные* фрагменты ДНК (молекулярные маркеры). Генетические маркеры служат опорными точками для картирования генов.

Раньше в качестве маркеров для экспериментов при составлении генетических карт использовали гены.

Теперь для построения генетических карт применяют генетические маркеры другого типа – *ДНК-маркеры*. Последние представляют собой генетические маркеры, обнаруживаемые с помощью молекулярных инструментов, оперирующих с самой ДНК, а не с продуктом гена или производным фенотипом.

При составлении карты генома человека используются ДНК-маркеры *четырёх* основных типов:

1) RFLP – полиморфизм длины рестрикта (*Restriction Fragment Length Polymorphism*);
2) VNTR – переменное число тандемных повторений (*Variable Number of Tandem Repeat*), называемое также *миниспутником* (*minisatellite*);
3) STR – короткое тандемное повторение (*Short Tandem Repeat*), называемое также *микроспутником* (*microsatellite*);
4) SNP – полиморфизм отдельного нуклеотида (*Single Nucleotide Polymorphism*). С помощью микроматриц ДНК может быть выполнена одновременная печать сотен SNP.

Локусы с варьирующим числом тандемных повторов (VNTRs), (минисателлиты). VNTRs содержат участки длиной 10-100 bp, 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* (*fingerprint*, генетические отпечатки пальцев). Ранее 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.

Следует иметь ввиду, что VNTRs (*Variable Number of Tandem Repeat*) – это характеристики генома, а RFLPs (*Restriction Fragment Length Polymorphism*) – это 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 (STR) (микросателлиты) – это участки из 2–5 bp, повторенных большое число раз (10–30). Существует много преимуществ использования STRs, одно из которых – достаточно равномерное их 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.

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

Контиг, или непрерывная карта клонов – это серия перекрывающихся ДНК клонов хромосомы в известной последовательности,

хранящиеся в дрожжевых клетках в YACs (Yeast Artificial Chromosomes) и 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 bp, и весь человеческий геном можно уместить в 10 000 YACs.

В бактериальных искусственных хромосомах (BAC) ДНК вставлена в плазмиду *E. coli*.

Плазмида – это отдельная маленькая двуспиральная ДНК, являющийся дополнением к основному геному, чаще всего она кольцевая.

Бактериальная искусственная хромосома может нести до 250 000 bp.

Несмотря на то, что это меньше, чем у YAC, BACs более предпочтительны; так как они стабильнее, и с ними проще работать.

Ярлык, определённый последовательностью (*sequence tagged site*, STS) – это короткий, секвенированный участок ДНК, обычно 200-600 bp в длину, локализованный в строго определённой 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 метода 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.

6.4. Методы секвенирования ДНК

Известно несколько методов определения порядка нуклеотидов в ДНК. Один из таких методов называют *секвенированием* с обрывом цепи, или *дидезокси-секвенированием*, или же (в честь его изобретателя Фредерика Сангера (*Frederick Sanger*)) методом полимерного копирования по Сангеру. В основной реакции секвенирования участвуют следующие реагенты: однонитевая матрица ДНК; праймер для инициации полимеризации синтезируемой цепи; четыре дезоксирибонуклеозидтрифосфата, дНТФ (дАТФ, дЦТФ, дГТФ и дТТФ); четыре дидезоксинуклеозидтрифосфата, ддНТФ (ддАТФ, ддЦТФ, ддГТФ и ддТТФ); фермент ДНК-полимераза, который встраивает комплементарные нуклеотиды в растущую нить ДНК, используя матричную нить в качестве шаблона.

Секвенирование ДНК по *методу дидезокси-терминации цепи*, предложенному Сангером, начинается с *денатурации* двойной спирали ДНК-фрагмента для того, чтобы получить одиночные матричные нити для синтеза ДНК *in vitro* (рисунок 32).

Синтетический олигодезоксинкулеотид используется в качестве праймера для *четырёх* независимых *реакций* полимеризации, каждая с использованием малой концентрации *одного из четырёх* ддНТФ в дополнение к высокой концентрации нормальных дНТФ.

В каждой из реакций ддНТФ случайным образом *присоединяется* к растущей цепи ДНК в позиции соответствующего дНТФ, *прекращая* дальнейшую полимеризацию в данной позиции. В каждой из четырёх реакционных смесей синтезируется набор фрагментов ДНК разной длины: с общим началом и концами в определённых (одного и того же вида, но стоящих в разных позициях последовательности) основаниях. Полученную в каждой реакции смесь укороченных фрагментов *денатурируют и анализируют* методом гель-электрофореза (рисунок 32).

Дидезокси-метод секвенирования ДНК полностью автоматизирован. Каждую реакционную смесь метят специфической флуоресцентной меткой (или на праймере, или на субстрате одного из нуклеотидов), что позволяет определять концевые основания всех фрагментов с помощью сканера. Затем все четыре смеси реагентов объединяют в общей ёмкости и

фрагменты ДНК разделяют путём электрофореза в полиакриламидном геле (*Polyacrylamide Gel Electrophoresis*, PAGE). Меньшие фрагменты ДНК движутся быстрее, чем более крупные.

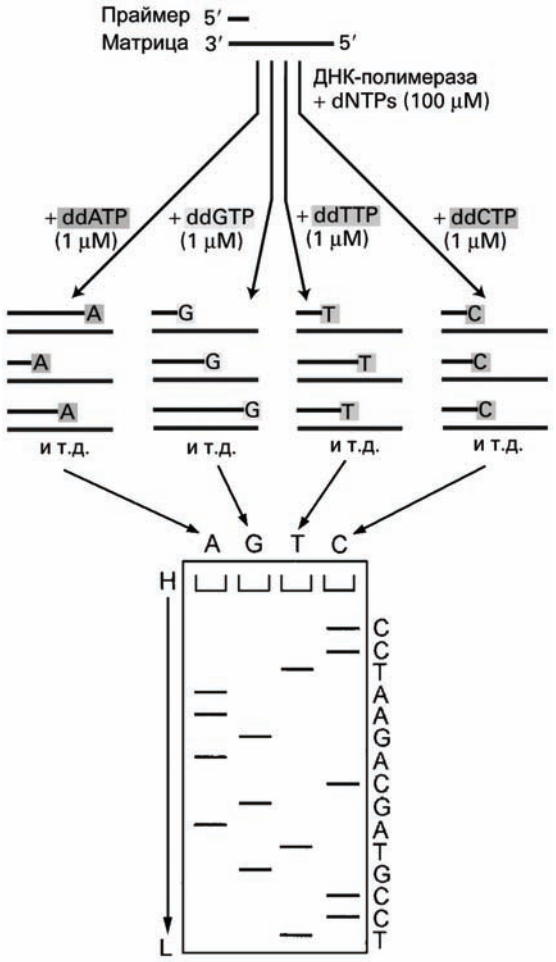


Рисунок 32 – Схема секвенирования ДНК по методу Сангера

Таким методом набор фрагментов ДНК разделяется по размеру. Разрешающая способность метода PAGE позволяет разделять полинук-

леотиды при разнице длин всего лишь в один остаток. Около конца дорожек сканер считывает флуоресцентную метку с проходящего мимо фрагмента ДНК, и эта информация преобразуется в данные сопоставления дорожек, представленные в виде графика, построенного из группы цветных пиков, соответствующих определённым основаниям (рисунок 33).

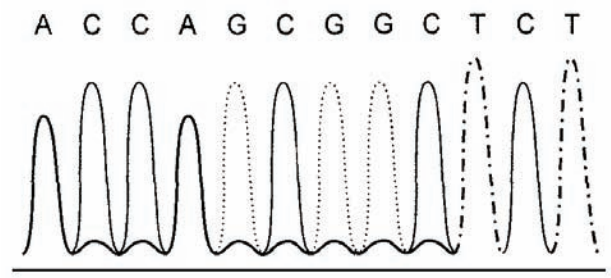


Рисунок 33 – Образец высококачественного графика сопоставления дорожек, где пики всех оснований обозначены достаточно наглядно. Пики обычно распечатываются разным цветом (на этом рисунке показаны линиями различного стиля) с целью облегчения визуальной интерпретации. Программное обеспечение типа Phred считывает пики и присваивает им качественные значения: A – толстая сплошная линия; C – тонкая сплошная; G – тонкая пунктирная; T – толстая штрих-пунктирная

Расшифрованные последовательности ДНК хранятся в базах данных. Существуют базы данных различных ДНК-последовательностей – геномной ДНК; комплементарной кДНК; рекомбинантной ДНК. Секвенирование генома выполняют с помощью метода дробовика или стратегии сборки UTR-клонов (*UnTranslated Region* – нетранслируемые области (НТО)). Для проверки качества расшифрованных последовательностей применяют многие различные программы, например: Phred, Vector_clip, CrossMatch, RepeatMaster, Phrap и Staden-Gap4.

Появление высокопроизводительной технологии автоматизированного секвенирования ДНК с флуоресцентными метками привело к быстрому накоплению информации о последовательностях; эта информация, в свою очередь, обеспечивает основу для получения данных о последовательностях белков вычислительными методами.

На анализе последовательности ДНК основываются множество видов исследований; например, к ним можно отнести: обнаружение филогенетических связей; генная инженерия и составление рестрикционных карт; определение структуры гена посредством предсказания интронов и экзонов; анализ кодирующей белок последовательности с помощью открытой рамки считывания ORF (*Open Reading Frame*) и т. д.

Согласно основной догме молекулярной биологии ДНК транскрибируется в РНК, которая затем транслируется в белок. В эукариотических системах экзоны формируют часть конечной кодирующей последовательности (КП) (*coding sequence*, CS), тогда как интроны, хотя и транскрибируются, но вырезаются механизмами сплайсинга прежде, чем мРНК принимает свою окончательную, зрелую форму (рисунок 34). Базы данных последовательностей ДНК обычно содержат информацию на уровне нетранслируемых геномных последовательностей, интронов и экзонов, мРНК, кДНК и продуктов трансляции.

Нетранслируемые области UTR встречаются как в ДНК, так и в РНК. UTR представляют собой отрезки транскрибируемой последовательности, которые с обеих сторон примыкают к кодирующей последовательности и не транслируются в белок. Нетранслируемая последовательность, особенно расположенная на 3'-конце кодирующей последовательности, весьма специфична как к самому гену, так и к биологическому виду, которому свойственно наличие этой кодирующей последовательности.

6.5. Открытая рамка считывания (ORF)

Открытыми рамками считывания называют отрезки последовательности ДНК, не прерываемые стоп-кодонами (которые привели бы к прекращению синтеза белка), и ограниченные соответствующими сигналами начала (старт-кодон) и конца трансляции (стоп-кодон). Таким образом, открытой рамкой считывания может считаться любая последовательность нуклеотидов до появления первого стоп-кодона (TGA, TAA или TAG), которая кодирует некоторое минимальное число аминокислот (около 100). Определение открытой рамки считывания у прокариотов не

представляет трудностей. У эукариот отыскание открытой рамки считывания усложнено наличием интронов.

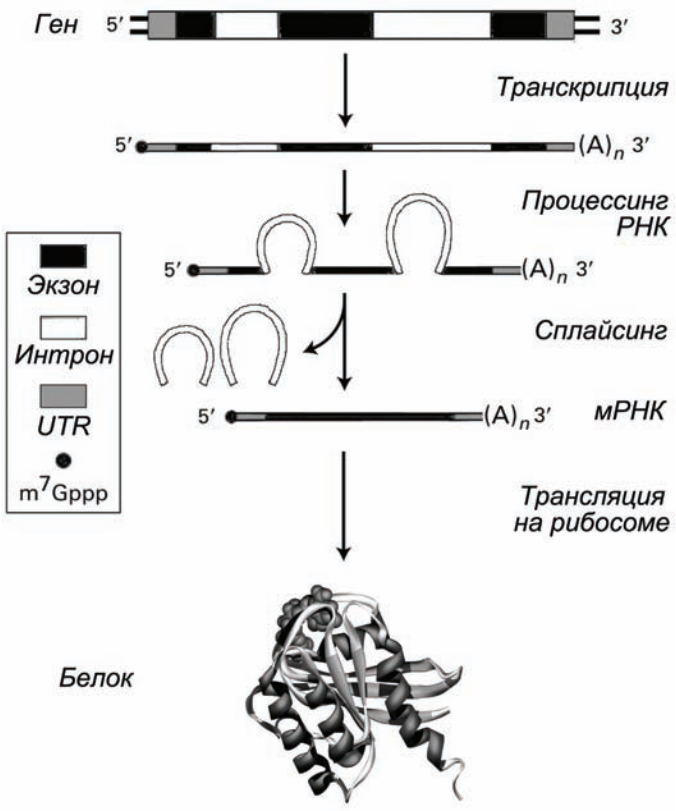


Рисунок 34 – Схема экспрессии гена

Какая рамка считывания является правильной для трансляции? Правильной рамкой считывания обычно считают *самую длинную* рамку, не прерываемую стоп-кодоном. Такую рамку называют открытой рамкой считывания (ORF). Найти конец открытой рамки считывания намного легче, чем отыскать её начало.

В качестве *индикаторов* областей ДНК, предположительно кодирующих белки, можно использовать *несколько* характеристик. Одна из таких характеристик – достаточная *длина* открытой рамки считывания. В точном определении начала кодирующей последовательности может быть полезно также распознавание примыкающих *последовательностей Козак*: (5')-ACCAUGG-(3') – специфического нуклеотидного окружения старт-кодона. "*Kozak sequence*" названа, в честь Марилин Козак, которая открыла её.

Кроме того, было установлено, что наборы используемых кодонов отличаются в кодирующих и некодирующих областях.

В частности, *частоты использования кодонов* для кодирования определённых аминокислот отличаются у организмов разных видов, а правила использования кодонов нарушаются в тех областях последовательности, которые не предназначены для трансляции.

Таким образом, статистический анализ частот использования кодонов может быть полезен для определения 5'- и 3'-UTR (а также для опознавания неправильных трансляций), потому что в этих областях наблюдается нехарактерно высокая встречаемость редко используемых кодонов.

Таблица 6 иллюстрирует значительную изменчивость в выборе кодонов, которые различные организмы используют для кодирования аминокислоты серин.

Таблица 6 – Частоты использования кодонов (в процентах) для кодирования серина, отмеченные у разнообразных опытных организмов

Кодон	*Escherichia coli*, Кишечная палочка	*Drosophila melanogaster*, Плодовая мушка	*Homo sapiens*, Человек	*Zea mays*, Кукуруза	*Saccharomyces cerevisiae*, Пивоваренные дрожжи
AGT	3	1	10	4	5
AGC	20	23	34	30	4
TCG	4	17	9	22	1
TCA	2	2	5	4	6
TCT	34	9	13	4	52
TCC	37	42	28	37	33

Для кодирования *серина* существует *шесть* возможных кодонов, которые в принципе могут использоваться с равной частотой всякий раз, когда в кодирующей последовательности определяется серин. В действительности, однако, организмы чрезвычайно *избирательны* в отношении кодонов. Отраженные в таблице 6 характерные различия в частотах встречаемости кодонов могут быть использованы в качестве дополнительного фактора в предсказании областей ДНК, предположительно кодирующих белки

Помимо характерной для каждого вида модели использования кодонов, многие организмы оказывают общее *предпочтение* нуклеотидам G или C над A или T в третьей позиции кодона (*wobble*-позиции). Закономерное отклонение частоты встречаемости нуклеотидов в этой позиции в сторону G или C также может внести вклад в предсказание ORF.

Хорошим средством опознавания ORF в области, расположенной выше старт-кодона генов прокариотов, является обнаружение *сайтов связывания рибосом* (которые помогают направлять рибосомы к правильным позициям начала трансляции).

Альтернативный сплайсинг у эукариот может привести к тому, что потенциальные продукты гена будут иметь разные длины, поскольку в конечной транскрибированной мРНК могут быть оставлены не все экзоны (хотя порядок расположения экзонов всегда сохраняется).

Если процесс редактирования мРНК приводит к трансляции полипептидов различной длины, то такие конечные белки называют *вариантами сращения* или *альтернативно сращёнными формами*. Таким образом, результаты поиска в базе данных по образцам кДНК или мРНК (информация транскрипционного уровня), обнаруживающие многочисленные пробелы в совпадениях с последовательностью запроса, могут быть следствием альтернативного сплайсинга.

6.6. Определение сиквенса клона

Клон – это скопированный фрагмент ДНК, который идентичен матрице, с которой он был получен. Процесс определения нуклеотидной

последовательности клонов позволяет выполнить анализ целой последовательности ДНК. По окончании эксперимента по клонированию некоторого гена, последовательность которого уже известна, необходимо удостовериться в том, что клонированная последовательность действительно идентична опубликованной расшифровке.

Исходный клон кДНК синтезируют с помощью матрицы мРНК. Затем этот клон секвенируют.

Расшифровка последовательностей клонов, взятых с физической карты генома, осуществляется путём сборки целого генома, секвенированного *методом дробовика*. Секвенирование целого генома методом дробовика проводят следующим образом (рисунок 35). Сначала на основании анализа уникальных перекрытий между считываниями последовательностей клонов строят отдельные контиги (рисунок 35(а)). Затем считывают участки спаренных концов контигов (рисунок 35(б)), в результате чего правильно упорядочивают и ориентируют контиги, а также перекрывают пропуски между ними и объединяют их в более крупные единицы, называемые *каркасами* (*scaffolds*) (рисунок 35(в)). .

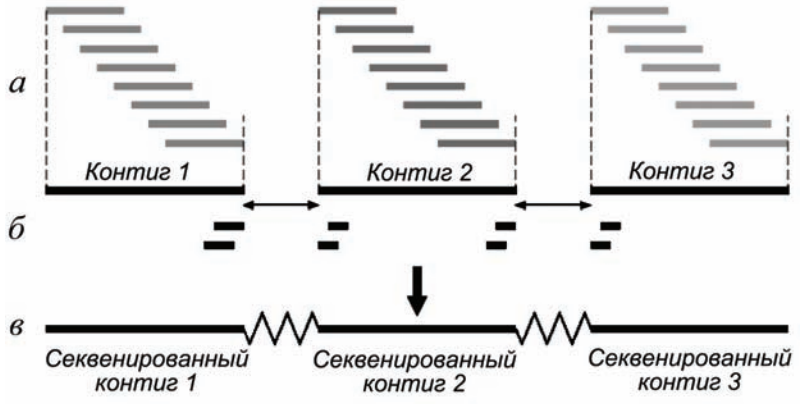


Рисунок 35 – Сборка каркаса (*scaffold*), секвенированного методом дробовика

Внедрение технологии флуоресцентного секвенирования привело к ускорению темпов накопления данных о последовательностях ДНК.

Теперь за тот же промежуток времени может быть выполнено большее число реакций секвенирования, а протоколы стали лучше отвечать условиям автоматизации. Если реакции протекают во флуоресцентном геле, то индуцированную лазером флуоресценцию непосредственно регистрирует компьютер.

Обычно гель-электрофорез проводят на 36 параллельных дорожках. Выходная информация представлена рядом закодированных цветом пиков, под которыми расположена строка знаков, обозначающих основания. Иногда интерпретирующее хроматограмму программное обеспечение не может определить, какое основание должно быть названо в определённой позиции. В таком случае появляется знак пробела "-". В конечном файле данных секвенирования такие неопределённые позиции обозначены буквой "N".

6.7. Ярлыки экспрессируемых последовательностей

Ярлык экспрессируемой последовательности EST (*Expressed Sequence Tags*) – это секвенированный отрезок последовательности клона, случайно отобранного из библиотеки кДНК, используемый для опознавания генов, экспрессируемых в определённой ткани.

Мы далеко не всегда располагаем расшифровками полных последовательностей ДНК; в основном накопленные к настоящему времени данные о ДНК состоят из отдельных отрезков последовательностей, большая часть которых представлена ярлыками экспрессируемых последовательностей (EST).

В анализе EST-последовательностей необходимо учитывать следующие моменты:

1) алфавит EST-последовательностей состоит из пяти знаков (a, c, g, t, u)
2) в последовательности могут присутствовать фантомные *всуды* (сокр. вставка/удаление – *insdel*, от англ. *insertion*/*deletion* – инсерция/делеция), приводящие к сдвигам рамки трансляции;

3) весьма вероятно, что EST-последовательность окажется подпоследовательностью какой-либо последовательности из баз данных;
4) EST-последовательность может вовсе не представлять отрезок кодирующей последовательности (КП) (*coding sequence*, CS) какого-либо гена.

Принцип секвенирования EST-последовательностей показан на рисунке 36.

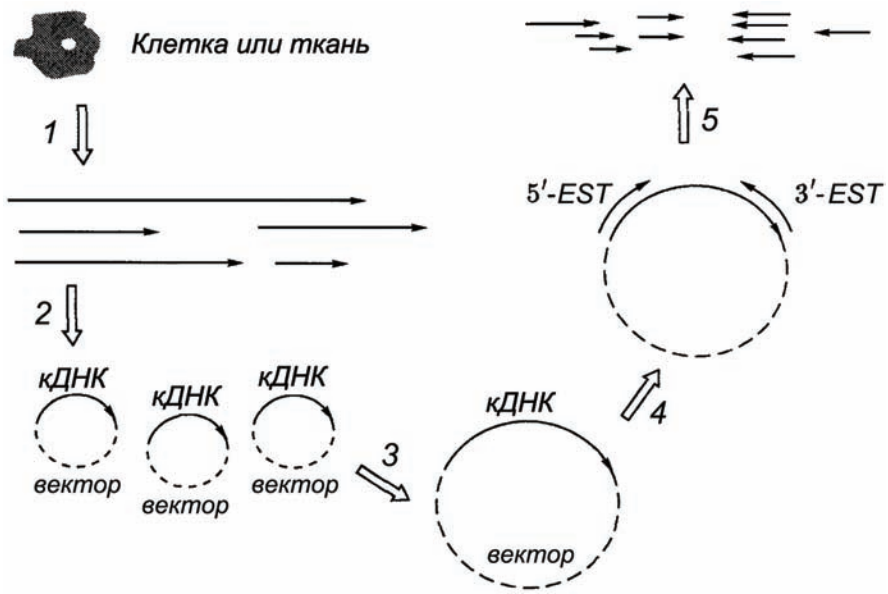


Рисунок 36 – Схема конструирования EST – ярлыков экспрессируемой последовательности: 1 – выделение мРНК и обратная транскрипция в кДНК; 2 – встраивание кДНК в вектор для размножения и создания библиотеки кДНК; 3 – отбор отдельных клонов; 4 – секвенирование 5'- и 3'-концов встроенной кДНК; 5 – помещение EST в базу данных dbEST (Database of Expressed Sequence Tags)

Из клеток интересующей ткани или клеточной линии создают библиотеку кДНК. Для этого из ткани или культуры клеток выделяют мРНК. Затем мРНК обратно транскрибируют в кДНК – обычно с помощью

праймера *олиго*-(дТ), так что один конец встройки кДНК получается транскрибированным с *поли-А* хвоста на конце мРНК. Другой конец кДНК обычно соответствует некоторому участку кодирующей последовательности или, если кодирующая последовательность коротка, – участку 5'-EST. Наконец, полученную кДНК клонируют с помощью вектора.

Отдельные клоны выбирают из библиотеки и синтезируют по одной последовательности с каждого конца встройки кДНК. Таким образом, каждый клон обычно представлен 5'-EST и 3'-EST. Поскольку EST-последовательности коротки, они обычно представляют только фрагменты генов, а не полные кодирующие последовательности. Типичный ярлык EST имеет длину от 200 до 500 нуклеотидов.

Как правило, процесс синтеза ярлыков EST в высокой степени автоматизирован и обычно предполагает использование флуоресцентной лазерной системы для считывания гелевых пленок. Для дальнейшего анализа расшифрованные последовательности загружаются в вычислительную систему.

Представляет ли такая EST-последовательность новый ген? Чтобы ответить на этот вопрос, необходимо произвести поиск в базе данных ДНК.

Если результат выравнивания *показывает существенное подобие* с некоторой последовательностью в базе данных, то нормальная процедура классификации совпадений определит, был ли найден действительно новый ген.

Если, однако, результат поиска *не показывает значительного подобия*, то мы не имеем достаточных оснований предполагать, что был обнаружен новый ген. Может оказаться и так, что данная EST-последовательность представляет некодирующую последовательность какого-либо известного гена, которую просто не успели поместить в базу данных.

Во многих мРНК (особенно у человека) на 5'- и 3'-концах кодирующей последовательности (КП) расположены длинные нетранслируемые области UTR (рисунок 34). Весьма вероятно, что данная EST-последовательность была целиком транскрибирована с одной из этих некодирующих областей. Если нам повезет, то в базе данных уже будет находиться

некоторые фрагменты нетранслируемой (некодирующей) последовательности. Если это так, то при поиске будет найдено прямое совпадение, поскольку нетранслируемые области UTR сильно консервативны и весьма специфичны к кодирующим генам.

В случае неблагоприятного исхода не будет найдено никакого совпадения, что указывает на одну из следующих двух возможностей:

1) данная EST-последовательность представляет некоторую кодирующую последовательность, для которой нет ни одной подобной последовательности в базе данных;
2) эта EST-последовательность представляет некодирующую последовательность, которая ещё не помещена в базу данных.

В интерпретации анализа EST-последовательностей важно четко различать эти две ситуации (рисунок 37).

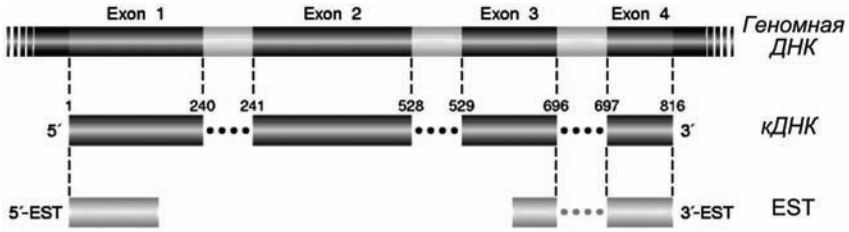


Рисунок 37 – Выравнивание полностью секвенированных последовательностей кДНК и ярлыков EST с геномной ДНК

На рисунке 37 толстые линии обозначают области выравнивания; на изображении кДНК это экзоны гена.

Точки между сегментами кДНК или EST-последовательностями обозначают области в геномной ДНК, которые не выравниваются с последовательностями кДНК или EST; это области интронов.

Числа над линией кДНК обозначают координаты (в нуклеотидах) последовательности кДНК, где нуклеотид №1 – ближайший к 5'-концу кДНК, а нуклеотид №816 – ближайший к 3'-концу кДНК.

Каждый ярлык EST представляет только короткую последовательность, считанную с 5'- или с 3'-конца соответствующей кДНК.

Таким образом, ярлыки EST устанавливают границы единиц транскрипции, но не дают никакой информации о внутренней структуре транскриптов, если только последовательности этих EST не пересекают интроны (как в случае 3'-EST, изображенной на рисунке 37).

6.8. Секвенирование белков

Современные методы секвенирования белков опираются на масс-спектрометрию – методику, позволяющую точно определить отношение массы иона к его заряду в вакууме (m/e или m/z) и по нему вычислить массу молекулы.

Структуру белка определяют путём рентгеноструктурного анализа или спектроскопии ядерного магнитного резонанса (ЯМР-спектроскопии). Рентгеноструктурный анализ заключается в восстановлении положений атомов на основании анализа дифракционной картины прохождения рентгеновских лучей через точно ориентированный кристалл белка. Рассеянные рентгеновские лучи вызывают положительную и отрицательную интерференцию и создают регулярную картину сигналов, или отражений. Результат зависит от *трёх* переменных:

1) амплитуда рассеяния,
2) фаза рассеяния (амплитуда и фаза зависят от числа электронов в каждом атоме),
3) длина волны падающих рентгеновских лучей.

Основанием метода ЯМР-спектроскопии послужил тот факт, что некоторые атомы, включая природные изотопы азота, фосфора и водорода, ведут себя подобно крошечным магнитам и изменяют свой спиновый магнитный момент в приложенном переменном магнитном поле. Эти процессы обусловлены поглощением коротковолнового электромагнитного излучения. Для определения структуры белка применяют также

некоторые другие методы: например, ЯМР-спектроскопию с магическим углом вращения и спектроскопию кругового дихроизма.

Предсказание вторичной структуры белка производят с помощью одного из *трёх* подходов:

1) эмпирические статистические методы, основанные на оценке параметров известных пространственных структур;
2) методы, опирающиеся на физико-химические критерии (такие как компактность свертки, гидрофобность, заряд, энергия водородной связи и т. д.);
3) алгоритмы предсказания, приписывающие полипептиду вторичную структуру по данным его сравнения с известными структурами гомологичных белков.

Одним из стандартных эмпирико-статистических методов является метод Чоу-Фасмена, основанный на оценке наблюдаемых в негомологичных белках конформационных предпочтений аминокислот. Однако, несмотря на то, что это "стандартный" подход, принятый во всех подобных методах, его надёжность в плане определения конформационных потенциалов аминокислот оказалась неудовлетворительной.

Что касается алгоритмов предсказания, то, напротив, за счет анализа данных множественного выравнивания последовательностей точность предсказаний в данной предметной области возрастает на несколько процентов.

Предсказание третичной структуры белка (особенно построенное на предсказанных вторичных структурах молекулы) все ещё лежит за пределом возможного современных компьютеров.

Контрольные вопросы

1. В каких случаях проводится секвенирование биологических последовательностей?
2. Чем занимается структурная геномика?
3. Чем занимается функциональная геномика?

4. Чем занимается сравнительная геномика?
5. В чём заключается метод дробовика?
6. Какие задачи решает протеомика?
7. Чем занимается структурная протеомика?
8. Чем занимается функциональная протеомика?
9. Каковы три основные стадии исследования протеома?
10. Что такое генетическая карта?
11. Что такое генетический маркер?
12. Какие четыре типа ДНК-маркеров используются при составлении карты генома человека?
13. Что такое метод полимерного копирования по Сангеру?
14. Что такое нетранслируемые области UTR?
15. Что такое открытая рамка считывания?
16. Как статистический анализ частот использования кодонов используют для определения нетранслируемых областей?
17. Как определяют сиквенс клона?
18. Что такое ярлык экспрессируемой последовательности?
19. Какие подходы используют для предсказания вторичной структуры белка?

7. Экспрессия генов

Экспрессия генов – это процесс, в котором наследственная информация от гена (последовательности нуклеотидов ДНК) преобразуется в функциональный продукт – РНК или белок. Фактически ген при экспрессии используется как своего рода план синтеза определённого белка.

Картины экспрессии гена дают ключи к раскрытию его биологической роли. Все функции клеток, тканей и органов управляются дифференциальной экспрессией генов.

Анализ экспрессии гена проводят с целью изучения его функции. Информация о том, какие гены экспрессируются в здоровых и больных тканях, позволяет определить как набор белков, характерный для нормальной функции, так и отклонения состава белков, соответствующие заболеваниям. Эти данные затем используются в разработке новых диагностических тестов различных заболеваний, а также новых лекарств, способных влиять на активность пораженных генов или белков.

Прежде экспрессию генов изучали на уровне РНК или белка, по принципу ген-за-геном, с помощью методов Нозерн- и Вестерн-блот анализа. Теперь известны способы анализа общей экспрессии, в которых все гены исследуют одновременно. Простой, но относительно дорогой методикой анализа на уровне РНК является прямая выборка последовательностей из наборов РНК, или библиотек кДНК, или даже из баз данных последовательностей.

В более совершенной методике, получившей название SAGE (Serial Analysis of Gene Expression – серийный анализ экспрессии генов, САЭГ), от каждой кДНК синтезируют очень короткие ярлыки последовательности (обычно 8-15 нуклеотидов), после чего их соединяют вместе по несколько сотен и таким образом формируют сцепку до начала секвенирования. В одной реакции секвенирования может быть получена информация об относительном содержании сотен различных мРНК. Каждый ярлык SAGE уникально обозначает каждый ген, и путём подсчета числа ярлыков могут быть определены относительные уровни экспрессии каждого гена.

7.1. Микроматрицы ДНК

Наиболее производительным является анализ экспрессии генов с помощью микроматриц ДНК.

Микроматрица ДНК – это матрица нитей ДНК (часто называемых признаками или ячейками), размещенных на миниатюрной подложке из нейлонового фильтра или предметного стекла.

Каждый элемент такой матрицы представляет собой множество тождественных нитей ДНК, представляющих определённый ген.

Явление гибридизации ДНК позволяет с помощью ДНК-зондов выделить молекулы ДНК из очень сложной смеси, например, из набора элементов целой молекулы ДНК или клеточной РНК. ДНК-зонд – это

отдельная молекула ДНК (или РНК в случае РНК-зонда), с которой ковалентно связана радиоактивная или флуоресцентная метка.

Матрицу обычно гибридизируют комплексным зондом РНК; такой зонд производят путём мечения совокупной смеси молекул РНК, полученных из клетки определённого типа. Таким образом, состав зонда отражает относительное число отдельных молекул РНК в клетке-источнике.

Если выполняется ненасыщаемая гибридизация, то интенсивность сигнала каждого элемента микроматрицы представляет относительное содержание соответствующей РНК в зонде и, следовательно, позволяет одновременно визуализировать относительные уровни экспрессии нескольких тысяч генов.

Наиболее широко применяют метод с автоматизированным нанесением отдельных клонов ДНК на подложку (покрытое специальным составом предметное стекло), например, с помощью струйного принтера (*ink-jet printing*, IJP). Такие IJP-матрицы ДНК могут иметь плотность до 5000 элементов на квадратный сантиметр. Элементы содержат молекулы двухнитевой ДНК (клоны из исследуемого генома или молекулы кДНК) до 400 bp длиной, которые должны быть денатурированы до начала гибридизации (рисунок 38).

Сначала клоны ДНК размножают и наносят на подложку, в результате чего получают микроматрицу (рисунок 38(а)). Одновременно эталонные фрагменты РНК и исследуемые фрагменты РНК образца обратно транскрибируют в ДНК-фрагменты и метят различными флуоресцентными красителями, которые высвечивают в разных областях спектра (красной и зелёной). Затем эти флуоресцентные зонды гибридизируют с ДНК на микроматрице (рисунок 38(а)).

После этого микроматрицу промывают водой, чтобы удалить негибридизировавшиеся зонды, при помощи лазерного возбуждения измеряют интенсивность флуоресценции каждого красителя на всех элементах (генах) микроматрицы и преобразуют эти данные к относительным уровням экспрессии генов в исследуемом образце по сравнению с эталоном.

В другом методе используются **ДНК-чипы**, в которых короткие олигонуклеотиды фотолитографически синтезируются *in situ* во время изготовления чипа. Такие батареи ячеек известны как *геночипы*. Они имеют плотность до 1 000 000 элементов на квадратный сантиметр, причем каждый элемент включает до 10^9 однонитевых олигонуклеотидов длиной 25 нуклеотидов.

Каждый ген на геночипе представлен *двадцатью* элементами (*двадцатью* перекрывающимися олигонуклеотидами); кроме того, для нормализации неспецифической гибридизации в него включены *двадцать контрольных несовпадений*.

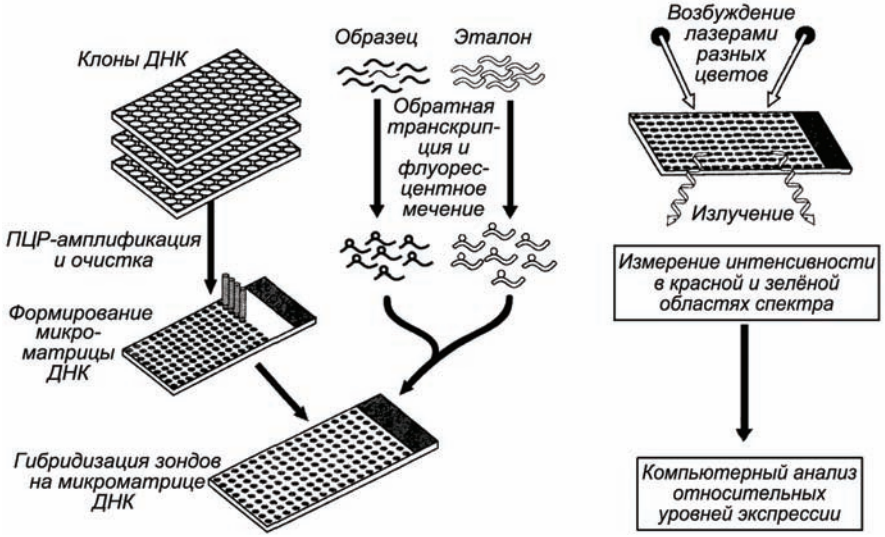


Рисунок 38 – Схема измерения дифференциальной экспрессии с помощью микроматрицы ДНК

Для скрининга матриц ДНК применяют *флуоресцентные РНК-зонды*, так как для мечения различных наборов РНК очень удобно использовать разные флуорофоры.

Флуоресцентно меченые РНК-зонды могут быть одновременно гибридизированы на одной матрице, что позволяет проводить непосредственное измерение дифференциальной экспрессии генов.

Гибридизацию геночипов проводят отдельными зондами на *двух идентичных* чипах, а интенсивности сигналов измеряют и *сравнивают* с помощью компьютера.

Исходные данные опытов на микроматрице состоят из изображений гибридизированных матриц. Точный характер изображения зависит от подложки матрицы (тип используемой матрицы). Матрицы ДНК могут содержать много тысяч элементов. Поэтому процессы сбора и анализа данных должны быть *автоматизированы*.

Программное обеспечение для предварительной обработки изображений обычно поставляется вместе со сканером. Оно позволяет определять границы отдельных пятен и измерять полную интенсивность (мощность) сигналов по яркости целых пятен. Интенсивность сигналов необходимо корректировать относительно интенсивности фона, и, кроме того, в матрицу должны быть включены эталоны для измерения неспецифической гибридизации и оценки разброса параметров гибридизации на различных матрицах.

Цель обработки данных состоит в преобразовании сигналов гибридизации в числа, которые могут быть использованы для получения матрицы экспрессии генов. Интерпретация данных гибридизации микроматрицы проводится с помощью их *группировки в кластеры* согласно подобным профилям экспрессии.

Группировкой называется способ упрощения больших наборов данных за счет объединения подобных данных в группы (*кластеры*). Для автоматизации методов анализа данных гибридизации микроматриц были разработаны различные варианты программных приложений, например:

- Ресурсы Стэнфордского университета – http://genome-www.stanford.edu/ – Stanford Microarray Database (SMD) – http://smd.stanford.edu/ – Microarray Resources : Software and Tools – http://smd.stanford.edu/resources/restech.shtml ;
- TM4 – http://www.tm4.org/ – Microarray Data Manager (MADAM), TIGR_Spotfinder, Microarray Data Analysis System (MIDAS), and Multiexperiment Viewer (MeV), as well as a Minimal Information About a Microarray Experiment (MIAME)-compliant MySQL database;

- GeneMaths XT – http://www.applied-maths.com/genemaths/genemaths.htm ;
- Eisen Lab – http://rana.lbl.gov/EisenSoftware.htm
- Microarray Image Analysis Software – http://www.statsci.org/micrarra/image.html ;
- The Gene Ontology – http://www.geneontology.org/GO.tools.microarray.shtml ;
- BioDiscovery – http://www.biodiscovery.com/ – Nexus Expression – http://www.biodiscovery.com/index/nexus-expression ;
- BxArrays – http://bioinforx.com/lims/microarray-gene-expression-data-analysis/bxarrays ;
- Array-Pro Analyzer – http://www.mediacy.com/index.aspx?page=ArrayPro ;
- Premier Biosoft – http://www.premierbiosoft.com/dnamicroarray/index.html ;
- J. Craig Venter Institute – http://www.jcvi.org/cms/research/software/ ;
- Bioconductor – http://www.bioconductor.org/ .

Микроматрицы ДНК применяют в следующих целях.

1. *Исследование состояний клеток и процессов, в них протекающих*. Анализ зависящей от состояния клетки дифференциальной экспрессии помогает в расшифровке механизмов таких процессов, как,например, образование спор или переход от аэробного метаболизма к анаэробному.
2. *Диагностика заболеваний*. Тест на присутствие мутаций может подтвердить диагноз предполагаемого генетического заболевания. Становится возможным обнаружение поздно проявляющихся симптомов, как, например, в случае болезни Хантингтона, и определение потенциально опасных для потомства генов у предполагаемых родителей (рекомендации при планировании семьи).
3. *Генетические предупредительные признаки*. Некоторые болезни не определяются исключительно генотипом, но вероятность их

развития зависит от поведения определённых генов и может быть оценена по картине их экспрессии. Осведомленный о предрасположении к той или иной болезни, человек в некоторых случаях может предупредить развитие заболевания, изменив свой образ жизни.

4. *Подбор лекарственных препаратов*. Установление генетических факторов, обусловливающих ответные реакции организма на воздействие медикаментов; у одних пациентов подобные эффекты делают лечение неэффективным, а у других даже вызывают опасные аллергические реакции.
5. *Классификация болезней*. По разным картинам экспрессии генов могут быть определены различные типы лейкемии. Знание точного типа болезни важно для подбора оптимальных методов лечения.
6. *Выбор мишени для разработки лекарства*. Белки, показывающие повышенный уровень транскрипции в определённых болезненных состояниях, могли бы быть потенциальными мишенями для фармакологического воздействия (при условии, что по другим данным будет показано, что усиленная транскрипция необходима для поддержания болезненного состояния или способствует ему).
7. *Сопротивляемость патогенам*. Сравнительный анализ генотипов или картин экспрессии у восприимчивых и стойких к антибиотику бактериальных штаммов позволяет обнаружить белки, вовлеченные в механизм сопротивляемости.

7.2. Анализ белков

Общепринятым биохимическим *методом* анализа белков, в котором белки разделяют по двум независимым параметрам: (1) по изоэлектрической точке (pI) (заряду) и (2) молекулярной массе является *двумерный электрофорез в полиакриламидном геле* (2d-PAGE, или SDS-PAGE – *Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis*).

Разделение *в первом измерении* выполняют с помощью *изоэлектрофокусировки* в неподвижном градиенте pH.

Градиент pH образуется рядом буферов, а неподвижный градиент pH создаётся ковалентным связыванием буферных групп с гелем, что предотвращает миграцию самого буфера в ходе электрофореза.

Изоэлектрофокусировка означает принудительную миграцию белков под действием электрического поля до тех пор, пока pH буфера не станет равной pI белка.

Изоэлектрическая точка белка – это величина pH, при которой белок не несет никакого избыточного заряда и поэтому не движется в приложенном электрическом поле.

По окончании миграции гель уравновешивают поверхностно-активным веществом *додецилсульфатом натрия* (ДСН, *sodium dodecyl sulfate*, SDS) (рисунок 39), который *однородно связывается* со всеми белками и *придаёт* им избыточный *отрицательный заряд*.

Благодаря этому может быть выполнено разделение *во втором измерении* – по молекулярной *массе*.

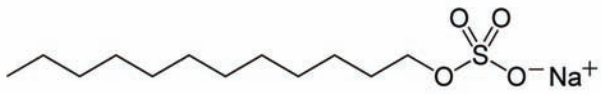


Рисунок 39 – Схема молекулы ДСН (SDS)

После разделения во втором измерении гель белка *окрашивают* универсальным красителем, чтобы *проявить* расположение всех белковых пятен.

Затем *для сравнения* уровней экспрессии белка могут быть выполнены *воспроизводимые сеансы* SDS-PAGE с подобными образцами (тканей). Таким методом получают *диагностический белковый индикатор* белка для данного образца.

На рисунке 40 представлен пример результата двумерного электрофореза в полиакриламидном геле. Каждое пятно соответствует определённому белку. Белки в образце были разделены по изоэлектрической точке (по горизонтали) и по молекулярной массе (по вертикали рисунка).

Затем окрашенный гель с белками *сканируют* и получают цифровое изображение, на котором находят и измеряют отдельные белковые пятна и корректируют интенсивность сигнала каждого пятна по интенсивности окружающего фона. Для выполнения этих операций было разработано несколько алгоритмов, основанных на Гауссовом приближении или определении лапласианов Гауссовых пятен. Пятна, морфология которых отклоняется от единой Гауссовой формы, могут быть интерпретированы с помощью *модели перекрывающихся форм*.

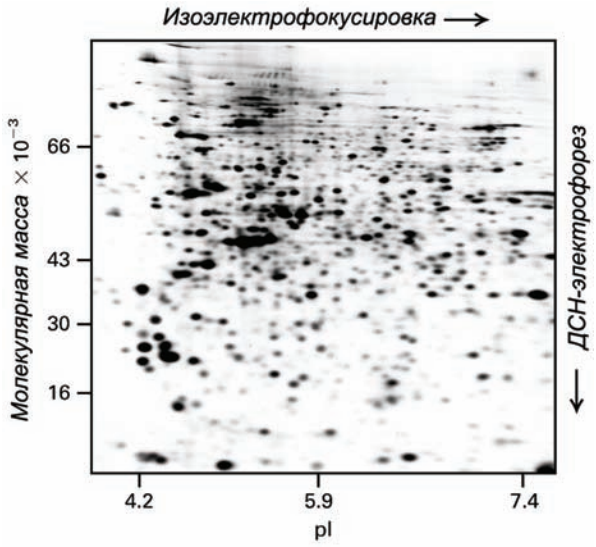


Рисунок 40 – Двумерная электрофореграмма метода SDS-PAGE

Более простой метод анализа полученного изображения – это *линейный цепной анализ*, в котором программа сканирует столбцы пикселей цифрового изображения и регистрирует пики плотности сигнала. Этот процесс повторяется для смежных столбцов пикселей, что позволяет алгоритмически определять как центры пятен, так и общую интенсивность сигнала каждого пятна.

Другой метод известен под названием "*преобразование водораздела*" (*watershed transformation*). В этом методе интенсивности пикселей

представлены в виде топографической карты, так что по ней могут быть определены холмы и долины. Этот метод полезен для отделения групп, цепей и маленьких пятен, перекрывающихся с большими (боковых пятен, или лепестков), а также для слияния областей одного пятна.

На выходе программы, опирающейся на любой из подобных методов, мы получаем *список пятен*.

Метод 2D-PAGE может быть использован также для анализа дифференциальной экспрессии белка. С его помощью можно определять белки, которые активируются или подавляются определённым курсом лечения или различными лекарствами, искать белки, связанные с теми или иными болезненными состояниями, или отслеживать изменения в экспрессии белка, происходящие в течение развития клетки или целого организма. После регистрации данные анализа экспрессии белка организуются в виде матрицы экспрессии этого белка. Результаты экспериментов 2D-PAGE хранятся в базах данных 2D-PAGE. Они могут быть найдены по адресу – WORLD-2DPAGE List Index to 2-D PAGE databases and services – http://www.expasy.ch/ch2d/2d-index.html.

7.3. Обнаружение генов

В последнее время значительные денежные средства выделяются на поиск генов, связанных с конкретными видами болезней. Цель этого поиска состоит в развитии новых методов терапии для борьбы с широким спектром распространенных функциональных и структурных расстройств, например, рака, туберкулеза, астмы и т. д.

В настоящее время есть *две* главные *стратегии* открытия белков, которые могут представлять собой молекулярные мишени, подходящие как для получения молекулярных препаратов, так и для развития генотерапии.

Одним из подходов к обнаружению связанных с болезнью генов является *метод позиционного клонирования*. Согласно этому методу изучают популяцию людей, в которой наблюдаются случаи рассматриваемого заболевания, и находят хромосому, связанную с развитием этой болезни. Затем устанавливают связь болезни с некоторой хромосомной областью,

после чего секвенируют большой отрезок хромосомы вблизи этой области (локуса) и получают последовательность ДНК длиной несколько сот тысяч пар нуклеотидов. В принципе такой локус может содержать множество генов, хотя, скорее всего, только один из них действительно вовлечен (прямо или косвенно) в болезнетворный процесс.

Для повышения эффективности распознавания генов в локусе могут быть использованы различные методы *поиска* последовательностей и *предсказания* генов, но так или иначе должны быть экспрессированы несколько генов, и для установления того, какой именно ген действительно вовлечен в болезнь, потребуется дальнейший анализ (или испытания). Хотя гены, обнаруженные этим способом, могут быть вполне удовлетворительными с академической точки зрения, они вовсе не обязательно будут хорошими мишенями для лекарственных препаратов (или точками терапевтического воздействия).

Другой подход к открытию генов, требующий намного меньших затрат на секвенирование и больше полагающийся на мощные поисковые возможности современных вычислительных систем, основан на *отыскании генов*, которые *фактически экспрессируются* в здоровых и больных тканях. Он позволяет проводить *сравнение уровней экспрессии* в двух состояниях и на основании такого анализа наиболее эффективно выбирать потенциальную мишень. Этот процесс анализирует те молекулы мРНК, которые используются для синтеза этих белков.

Как правило, в обнаружении генов участвуют следующие элементы:

- участки сращения,
- старт- и стоп-кодоны,
- точки разветвления,
- промоторы и терминаторы транскрипции,
- участки полиаденилирования,
- участки прикрепления рибосом,
- участки связывания с топоизомеразой II,
- участки расщепления топоизомеразой I,
- участки связывания с различными факторами транскрипции.

Такие локальные участки называют "*сигналами*" и обнаруживают с помощью "датчиков сигналов". Напротив, удлиненные последовательности и последовательности переменной длины (например, экзонов и интронов) называют "*содержанием*" и обнаруживают посредством "*датчиков содержания*". Наиболее сложные из применяемых датчиков сигналов – нейросети. Типичный датчик содержания – тот, который предсказывает кодирующие области.

Для определения *полной* структуры гена было создано несколько систем, комбинирующих датчики сигналов и содержания. Такие системы способны распознавать более сложные взаимозависимости между свойствами генов. Одной из первых комплексных программ поиска генов, разработанных на сегодняшнее время, является программа Genelang (http://arete.ibb.waw.pl/PL/html/gene_lang.html). Построенная на принципе динамического программирования, эта программа комбинирует отобранные экзоны и другие области или участки с назначаемым счётом и предсказывает целый ген с максимальным полным счетом.

Главная особенность динамического программирования – модель, которая содержит скрытую, или ненаблюдаемую переменную, привязанную к каждому нуклеотиду и отражающую функциональную роль или позицию этого нуклеотида. Такие модели называют *скрытыми марковскими моделями* (СММ).

Скрытая марковская модель (СММ) (*Hidden Markov model*, HMM) – это статистическая модель, имитирующая работу процесса, похожего на марковский процесс с неизвестными параметрами, и задачей ставится разгадывание неизвестных параметров на основе наблюдаемых. Полученные параметры могут быть использованы в дальнейшем анализе, например, для распознавания образов. Первые заметки о скрытых марковских моделях опубликовал Баум в 1960-х, и уже в 70-х их впервые применили при распознавании речи. С середины 1980-х СММ применяются при анализе биологических последовательностей, в частности, ДНК.

Марковский процесс – это случайный процесс, эволюция которого после любого заданного значения временного параметра t не зависит от эволюции, предшествовавшей t, при условии, что значение процесса в

этот момент фиксировано ("будущее" процесса не зависит от "прошлого" при известном "настоящем"; или по определению Вентцеля: "будущее" процесса зависит от "прошлого" лишь через "настоящее").

Марков, Андрей Андреевич (1856–1922) – выдающийся русский математик, внёсший большой вклад в теорию вероятностей, математический анализ и теорию чисел. Основное применение СММ получили в области распознавания речи, письма, движений и биоинформатике. Кроме того, СММ применяются в криптоанализе, машинном переводе.

Определяющее марковский процесс свойство принято называть *марковским*; впервые оно было сформулировано А.А. Марковым, который в работах 1907 г. положил начало изучению последовательностей зависимых испытаний и связанных с ними сумм случайных величин. Это направление исследований известно под названием теории цепей Маркова (*Markov Chain*).

Марковские модели активно используются при геномных исследованиях. Самые популярные статистические методы, используемые для поиска генов, – марковские модели, реализованы, например, в программах:

- Genemark – http://exon.biology.gatech.edu/ и http://opal.biology.gatech.edu/GeneMark/gmhmm2_prok.cgi ;
- GlimmerM – http://www.cbcb.umd.edu/software/glimmerm/ ;
- Critica – http://www.ttaxus.com/software.html ;
- AMIGene – http://www.genoscope.cns.fr/agc/tools/amigene/Form/form.php ;
- EasyGene – http://www.cbs.dtu.dk/services/EasyGene/ .

У *прокариотов* локус гена все ещё принято определять путём тривиального поиска открытой рамки считывания. Такой способ, конечно, не пригоден для высших *эукариотов*. Для различения кодирующих и некодирующих областей у высших эукариотов применяют *датчики содержания экзонов*, построенные на статистических моделях частот использования нуклеотидов и проводящие статистическую оценку некоторых зависимостей, наблюдаемых в структуре кодона.

7.4. Анализ экспрессии генов

Геном человека невероятно сложен и состоит приблизительно из трёх миллиардов пар нуклеотидов ДНК. При этом лишь только 3% ДНК является кодирующей последовательностью (то есть той частью генома, которая транскрибируется в РНК и затем транслируется в белок).

Остальная часть генома состоит из областей, необходимых для компактного хранения хромосом, их репликации во время деления клетки, управления транскрипцией и т. д. Основная часть работы по анализу последовательности генома приходится на исследование продуктов клеточных механизмов транскрипции и трансляции, то есть на анализ белковых последовательностей и структур.

В последнее время значительные усилия направлены на автоматизацию процессов исследования мРНК; частично это связано с тем, что смысловая машинная трансляция мРНК в последовательность белка может быть легко реализована алгоритмически, но главная причина состоит в том, что молекулы мРНК представляют ту часть генома, которая экспрессируется в клетках определённого типа на определённом этапе их развития.

Таким образом, можно выделить *три уровня геномной информации*:

1) геном хромосом (собственно *геном*) – генетическая информация, общая для всех клеток организма;
2) экспрессируемый геном (*транскриптом*) – та часть генома, которая экспрессируется в клетке на определённой стадии её развития;
3) *протеом* – совокупность белковых молекул, взаимодействие которых придаёт клетке её индивидуальные качества.

Каждый уровень требует различные аналитические методы и объяснительные алгоритмы. На разных стадиях развития и уровнях биологической активности клетки экспрессируют различный набор генов.

Такой характеристический набор экспрессируемых генов называют *профилем экспрессии* этой клетки.

Зарегистрировав профили экспрессии некоторой клетки, мы можем воссоздать картину уровней экспрессии генов в нормальном или ненормальном состоянии клетки, а также картину относительных уровней экспрессии всех генов, транскрибируемых в этой клетке. Кроме того, анализ зарегистрированных профилей экспрессии позволяет обнаруживать новые гены, тем самым дополняя другие методы, используемые в глобальных проектах секвенирования генома.

Процесс *регистрации профиля экспрессии* состоит в следующем. Сначала отбирают культуру клеток, затем из этих клеток извлекают РНК и стабилизируют её посредством обратной транскриптазы, с помощью которой с матрицы РНК синтезируют кДНК. Наконец, кДНК преобразуют в библиотеку (библиотеку кДНК), подходящую для использования в реакциях быстрого секвенирования.

Выборка клонов отбирается из библиотеки наугад – например, 10000 из библиотеки объёмом 2 миллиона клонов. Для того, чтобы инициировать 10000 реакций секвенирования и затем провести их на автосеквенаторах, выполняется сложная автоматизированная операция секвенирования. Итоговые данные загружаются в базу данных для дальнейшего анализа.

Идеальный результат – это набор из 10000 последовательностей; каждая из них имеет длину 200-400 нуклеотидов и представляет некоторую часть последовательности каждого из 10000 клонов.

В действительности некоторые реакции секвенирования вообще не получатся, некоторые производят недостаточно содержательные данные, а некоторые выдают данные неприемлемого качества. Последовательности, которые успешно миновали весь этот процесс, и называют ярлыками экспрессируемых последовательностей EST (*expressed sequence tags*).

Полученные EST-ярлыки помещают в GenBank, EMBL и DDBJ. К EST-ярлыкам открыт доступ через все эти базы данных. Те же самые EST-ярлыки находятся в базе данных dbEST, поддерживаемой NCBI.

7.5. Использование результатов секвенирования

Открытия, совершённые в различных программах исследования генома, найдут свое применение в следующих областях человеческой деятельности:

Молекулярная медицина:

- совершенствование диагностики заболеваний;
- обнаружение наследственного предрасположения к болезням;
- разработка лекарств на основе молекулярной информации и индивидуальных генетических профилей пациентов;
- развитие генотерапии.

Геномика микробов:

- быстрое обнаружение и уничтожение патогенов;
- разработка новых видов биологического топлива;
- защита граждан от последствий применения бактериологического и химического оружия;
- безопасная и эффективная очистка токсических отходов.

Оценка риска:

- оценка уровня риска для здоровья индивидуумов, которые подвергаются радиоактивному излучению или воздействию мутагенов;
- обнаружение загрязняющих веществ и наблюдение за состоянием окружающей среды.

Антропология и эволюция:

- изучение эволюции, обусловленной мутациями эмбрионов;
- изучение миграции различных групп населения;

- изучение мутации Y-хромосомы, чтобы проследить происхождение и миграцию мужчин.

Опознавание с помощью ДНК:

- установление личности преступников, ДНК которых может соответствовать вещественным уликам, оставленным на месте преступления;
- оправдание людей, ошибочно обвиненных в преступлениях;
- установление отцовства и других отношений родства;
- выявление биологических видов, находящихся под угрозой исчезновения и вне опасности вымирания;
- обнаружение бактерий и других организмов, которые могут загрязнять окружающую среду;
- определение соответствия доноров реципиентам при проведении операций по пересадке органов;
- определение родословной селекционного семенного материала или племенного скота.

Земледелие и животноводство:

- выращивание зерновых культур, устойчивых к болезням и засухе;
- повышение производительности;
- разведение домашнего скота;
- разработка и применение биопестицидов.

Контрольные вопросы

1. Что называется экспрессией генов?
2. Что такое микроматрица ДНК?
3. Опишите процесс гибридизации микроматриц ДНК.
4. Чем ДНК-чипы отличаются от микроматриц ДНК?
5. В каких случаях используют микроматрицы ДНК?

6. В чем состоит анализ экспрессии белка?
7. Что такое изоэлектрофокусировка?
8. Что такое изоэлектрическая точка белка?
9. Какие выделяют две главные стратегии открытия генов, экспрессирующих белки?
10. В чём заключается метод позиционного клонирования?
11. Какие элементы используются для обнаружения генов, анализируя пул экспрессируемых мРНК?
12. В чём заключается отличие участков-"сигналов" и участков-"содержания"?
13. Что такое скрытая марковская модель?
14. Что такое марковский процесс?
15. Какие выделяют три уровня геномной информации?
16. Что называется профилем экспрессии клетки?
17. Как результаты секвенирования могут быть использованы в молекулярной медицине?
18. Как результаты секвенирования могут быть использованы в геномике микроорганизмов?
19. Как результаты секвенирования могут быть использованы при оценке рисков?
20. Как результаты секвенирования могут быть использованы в антропологии и теории эволюции?
21. Как результаты секвенирования могут быть использованы при опознавании с помощью ДНК?
22. Как результаты секвенирования могут быть использованы в земледелии и животноводстве?

Список литературы

Основная

1. Fulekar M.H. Bioinformatics: Applications in Life and Environmental Sciences / M.H. Fulekar. – Berlin : Springer, 2009. – 247 p.
2. Griffiths J.F. An Introduction to Genetic Analysis / Griffiths J.F., Wessler S.R., Lewontin R.C., Gelbart W.M., Suzuki D.T., Miller J.H. – New York : W. H. Freeman Publishers, 2005. – 706 p.
3. Liebler D.C. Introduction to Proteomics. Tools for the New Biology / D.C. Liebler. – Totowa : Humana Press, 2002. – 198 p.
4. Lesk A.M. Introduction to Bioinformatics / Lesk A.M. – Oxford : Oxford University Press, 2002. – 255 p.
5. Marcus F.B. Bioinformatics and Systems Biology. Collaborative Research and Resources / F.B. Marcus. – Berlin : Springer, 2008. – 287 p.
6. Molecular Biomethods. Handbook / Ed. by J.M. Walker, R. Rapley. – Totowa : Humana Press, 2008. – 1124 p.
7. Stephenson F.H. Calculations for Molecular Biology and Biotechnology / F.H. Stephenson. – Amsterdam : Elsevier, 2003. –302 p.
8. Ramsden J. Bioinformatics. An Introduction / J. Ramsden. – Berlin : Springer, 2009. – 271 p.
9. Selzer P.M. Applied Bioinformatics. An Introduction / P.M. Selzer, R.J. Marhöfer, A. Rohwer. – Berlin : Springer, 2008. – 287 p.

10. Бородовский М. Задачи и решения по анализу биологических последовательностей / М. Бородовский, С. Екишева. – М.-Ижевск : РХД, 2008. – 440 с.
11. Винер Н. Кибернетика или управление и связь в животном и машине / Н. Винер. – М. : Советское радио, 1968. – 326 с.
12. Дурбин Р. Анализ биологических последовательностей / Р. Дурбин, Ш. Эдди, А. Крог, Г. Митчисон. – М.-Ижевск : РХД, 2006. – 480 с.
13. Игнасимуту С. Основы биоинформатики / С. Игнасимуту. – Ижевск : Институт компьютерных исследований, 2007. – 320 с.
14. Каменская М.А. Информационная биология / М.А. Каменская. – М. : Академия, 2006. – 368 с.
15. Корогодин В.И. Информация как основа жизни / В.И. Корогодин, В.Л. Кологодина. – Дубна : Изд. центр "Феникс", 2000. – 208 с.
16. Мёллер Г.Д. Избранные работы по генетике / Г.Д. Мёллер. – М.-Л. : Огиз-Сельхозгиз, 1937. – 350 с.
17. Николис Г. Познание сложного / Г. Николис, И. Пригожин. – М. : Мир, 1990. – 344 с.
18. Огурцов А.Н. Введение в биофизику. Физические основы биотехнологии / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2008. – 320 с.
19. Огурцов А.Н. Основы молекулярной биологии. Ч. 1. Молекулярная биология клетки / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2010. – 304 с.
20. Тимофеев-Ресовский Н.В. О природе генных мутаций и структуре гена / Н.В. Тимофеев-Ресовский, Л.Г. Циммер, М. Дельбрюк / Н.В. Тимофеев-Ресовский. Избранные труды. – М. : Медицина, 1996. – С. 105–153.
21. Хакен Г. Синергетика / Г. Хакен. – М. : Мир, 1985. – 423 с.
22. Чернавский Д.С. Синергетика и информация. Динамическая теория информации / Д.С. Чернавский. – М. : Либроком, 2009. – 304 с.
23. Шеннон Л. Работы по теории информации и кибернетике / К. Шеннон, Е. Бандвагон. – М. : Иностр. лит., 1963. – С. 667.
24. Шредингер Э. Что такое жизнь? Физический аспект живой клетки / Э. Шредингер. – М.-Ижевск : РХД, 2002. – 92 с.
25. Эйген М. Гиперцикл. Принципы самоорганизации макромолекул / М. Эйген, П. Шустер. – М. : Мир, 1982. – 270 с.

Дополнительная

26. Attwood T.K. Introduction to bioinformatics / T.A. Attwood, D.J. Parry-Smith. – Harlow : Addison Wesley Longman Ltd., 1999. – 218 p.
27. Brown S.M. Bioinformatics-A Guide to Biocomputing and the Internet / S.M. Brown. – Natick : Eaton Publishing, 2000. – 188 p.
28. Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins / Ed. by A.D. Baxevanis, B.F.F Ouellette. – New York : Wiley, 2001. – 470 p.
29. Pevsner J. Bioinformatics and Functional Genomics / J. Pevsner. – Hoboken : Wiley, 2003. – 753 p.
30. Bioinformatics. Managing Scientific Data Ed. by Z. Lacroix, T. Critchlow. – San Francisco : Morgan Kaufmann Publishers, 2003. – 441 p.
31. Bioinformatics for Geneticists / Ed. by M.R. Barnes, I.C. Gray. – Hoboken : Wiley, 2003. – 408 p.
32. Introduction to Bioinformatics. A Theoretical and Practical Approach / Ed. by S.A. Krawetz, D.D. Wombe, 2003. – 746 p.
33. Structural Bioinformatics / Ed. by P.E. Bourne, H. Weissig. – New York : Wiley, 2003. – 649 p.
34. Batiza A.F. Bioinformatics, Genomics, And Proteomics. Getting the Big Picture / A.F. Batiza. – New York : Infobase Publishing, 2006. – 196 p.
35. Isaev A. Introduction to mathematical methods in bioinformatics / A. Isaev. – Berlin : Springer, 2006. – 294 p.
36. Böckenhauer H.-J. Algorithmic Aspects of Bioinformatics / H.-J. Böckenhauer, D. Bongartz. – Berlin : Springer, 2007. – 396 p.
37. Bioinormatics. From Genomes to Therapies / Ed. by T. Lengauer. – New York, Wiley, 2007. – 1781p.
38. Plant Bioinformatics. Methods and Protocols / Ed. by D. Edwards. – Totowa : Humana Press, 2007. – 552 p.
39. Polanski A. Bioinformatics / A. Polanski, M. Kimmel. – Berlin : Springer, 2007. – 376 p.

40. Xia X. Bioinformatics and the Cell. Modern Computational Approaches in Genomics, Proteomics and Transcriptomics / X. Xia. – Berlin : Springer, 2007. – 349 p.
41. Sharma K.R. Bioinformatics.-Sequence Alignment and Markov Model / K.R. Sharma. – Ney York : McGraw-Hill, 2009. – 320 p.
42. Carugo O. Data Mining Techniques for the Life Sciences / O. Carugo, F. Eisenhaber. – New York : Humana Press, 2009. – 407 p.
43. Kangueane P. Bioinformation Discovery. Data to Knowledge in Biology / P. Kangueane. – Berlin : Springer, 2009. – 166 p.
44. Statistical Bioinformatics. A Guide for Life and Biomedical Science Researchers / Ed. by J.K. Lee. – New York : Wiley, 2010. – 350 p.
45. Хакен Г. Информация и самоорганизация / Г. Хакен. – М. : Мир, 1991. – 240 с.
46. Гасфилд Д. Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология / Д. Гасфилд. – СПб. : БХВ-Петербург, 2003. – 654 с.
47. Компьютеры и суперкомпьютеры в биологии / Под ред. В.Д. Лахно, М.Н. Устьин. – Ижевск : Институт компьютерных исследований, 2002. – 528 с.
48. Леск А. Введение в биоинформатику / А. Леск. – М. БИНОМ, 2009. – 318 с.
49. Хакен Г. Тайны природы. Синергетика: учение о взаимодействии / Г. Хакен. – Ижевск : Инст. компьютерных исследований, 2003. – 320 с.
50. Eigen M. The Hypercycle: A principle of natural self-organization / M. Eigen, P. Schuster. – Berlin : Springer 1979, – 92 p.
51. Огурцов А.Н. Механизмы мембранных процессов / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2006. – 139 с.
52. Огурцов А.Н. Молекулярная биология клетки. Основы клеточной организации / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2006. – 169 с.
53. Огурцов А.Н. Кинетика ферментативных реакций / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2007. – 146 с.
54. Огурцов А.Н. Молекулярная биология клетки. Молекулярные основы генных технологий / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2008. – 104 с.

55. Огурцов А.Н. Структурные принципы бионанотехнологии / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2008. – 140 с.
56. Огурцов А.Н. Введение в молекулярную биотехнологию / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2008. – 152 с.
57. Огурцов О.М. Молекулярна біофізика ферментів / А.Н. Огурцов. – Х. : НТУ "ХПІ", 2009. – 192 с.
58. Огурцов А.Н. Молекулярная биология клетки. Основные молекулярные генетические механизмы / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2007. – 120 с.
59. Огурцов А.Н. Молекулярная биоэнергетика клетки / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2009. – 112 с.
60. Огурцов А.Н. Молекулярная биотехнология клетки / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2009. – 120 с.
61. Огурцов А.Н. Функциональные принципы бионанотехнологии / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2009. – 146 с.
62. Огурцов А.Н. Структура, функции и аналитические методы исследования биомембран / А.Н. Огурцов, Н.Ю. Масалитина. – Х. : НТУ "ХПИ", 2010. – 240 с.
63. Огурцов А.Н. Электрогенез биомембран и механизмы мембранной сигнализации / А.Н. Огурцов, О.Н. Близнюк. – Х. : НТУ "ХПИ", 2010. – 224 с.
64. Огурцов А.Н. Введение в бионанотехнологию / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2010. – 136 с.
65. Огурцов А.Н. Ферментативный катализ / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2010. – 304 с.
66. Огурцов А.Н. Нанобиотехнология. Основы молекулярной биотехнологии / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2010. – 384 с.
67. Огурцов А.Н. Основы молекулярной биологии. Ч. 2. Молекулярные генетические механизмы / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2011. – 240 с.
68. Огурцов А.Н. Введение в молекулярную биофизику / А.Н. Огурцов. – Х. : НТУ "ХПИ", 2011. – 160 с.

Словарь терминов

A priori – от предшествующего – *априори* – знание, полученное до опыта и независимо от него (*знание априори*, *априорное знание*).

Ab initio – от начала, из первых принципов – моделирование, основанное на законах физики без использования эмпирических данных.

In silico – термин, обозначающий компьютерное моделирование (симуляцию) биологического эксперимента. Фраза была создана по аналогии с фразами *in vivo* (в живом организме) и *in vitro* (в пробирке), которые часто используются в биологии.

Omics – омики – англоязычный неологизм – общее обозначение биологических наук, оканчивающихся на "омика", таких, как геномика, протеомика, метаболомика, транскриптомика и пр.

Алгоритм – набор правил с заданной логической последовательностью шагов (стандартных операций), представляющих собой этапы решения задачи.

Алгоритм Нидлмена-Вунша – основанный на методе динамического программирования алгоритм поиска *глобального* выравнивания последовательностей.

Алгоритм Смита-Уотермана – основанный на методе динамического программирования алгоритм поиска *локального* выравнивания последовательностей. Главная особенность состоит в том, что все отрицательные счета, вычисленные по матрице динамического программирования, заменяются нулями, чтобы избежать продолжения выравниваний с плохим счетом и облегчить поиск (по этой

матрице) локальных выравниваний, начинающихся и заканчивающихся в любых заданных позициях.

Аллели – формы одного и того же гена, находящиеся в одном и том же месте (локусе) гомологичных (парных) хромосом и влияющие на проявление одного альтернативного признака.

Алфавит – общее число знаков (букв), принятых для описания последовательностей. Алфавит ДНК содержит 4 буквы, а алфавит белков – 20 букв.

Аналоги – в филогенетике к ним относят негомологичные белки, которые обладают подобным строением глобул или подобными функциональными участками и которые, как полагают, возникли в результате сходящейся эволюции.

Аннотация – совокупность комментариев, примечаний, ссылок и справок, которые организованы в произвольном формате или в виде управляемого словаря и в совокупности описывают всю экспериментальную информацию и теоретические выводы о гене или белке.

Апплет – маленькая прикладная программа, загружаемая с сервера через HTML-страницы.

База данных – собрание записей данных, представленное либо единственным файлом, либо совокупностью отдельных файлов.

Библиотека клонов – неупорядоченное собрание клонов (то есть клонированная ДНК некоторого организма), полученных из геномной ДНК или кДНК.

Библиотека кДНК – библиотека генов, собранная из фрагментов кДНК, синтезированных по матрице мРНК с помощью обратной транскриптазы.

Биоинформатика – наука о хранении, извлечении, организации, анализе, интерпретации и использовании биологической информации.

Бит – двоичный знак (от англ. *binary digit*) – один разряд двоичного кода (двоичная цифра). Может принимать только два взаимоисключающих значения: да/нет, 1/0, включено/выключено.

Бит информации – минимальное количество информации, необходимое для обозначения двух равновероятных возможностей (определение из теории информации). Число битов информации N, требуемое для передачи сообщения, содержащего M возможностей, равно $\log_2 M = N$ битов. Впервые слово "bit" было использовано Шенноном для логарифмической единицы информации в 1948 г. в статье "*A Mathematical Theory of Communication*".

БЛАСТ – программа для поиска подобия в базах данных последовательностей, BLAST.

Блок – не содержащий пропусков, мотив выравнивания, состоящий из сегментов последовательности, которые группируют, чтобы уменьшить кратный вклад от групп сильно подобных или идентичных последовательностей.

Бренность информации – свойство, определяемое фиксируемостью информации. Свойство бренности позволяет говорить о сроке жизни информации, который зависит от состояния её носителя.

Варианты сращения – белки различной длины, синтезированные путём трансляции молекул мРНК, собранных из произвольных разновеликих выборок экзонов из матричной ДНК.

Вверх – направление считывания последовательности молекулы ДНК, обратное направлению транскрипции.

Вдова – остаток аминокислоты, отделенный от соседних остатков ложными пропусками; как правило, результат чрезмерно усердной вставки пропусков программами автоматического выравнивания.

Вектор экспрессии – клонирующий вектор, который конструируют для осуществления экспрессии белка с кДНК.

Вестерн-блоттинг – метод опознавания антигенов в смеси белков с помощью специфических антител.

"Всемирная паутина" (www) – информационная система ресурсов, доступных через сеть Интернет, использующая HTTP в качестве основной среды передачи информации.

Вставка – или инсерция (от англ. *insertion*) – область выравнивания последовательностей, в которой одна последовательность очевидно имеет дополнительные мономеры по сравнению с другой последовательностью.

Всуд – вставка/удаление (*insdel*, от англ. *insertion*/*deletion*) в последовательности ДНК или белка.

Вторичная база данных – база данных, которая содержит информацию, полученную путём обработки первичных данных о последовательности и представленную, как правило, в форме регулярных выражений (комбинаций), индикаторов, блоков, профилей или скрытых марковских моделей. Эти абстракции представляют собой экстракт наиболее консервативных особенностей множественных выравниваний и, таким образом, вполне могут быть надёжными дискриминаторами для определения принадлежности недавно расшифрованных последовательностей к семействам.

Вторичная структура белка – упорядоченное строение полипептидных цепей, обусловленное внутрибелковыми водородными связями между группами C=O и N–H разных аминокислот.

Вторичная структура РНК – обусловлена комплементарным связыванием между участками самой одноцепочечной молекулы РНК или между двумя молекулами РНК. Вторичная структура РНК не так регулярна, как у ДНК.

Выборка – множество элементов, выбранных для исследования из множества данных с помощью определённой процедуры.

Выравнивание – взаимное расположение двух или более последовательностей нуклеотидов или аминокислотных остатков, при котором число совпадающих мономеров является возможно максимальным.

Выравнивание последовательностей – линейное сравнение последовательностей, в которые вводятся пропуски, чтобы передвинуть совпадающие позиции в сопоставленных последовательностях в правильный столбец так, чтобы по возможности большее количество одинаковых или близких мономеров оказались друг над другом. При этом разрешается удалить некоторые фрагменты сравниваемых последовательностей, удаленные фрагменты называются делециями. Выравнивания составляют основу методов анализа последовательностей и используются для точного определения места появления консервативных мотивов.

Гаплоид – организм (клетка, ядро) с одинарным (гаплоидным) набором хромосом.

Ген – фундаментальная физическая и функциональная единица наследственности. Ген представляет собой упорядоченную последовательность нуклеотидов, расположенную в определённой области определённой хромосомы и кодирующую специфический функциональный продукт (то есть молекулу белка или РНК).

Геном – наследственный аппарат организма; совокупность всех участков ДНК. Геном представляется в виде одной непрерывной последовательности нуклеотидов. Задача картирования (аннотирования) генома состоит в теоретическом определении границ кодирующих (белки или РНК) участков, координат регуляторных, связывающих и других функциональных областей.

Генерация информации – выбор варианта, сделанный случайно (без подсказки извне) из многих возможных и равноправных (т. е. из принадлежащих одному множеству) вариантов.

Генетическая карта – изображение относительных позиций известных генов, или маркеров.

Генетический алгоритм – особый алгоритм поиска, созданный по аналогии с механизмами эволюции. Алгоритм кодирует совокупность первичных решений, измеряет предопределённую пригодность

каждого решения и выбирает из первичной совокупности решения с самой высокой пригодностью для воспроизведения.

Генетический код – правила соотнесения четырёх оснований ДНК или РНК с 20 аминокислотами. С помощью трёх оснований (триплета, или кодона) можно закодировать 64 возможных последовательности. Каждый триплет уникально определяет одну аминокислоту, но одна аминокислота может быть закодирована несколькими кодонами (от одного до шести). Поэтому генетический код называют вырожденным.

Генный полиморфизм – обусловленные мутациями варианты кодонов в виде однонуклеотидных замен; вариабельность генов (популяционная и индивидуальная) в пределах одного вида.

Геном – весь генетический материал в хромосомах организма определённого биологического вида; размер генома в целом определяется общим количеством пар нуклеотидов.

Гибридизация – процесс соединения двух комплементарных нитей ДНК или нити ДНК с нитью РНК, приводящий к образованию молекулы в виде двойной спирали.

Гиперссылка – активная перекрестная ссылка HTTP, которая связывает веб-документы посредством сети Интернет.

Гипертекст – текст, содержащий вложенные ссылки (гиперссылки) на другие документы.

Гиперцикл – средство объединения самовоспроизводящихся единиц в новую устойчивую систему, способную к эволюции. Он построен из автокатализаторов, которые сочленены посредством циклического катализа, т. е. посредством ещё одного автокатализа, наложенного на систему.

Глобальное выравнивание – процедура приведения в соответствие как можно большего числа знаков, распространяющаяся на всю длину двух и более последовательностей.

Гомология – родство, обусловленное эволюционным процессом расхождения от общего предка. Гомология не является синонимом подобия. Гомология означает, конкретно, что последовательности и организмы, в которых они обнаружены, являются потомками общего предка.

Групповой анализ – метод объединения в группы наиболее подобных объектов, выбранных из более крупной группы соотнесенных объектов. Отношения определяются по некоторому критерию подобия или различия.

Гугол – (от англ. *googol*) – число равное 10^{+100}, такое, что никакая физическая величина не может иметь значение превышающее гугол.

Действенность информации – будучи включенной в свою информационную систему, информация может быть использована для построения того или иного *оператора*, который может совершать определённые целенаправленные действия. Оператор, таким образом, выступает в роли посредника, необходимого для *проявления* действенности информации.

Делеция – выпадение и потеря срединного участка хромосомы.

Динамическое программирование – метод сравнения и выравнивания строк или последовательностей по принципу, допускающему эффективное в вычислительном отношении введение пропусков.

Дискриминатор – математическая абстракция консервативного мотива или набора мотивов (например, образец регулярного выражения, профиль или индикатор), используемая для поиска идентичного или подобного мотива (мотивов) или в отдельно взятой последовательности запроса, или в целой базе данных.

Длина ветви – в анализе последовательностей длина ветви отражает число изменений последовательности, произошедших в ходе эволюции, направленной по данной ветви филогенетического дерева.

Домашняя страница – составленный на HTML документ, который служит первым пунктом контакта между программой-обозревателем и сервером.

Дублирование гена – генетическое изменение, при котором отдельный сегмент ДНК повторяется. Дублирования могут появиться где угодно; если дублированный сегмент примыкает к оригинальному, то такое дублирование называют тандемным.

E-значение – (*E-value* от англ. *expectation value*) – величина, предсказывающая вероятность того, что последовательности после выравнивания будут подобными при отсутствии эволюционного родства между ними. Если последовательности идентичны, то E = 0,0.

Задача свертывания белка – задача определения механизма (алгоритма) свертывания белка в его конечную пространственную структуру, определяемую лишь информацией, закодированной в его первичной последовательности.

Идиотип – карта упорядоченных по номеру и размеру хромосом клетки организма.

Изменчивость информации – это свойство информации, позволяющее изменять количество и (или) семантику информации, но сохраняющее её смысл.

Инвариантность информации – это возможность фиксации информации (записи) на любом языке, любым алфавитом.

Индикатор – группа непрерывных мотивов, используемая для построения характеристических сигнатур принадлежности к определённому семейству. Такие группы непрерывных мотивов опознают в выравниваниях последовательностей при диалоговом поиске в первичной (или смешанной) базе данных.

Инициатор – стартовый кодон, с которого начинается белок-кодирующий участок ДНК.

Инсерция – перемещение (вставка) участков внутри хромосом.

Информация – запомненный выбор одного варианта из нескольких возможных и равноправных.

Истинное несовпадение – несовпадение, правильно распознанное дискриминатором в качестве такового.

Истинное совпадение – совпадение, правильно распознанное дискриминатором в качестве такового.

Истинность информации – свойство, которое выявляется в ходе реализации полезности информации. Критерий истинности – практика. Из свойства полипотентности информации следует относительность её истинности, т.е. зависимость от ситуации и цели.

Итеративный – предполагающий многократное выполнение последовательности операций.

Кариотип – схематически представленная характерная для вида совокупность морфологических признаков хромосом (число, размер, форма, детали строения и т. д.).

Кладограмма – древовидная диаграмма, в которой каждый узел имеет две ветви; представляет эволюционную историю как процесс видообразования путём раздвоения эволюционных линий.

Клон – генетически однородное вегетативное потомство одной особи.

Клонирование – процесс производства идентичных копий некоторого фрагмента ДНК (который может кодировать целый ген), вырезанного из единственной матричной ДНК; также процесс создания идентичных копий клеток – потомков общего предка.

Клонирующий вектор – молекула ДНК, сконструированная из части генетического материала вируса, плазмиды или клетки высшего организма, в которую может быть встроен фрагмент ДНК без нарушения способности вектора к саморепликации.

Кодирующая последовательность – область ДНК или РНК, последовательность которой определяет последовательность аминокислот в белке.

Кодон – последовательность трёх смежных нуклеотидов, которая кодирует либо определённый мономер (азотистое основание или остаток аминокислоты), либо старт- или стоп-участок для механизма считывания.

Компетенция – состояние организма (или органа) в котором он подготовлен к следующему акту морфогенеза. В терминах теории динамических систем компетенция соответствует состоянию, близкому к бифуркации.

Комплементарное связывание – это связывание нуклеотидов, расположенных в разных цепях ДНК, которое обеспечивает связь двух полинуклеотидных цепей. Это связывание очень специфично: гуанин связывается только с цитозином, а аденин – с тимином. Двухцепочечная молекула ДНК содержит две последовательности, каждая из которых получается из противоположной перекодировкой всех A на T, всех T на A, всех G на C и всех C на G.

Консенсус (консенсусная последовательность) – обычно записывается под набором выровненных последовательностей и состоит из символов, характеризующих наиболее часто встречающиеся в колонке над этим символом варианты. Другими словами, это однозначный паттерн серии последовательностей.

Консервативная последовательность – последовательность оснований в молекуле ДНК (или последовательность аминокислот в белке), которая в ходе эволюции оставалась фактически неизменной.

Контиг – *contig* (от англ. *contiguous* – смежный, прилегающий) – непрерывный фрагмент молекулы ДНК, собранный из набора клонированных фрагментов, непрерывно перекрывающих в известном порядке часть генома или весь геном. Контиг формируется, например, при сборке бактериальной искусственной хромосомы.

Конформация – взаимное пространственное расположение атомов и связей в молекуле, обусловливающее её строение и, следовательно, функцию.

Корневое дерево – филогенетическое дерево, в котором наименее общий предок всех биологических видов выделен в виде отдельного порождающего узла.

k-Кортежи – короткие идентичные отрезки последовательностей, называемые также словами.

КОСА – (количественное отношение структура-активность). Математическая функция, описывающая взаимосвязь между структурными особенностями молекулы и её биологической функцией.

Кроссинговер – механизм обмена генами или комплексами генов гомологичных хромосом.

Лиганд – любая маленькая молекула, которая связывается с белком или рецептором.

Линейный штраф за пропуски – счет штрафа за пропуски, определяемый линейной функцией длины пропуска и состоящий из штрафа за введение пропуска и штрафа за продолжение пропуска, умноженного на длину пропуска.

Логарифмический счет шансов – логарифм счета шансов.

Локальное выравнивание – процедура выравнивания принадлежащих последовательностям областей с наивысшей плотностью совпадений, распространяющаяся на отдельные короткие отрезки обеих последовательностей.

Локус – участок генома, в котором расположен ген или генетический маркер.

Ложное несовпадение – истинное совпадение, которое неправильно распознаётся дискриминатором в качестве несовпадения.

Ложное совпадение – истинное несовпадение, неправильно распознаваемое дискриминатором в качестве совпадения.

Маркер – ген или участок ДНК с известной локализацией на хромосоме с определённым фенотипом, проявляющимся как свойства организма или самой молекулы ДНК.

Множественное выравнивание – это взаимное выравнивание многих последовательностей.

Максимальное правдоподобие – наиболее вероятный исход (дерево или выравнивание) при вероятностном моделировании эволюционных изменений в последовательностях ДНК.

Максимальная экономичность – минимальное число эволюционных шагов, необходимых для воспроизведения наблюдаемых изменений в наборе последовательностей; определяют путём сравнения числа шагов во всех возможных филогенетических деревьях.

Матрица BLOSUM – BLOcks Substitution Matrix – матрица, полученная с помощью локальных множественных выравниваний более отдаленно связанных последовательностей. Их применяют для оценки подобия последовательностей при построении выравниваний.

Матрица счетов PAM – Percent Accepted Mutation – матрица процентов точечных мутаций (ПТМ) описывает вероятность замен оснований или аминокислот в ходе эволюции. Матрицы PAM аминокислот получают из семейств близкородственных последовательностей и используют для оценки подобия последовательностей при построении выравниваний.

Метод дробовика – Shotgun sequencing – метод клонирования фрагментов ДНК, полученных путём произвольного дробления генома.

Метод объединения соседей – метод объединения в группы подобных пар из набора родственных объектов; позволяет построить дерево, ветви которого отражают степени различия между этими объектами (гены с подобными последовательностями).

Меченый участок последовательности STS – Sequence-Tag Site (МУП) – короткий (200-500 bp) отрезок последовательности ДНК, который присутствует в геноме человека в единичном экземпляре; его местоположение и последовательность оснований известны. МУПы опознаются в ходе полимеразной цепной реакции (ПЦР) и помогают направлять и ориентировать картографирование, а также со-

относить данные о последовательности, сообщаемые из многих лабораторий, то есть служат ориентирами на развивающейся физической карте генома человека. Ярлык экспрессируемой последовательности EST (ЯЭП) – это STS (МУП), полученный из кДНК.

Микроматрица – миниатюрный прибор, называемый также чипом (микрочипом, биочипом), который содержит сотни или тысячи различных молекул, закрепленных на подложке в узлах регулярной сетки.

Монте-Карло – метод статистических испытаний, в котором общее решение сложной задачи отыскивается по совокупности возможных частных решений (случайных проб).

Мотив – непрерывная цепь следующих друг за другом аминокислот в последовательности белка, общий характер которой повторяется (или сохраняется) в некоторой постоянной позиции всех последовательностей во множественном выравнивании.

Морфоген – фактор, способствующий образованию той или иной формы состояния компетенции. В качестве морфогенов могут выступать химические соединения (часто очень простые) и физические воздействия. Часто к одинаковому результату приводит несколько различных морфогенов.

Морфогенез – образование новых пространственных форм при онтогенезе.

Мультипликативность информации – возможность одновременного существования одной и той же информации в виде идентичных копий на одинаковых или разных носителях.

Мутация – скачкообразное изменение наследственного признака вследствие изменения генетического материала.

Нейронная сеть – применяется в алгоритмах с элементами искусственного интеллекта. Абстрактная структура, состоящая из множества простых единиц, которым присвоены численные веса и которые содержат символьные данные. Каждая единица работает только с

символьными данными, поступившими на её вход по связям с другими единицами.

Нормализованная библиотека – библиотека кДНК, которая организована таким образом, что все гены в ней представлены с одинаковой частотой.

Объектно-ориентированная база данных – база данных, в которой данные хранятся в виде абстрактных объектов, связанных абстрактными же отношениями. Единицей хранения информации является не запись, как в реляционных базах данных, а объект. Форматы представления данных могут быть самыми разными, включая, например, строки знаков, оцифрованные изображения, таблицы и т. д. Комплексный объект может включать в себя множество других объектов, причем объектно-ориентированная база данных позволяет осуществлять выборку таких объектов как цельных элементов. Благодаря гибкой системе представления данных и возможности объединять объекты в группы, объектно-ориентированные базы данных являются мощными информационными системами.

Однонуклеотидные замены – генные полиморфизмы, SNP.

Онтогенез – последовательность всех этапов развития особи от возникновения зародыша до смерти особи или до полного отмирания всего её вегетативного потомства. Для начальных стадий онтогенеза принят термин эмбриогенез.

Онтогенетическое состояние – определённый этап онтогенеза, характеризующийся специфическим физиолого-биохимическим состоянием, наличием ряда индикаторных морфологических и биологических признаков. Каждый этап онтогенеза характеризует биологический возраст особи.

Онтология – описание отношений между объектами (особенно в системах искусственного интеллекта).

Операционная система – программа или комплект программ для управления работой компьютера, контроля операций ввода/вывода, прерывания пользовательских запросов и т. д. (например: Windows, Unix, …).

Оперон – участок ДНК, состоящий из регуляторных элементов – промотора, оператора и структурных генов. Детерминирует синтез белков-ферментов, осуществляющих последовательные биохимические реакции в организме.

Описатель – слово (фраза), содержащее данные об отдельной последовательности или о наборе последовательностей; объём такой информации зависит от места описателя в записи.

Оптимальное выравнивание – выравнивание с наивысшим счетом, построенное алгоритмом, способным к нахождению кратных решений. Это возможно лучшее выравнивание, которое может быть найдено для любого параметра, заложенного пользователем в программу выравнивания последовательностей.

Ортологи – гомологичные белки, которые синтезируются в организмах различных биологических видов, но выполняют аналогичные функции.

Ортологичные гены – гомологичные гены, которые неодинаково эволюционировали у разных видов, имеющих общего предка.

Открытая рамка считывания (ORF) – ряд кодонов ДНК (в том числе старт-кодон и стоп-кодон), кодирующий предполагаемый или известный ген.

Паралоги – синтезируемые в одном организме гомологичные белки, которые выполняют различные, но связанные функции.

Паралогичные гены – гомологичные гены, возникшие в результате дупликации и эволюционировавшие параллельно в одном и том же организме.

Параметрическое выравнивание последовательностей – метод, позволяющий найти диапазон возможных выравниваний путём варьирования параметров системы очков за совпадения и несовпадения, а также штрафов за пропуски.

Паттерн – (от англ. *pattern* – образец, шаблон, модель) – это либо фрагмент последовательности, либо (реже) некий стандартный набор процедур, применяемый к разным объектам.

Первичная база данных – база данных, содержащая последовательности биомолекул (белков или нуклеиновых кислот) и сопутствующие аннотации (организм, биологический вид, функция, мутации, связь с определёнными заболеваниями, функциональные и структурные комбинации, библиографические ссылки и т. д.).

Первичная структура белка – последовательность расположения аминокислотных остатков в полипептидной цепи.

Плоский файл – файл данных, предназначенный для обмена информацией между базами данных и представленный в удобном для человека формате. Плоские файлы могут быть созданы на базе реляционных баз данных и приведены к формату, подходящему для их загрузки в другие базы данных.

Позиционная матрица счетов (ПМС) – (Position Weight Matrix (PWM), или Position-Specific Weight Matrix (PSWM), или Position-Specific Scoring Matrix (PSSM)) – отражает изменения, отмеченные в столбцах выравнивания множества родственных последовательностей. Каждый последующий столбец матрицы соответствует очередному столбцу в выравнивании, а каждая строка соответствует определённой последовательности знаков.

Поиск мотивов совпадения – поиск совпадения короткой последовательности в одном или более отрезках длинной последовательности.

Полезность информации – свойство информации, предполагающее, что она кому-нибудь нужна, может быть с пользой применена для некоторых целенаправленных действий. На основании свойства полипотентности можно утверждать, что полезной может оказаться любая информация.

Полиморфизм отдельного нуклеотида (SNP) – изменение отдельного нуклеотида в последовательности ДНК.

Полипотентность информации – это возможность использования оператора, закодированного данной информацией, для осуществления различных действий (т. е. для достижения разных целей).

Попарное выравнивание – выравнивание последовательностей по парам.

Последовательность запроса – последовательность ДНК, РНК или белка, используемая в качестве образца для поиска в базах данных последовательностей с целью отыскать близкие или отдаленные члены семейства с известной функцией.

Правило – короткое регулярное выражение (обычно длиной 4-6 остатков), используемое для опознавания кодируемых геномом (без установленной принадлежности к семействам) комбинаций в последовательностях белка. Правила тяготеют к кодированию определённых функциональных участков (например, участков присоединения сахаров, фосфорилирования, гидроксилирования, сульфатации и т. д.). Однако их небольшой размер означает, что регулярные комбинации не могут быть хорошими дискриминаторами, а могут лишь служить подсказкой при решении вопроса о возможности существования в последовательности того или иного функционального участка.

Праймер – короткий олигонуклеотид, к которому с помощью ДНК-полимеразы можно добавлять новые нуклеотиды.

Предсказание структуры – процесс алгоритмического восстановления вторичной, третичной и даже четвертичной структуры белка по последовательности аминокислотных остатков.

Продукт гена – белок, синтезируемый в ходе экспрессии гена. В некоторых случаях продуктом гена может быть молекула РНК, которая никогда не транслируется.

Прокариот – организм, для которого характерно отсутствие ограниченного мембраной, структурно обособленного ядра (и других внутриклеточных полостей). Примером таких доядерных организмов служат бактерии.

Промотор – регуляторный участок молекулы ДНК (80–90 пар нуклеотидов), с которым связывается РНК-полимераза, инициирующая транскрипцию.

Пропуск – область выравнивания последовательностей, в которой одна из последовательностей не содержит никакого мономера.

Протеом – вся совокупность белков, синтезируемых с данного генома; сюда же относятся варианты одного и того же базового белка, появляющиеся в результате посттрансляционных модификаций.

Протокол связи – согласованный набор правил для стандартизации связи между программами.

Протягивание – метод предсказания структуры белка, состоящий в выравнивании последовательности белка неизвестной структуры с моделью известной пространственной структуры; позволяет определить пространственную и химическую совместимость последовательности аминокислот с моделью известной структуры.

Профиль – позиционная таблица счетов, в которую сведена информация о полном выравнивании последовательностей. Профили показывают, какие остатки могут находиться в данных позициях; какие позиции консервативны, а какие вырождены; которые позиции, или области, допускают вставки. В дополнение к данным, полученным из выравнивания, система очков может включать в себя эволюционные веса и результаты анализа структур. Дифференциальные штрафы предназначены для компенсации вставок и удалений, встречающихся в элементах вторичной структуры.

Профиль экспрессии – характеристический набор генов, экспрессируемых в различных стадиях развития и функционирования клетки.

Процент подобия – счет выравнивания последовательностей аминокислот; счета замен различных аминокислот ранжированы с помощью матрицы замен.

Регулярная комбинация – обнаруживаемые в биомолекулах регулярные комбинации, как правило, образованы мономерами, составляющими последовательность гена или белка.

Регулярное выражение – отдельное согласованное выражение, полученное из консервативной области выравнивания последовательностей и используемое в качестве характеристической сигнатуры принадлежности к семейству. Синонимичные термины: правило, регулярная комбинация.

Регулятивная область или последовательность – область в последовательности ДНК (или целая последовательность), которая управляет экспрессией гена.

Реляционная база данных – база данных, построенная на реляционной модели данных (основанной на отношениях); данные организованы в двумерные таблицы. Таблицы описывают различные характеристики или свойства данных, но содержат избыточную информацию.

Репрессор – регуляторный белок, контролирующий синтез (трансдукцию) мРНК с определённого оперона.

Рестрикционная карта – вид физической карты, на которой указан порядок следования и расстояния между сайтами расщепления ДНК рестриктазами (обычно участок узнавания рестриктазы 4–6 bp). Маркерами этой карты являются рестрикционные фрагменты/сайты рестрикции.

Рестрикционные фрагменты – фрагменты молекулы ДНК, которые образуются при разрезании её ферментами рестрикционными эндонуклеазами (рестриктазами).

Рестрикционный портрет – характеристика молекулы ДНК по набору рестрикционных фрагментов, размер которых определяют на электрофореграмме, не составляя рестрикционой карты.

Сайт связывания рибосомы (ССР) – участок вблизи 5'-конца матричной РНК, который служит рибосоме для прикрепления к мРНК в нача-

ле процесса считывания белка с матрицы мРНК. Прикрепившись, рибосома начинает движение по мРНК к 3'-концу, попутно транслируя каждый триплет в аминокислоту. ССР в прокариотах называется ещё сайтом Шайна-Дальгарно (по имени ученых, открывших его существование). Связывание осуществляет определённый тип рибосомальных РНК (так называемый 16S-рРНК), у которой на 3'-конце есть комплементарный ССР участок.

Сборка – процесс выравнивания перекрывающихся фрагментов последовательности в одну НПО или в ряд НПО.

Сдвиг рамки считывания – изменение смысла считывания ДНК, возникающее в результате вставки или выпадения основания, при котором рамка считывания всех последующих кодонов смещается соответственно числу произошедших изменений (например, если к началу исходной последовательности, в которой читаются кодоны UCU-CAA-AGG-UUA добавить одно основание U, то новая последовательность будет читаться как UUC-UCA-AAG-GUU, и т.д.). Сдвиг рамки считывания может быть вызван появлением случайных мутаций или ошибок в чтении результатов секвенирования.

Секвенирование – определение порядка нуклеотидов (последовательности оснований) в молекуле ДНК или РНК, либо порядка аминокислот в белке.

Семантика – (от др.-греч. σημαντικός – обозначающий) – раздел языкознания, изучающий смысловое значение единиц языка.

Семантика (или содержательность) информации – проявляется в специфике кодируемого информацией оператора, причем каждая данная информация однозначно определяет оператор, для построения которого она использована.

Семейства генов – группы тесно связанных генов, которые кодируют подобные белковые продукты.

Сиквенс – последовательность нуклеотидов в фрагменте ДНК (от англ. *sequence* – последовательность).

Синтения – сохранение порядка следования генов в геномах (относительно) родственных организмов.

Система выборки последовательностей (SRS) – Sequence Retrieval System – средство выборки данных.

Скрытая марковская модель (HMM) – Hidden Markov model – вероятностная модель, состоящая из множества взаимосвязанных состояний. Подобно «Профилям», программа «HMM» кодирует полные выравнивания доменов. HMM представляют собой существенно линейные цепи состояний "совпадение", "удаление" или "вставка"; состояние совпадения обозначает консервативный столбец в выравнивании; состояние вставки, напротив, допускает вставки; состояния удаления позволяют пропускать позиции совпадения.

Смешанная база данных – база данных, которая объединяет в себе множество первичных источников и использует набор заданных критериев, определяющих приоритет включения различных источников и необходимый порог избыточности.

Смысловая трансляция – вычислительный процесс интерпретации смыслового содержания последовательности нуклеотидов мРНК и кодирования его с помощью генетического кода в последовательность аминокислот, которая далеко не во всех случаях будет описывать белок.

Согласованная последовательность – псевдопоследовательность, которая содержит сводную информацию о расположении остатков, содержащуюся во множественном выравнивании.

Сплайсинг – (от англ. *to splice* – сплетать, сращивать) – процесс удаления интронов из РНК-транскрипта и соединения экзонов с образованием мРНК.

Сравнительная геномика – наука о сравнении геномов различных организмов по числу генов, локусам и биологическим функциям генов; одна из целей состоит в определении групп генов, кодирующих уникальные для каждого организма биологические функции.

Сравнительное моделирование – процесс предсказания структуры белка на основании сравнения с последовательностью родственного белка с известной структурой.

Сходство – это наличие или измерение сходства и различия, независимо от источника сходства.

Счёт выравнивания – счет, алгоритмически вычисленный по числу совпадений, замен, вставок и удалений (пропусков) в выравнивании. Счета выравниваний выражают в единицах логарифмов шансов, часто в двоичных единицах (логарифм по основанию 2).

Счёт шансов – отношение правдоподобий двух событий, или исходов. Счёт шансов на соответствие знаков двух последовательностей (применяемый при оценке выравниваний последовательностей и получении матриц счетов) равен отношению частоты наступления событий выравнивания знаков в родственных последовательностях к частоте абсолютно случайного выравнивания тех же двух знаков при условии совместного появления этих знаков в последовательностях. Счета шансов для множества отдельных выровненных позиций вычисляют путём перемножения счетов шансов, найденных для каждой отдельной позиции. Счета шансов часто преобразуют к логарифмам и таким образом получают логарифмические счета шансов; тогда логарифмический счет шансов выравнивания целых последовательностей вычисляют путём суммирования отдельных логарифмических счетов шансов, что намного удобнее.

Тандемные повторы – множественные копии идентичных последовательностей нуклеотидов, следующих один за другим в конкретном участке хромосомы.

Теория информации – отрасль математики, предметом которой служит измерение количества информации в битах; бит определяет минимальное количество структурной сложности, необходимое для кодирования единичного объёма информации.

Терминатор – стоп-кодон, которым заканчивается белок-кодирующий участок ДНК.

Точечная матрица – анализ диаграммы точечной матрицы представляет собой графический метод сравнения двух последовательностей.

Транскрипт – однонитевая цепь мРНК, синтезированная по матрице-гену; набор полинуклеотидных цепей мРНК, представляющих собой комплементарную копию последовательности оснований цепи ДНК-матрицы.

Транскрипция – синтез комплементарной РНК по последовательности ДНК (гена); первый шаг экспрессии гена.

Транслокация – любое перемещение хромосомных сегментов в наборе хромосом. Может быть внутрихромосомной или межхромосомной.

Трансляция – процесс, в котором генетический код, заложенный в мРНК, направляет синтез белков из аминокислот.

Трансляция с шестью рамками – трансляция отрезка ДНК, включающая в себя три трансляции в прямом и три – в обратном направлении.

Транслируемость информации – это свойство, противостоящее бренности информации, это возможность передачи информации с одного носителя на другой, т. е. размножение информации.

Трансмембранный домен – пронизывающая мембрану область белковой последовательности.

Третичная база данных – база данных, содержащая обработанную информацию из вторичных баз данных (регулярных комбинаций). Ценность таких ресурсов состоит в том, что они обеспечивают альтернативную схему назначения счетов для фактически тех же самых первичных данных и дают возможность выявить отношения, которые могли быть пропущены при сравнительном анализе данных в исходной форме.

Третичная структура – общая конфигурация последовательности белка, всех α-спиралей и β-структур белка, распределение в пространстве всех атомов белковой молекулы.

Унифицированный указатель ресурса (URL) – адрес источника информации. Указатель URL состоит из четырёх частей – протокола, имени хоста, пути к директории и имени файла, например: http://www.ilt.kharkov.ua/bvi/ogurtsov/ogurtsov.htm .

Файл – обособленный набор байтов, которым можно манипулировать как цельным объектом.

Фантомные всуды – ложные вставки или удаления, которые появляются, когда из-за физических неоднородностей в геле для секвенирования программа считывания или регистрирует какое-либо основание преждевременно, или же вообще пропускает проходящее мимо основание.

Фармакоинформатика – отрасль информатики, изучающая вопросы управления биологической и химической информацией в фармацевтической промышленности.

Феном – набор проявившихся наследственных признаков организма.

Фермент – белок-катализатор, который увеличивает скорость протекания биохимической реакции, но при этом не изменяет её направление и характер.

Физическая карта – графическое представление порядка следования физических маркеров (фрагментов молекулы ДНК), расстояние между которыми определяется в парах нуклеотидов.

Фиксируемость информации – это свойство, благодаря которому любая информация, не будучи ни материей, ни энергией, может существовать не в свободном виде, а только в зафиксированном состоянии – в виде записи на каком-либо физическом носителе.

Филогенетический анализ – изучение эволюционных отношений между организмом определённого вида и его предшественниками (например, с помощью филогенетических деревьев).

Филогенетическое дерево – графическое представление предполагаемых эволюционных отношений между группами организмов; такие отношения могут быть установлены, например, путём множественного выравнивания последовательностей белков или нуклеиновых кислот.

Функциональная геномика – наука, занимающаяся оценкой функции генов, опознанных путём сравнения геномов. Функцию недавно опознанного гена проверяют путём введения в этот ген мутаций и последующего анализа изменений, произошедших в фенотипе полученного мутантного организма.

Характеристика – аннотация на определённый отрезок последовательности.

Хромосомы – носители генов; нуклеопротеидные нитевидные самовоспроизводящиеся структуры ядра клетки.

Центральная догма – фундаментальный принцип молекулярной биологии, провозглашенный Френсисом Криком в 1958 году. Центральная догма гласит, что передача информации между нуклеиновыми кислотами или от нуклеиновых кислот белкам возможна, а передача между белками или от белков нуклеиновым кислотам – невозможна.

Четвертичная структура белка – это агрегация двух или большего числа полипептидных цепей, имеющих третичную структуру, в олигомерную функционально значимую композицию.

Штраф за пропуски – штраф, который вычитают из счета подобия последовательностей с целью учета пропусков в выравнивании последовательностей.

Штрафы – очки, или веса, которые в программах построения выравниваний последовательностей служат в качестве коэффициентов для вычисления счетов; такие веса обычно заданы в виде переменных параметров и могут быть изменены пользователем.

Эвристический алгоритм – экономичная стратегия поиска решения задачи, для которой вычисление точного решения практически неосуществимо.

Эдмана метод расщепления – применяемый в секвенировании полипептидов метод, с помощью которого остатки аминокислот последовательно отщепляют от N-конца посредством реакции с фенилизотиоцианатом, приводящей к образованию фенилтиокарбамилпептида (ФТК-пептида). Это соединение расщепляют в безводной кислоте, высвобождая промежуточное вещество тиазолинон и остаток пептида.

Экзон – кодирующая последовательность гена, входящая в первичный транскрипт.

Эквивалент генома – такое количество рекомбинантных клонов, суммарная длина вставок в которых равна длине генома.

Экспрессия гена – процесс, в ходе которого закодированная в гене информация преобразуется в структурные и функциональные элементы клетки. К экспрессируемым относятся гены, которые транскрибируются в мРНК и затем транслируются в белок, а также гены, которые транскрибируются в РНК, но не транслируются в белок (например, гены транспортной и рибосомной РНК).

Эмбриогенез – начальная, зародышевая стадия онтогенеза.

Энциклопедия клонов – контиг, перекрывающий весь геном.

Язык разметки гипертекста (HTML) – HyperText Markup Language – язык, применяемый для создания веб-документов, которые могут интерпретироваться и отображаться программой-обозревателем.

Ярлык экспрессируемой последовательности (EST) – Expressed Sequence Tag – отрезок последовательности клона, произвольно выбранного из библиотеки кДНК, и используемый для опознавания генов, экспрессируемых в определённой ткани.

Список условных обозначений

ASCII	American Standard Code for Information Interchange – Американский стандартный код обмена информацией (АСКОИ) определяет 128 знаков, которым присвоены номера 0–127.
BAC	Bacterial Artificial Chromosome – Бактериальная искусственная хромосома.
BIOML	Biopolymer Markup Language – Язык разметки биополимеров (БИОМЛ).
BIOS	Basic Input-Output System – базовая система ввода-вывода (БИОС).
BLAST	Basic Local Alignment Search Tool.
BLOSUM	BLOcks Substitution Matrix – Матрица БЛОСУМ для расчёта весов для замен в аминокислотных последовательностях.
bp	base pair – пара азотистых оснований (по).
BSML	Bioinformatic Sequence Markup Language – Язык разметки последовательностей в биоинформатике (БСМЛ).
CERN	От фр. Conseil Européen pour la Recherche Nucléaire – Европейский совет по ядерным исследованиям (ЦЕРН) http://public.web.cern.ch/public/
COG	Cluster of Ortologous Gene) – кластер ортологичных генов, представляет собой группу генов-ортологов или ортологичные группы паралогов из геномов разных организмов.
CORBA	Common Object Request Broker Architecture – Общая архитектура брокеров объектных запросов (ОАБОЗ).

dbEST Database of Expressed Sequence Tags – База данных ярлыков экспрессируемых последовательностей http://www.ncbi.nlm.nih.gov/nucest

DBMS Database Management System – Система управления базами данных (СУБД).

DDBJ DNA DataBank of Japan – Японский банк ДНК (ЯБД) http://www.ddbj.nig.ac.jp/Welcome.html.en

DNS Domain Name System – Система имен доменов (СИД).

EMBL European Molecular Biology Laboratory – Европейская лаборатория молекулярной биологии: http://www.embl.de/

EST Expressed Sequence Tags – Маркерные экспрессирующихся последовательностей или ярлыки экспрессируемых последовательностей (ЯЭП).

ExPASy Expert Protein Analysis System – Экспертная система анализа белков – первый веб-сервер молекулярной биологии http://www.expasy.org

FASTA Fast Allignment – быстрое выравнивание (ФАСТА).

FASTP Алгоритм "ФАСТП".

FISH Fluorescence In Situ Hybridization – Флуоресцентная гибридизация *in situ* – метод, который применяют *in situ* для детектирования и определения положения специфической последовательности ДНК на хромосомах.

FTP File Transfer Protocol – Протокол передачи файлов (ППФ).

GCG Genetics Computer Group – Фирма "Genetics Computer Group".

GI Genlnfo Identifiers – регистрационные номера "ГенИнфо" биологических последовательностей, внесенных в Entrez NCBI.

HGP Human Genome Project – Проект "Геном человека".

HMM Hidden Markov model – Скрытая марковская модель (СММ).

HTML HyperText Markup Language – Язык разметки гипертекста.

HTTP Hypertext Transfer Protocol – Протокол передачи гипертекстовых файлов.

HUGE HUman Genome Equivalents – количество информации, содержащееся в геноме человека.

IDL Interface Definition Language – Язык описания интерфейсов, (ИДЛ).

IHGSC International Human Genome Sequencing Consortium – Международный консорциум секвенирования генома человека.

NBRF National Biomedical Research Foundation – Национальный фонд биомедицинских исследований США.

NCBI National Center for Biotechnology Information – Национальный центр биотехнологической информации.

NHGRI National Human Genome Research Institute – Национальный институт исследования генома человека.

NIH National Inatitute of Health – Национальный институт здоровья: http://www.nih.gov/

NLM National Library of Medicine – Национальная медицинская библиотека.

ORB Object Request Broker – Брокер объектных запросов (БОЗ).

ORF Open Reading Frame – Открытая рамка считывания (ОРС).

PAGE Polyacrylamide Gel Electrophoresis – Электрофорез в полиакриламидном геле (ЭПААГ) .

PAM Percent Accepted Mutation – проценты точечных мутаций (ПТМ).

PCR Polymerase chain reaction – Полимеразная цепная реакция (ПЦР).

PDB Protein data bank – Банк белковых данных: http://www.pdb.org/

PERL Practical Extraction and Reporting Language – Практический язык извлечения данных и формирования отчетов (ПЕРЛ).

PFGE Pulsed Field Gel Electrophoresis – Электрофорез в пульсирующем электрическом поле. Метод разделения крупных (до 10 Mbp) фрагментов ДНК.

PHI-BLAST Pattern Hit Initiated BLAST – Программа поиска белков, содержащих определённый пользователем паттерн.

PIR Protein Information Resource – Ресурс информации о белках (РИБ): http://pir.georgetown.edu/

PRF База белковых данных Protein Research Foundation (Osaka): http://www.genome.jp/dbget-bin/www_bfind?prf

PSI-BLAST Position specific iterative BLAST – Программа сравнения с целью поиска последовательностей, обладающих незначительным сходством.

RFLP Restriction Fragment Length Polymorphism – Полиморфизм длины рестрикта (ПДР) – полиморфизм длин рестрикционых фрагментов в одинаковых локусах гомологичных хромосом.

SAGE Serial Analysis of Gene Expression – Серийный анализ экспрессии генов (САЭГ), http://www.sagenet.org/

SDS-PAGE Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis – Двумерный электрофорез белков в полиакриламидном геле в присутствии додецилсульфата натрия.

SeqDB The Sequence Database – База данных нуклеотидных последовательностей: http://pgrc-35.ipk-gatersleben.de/portal/page/portal/PG_BICGH/P_BICGH/P_BICGH_RESOURCES/SeqDB

SGML Standard Generalized Markup Language – Стандартный обобщённый язык разметки.

SNP Single Nucleotide Polymorphism – Однонуклеотидный полиморфизм (ОНП) или Полиморфизм отдельного нуклеотида (ПОН) – отличия последовательности ДНК размером в один нуклеотид.

SRS Sequence Retrieval System – Система выборки последовательностей (СВП): http://srs.ebi.ac.uk/

STR Short Tandem Repeat – Короткое тандемное повторение (КТП) – короткие (обычно до 5 нуклеотидов) тандемные повторы, находящиеся в одинаковых локусах гомологичных хромосом, но содержащие различное число повторов данного типа.

STS Sequence Tagged Site – Ярлык, определённый последовательностью или Меченый участок последовательности (МУП).

Swiss-Prot База данных белковых последовательностей Шведского Института биоинформатики (Swiss Institute of Bioinformatics, http://www.isb-sib.ch/): http://www.expasy.org/sprot/

TCP Transmission Control Protocol – Протокол управления передачей.

TCP/IP Transmission Control Protocol / Internet Protocol – Протокол управления передачей (данных) / Интернет-протокол (ПУП/ИП).

URL Uniform Resource Locator – Унифицированный указатель (информационного) ресурса.

UTR UnTranslated Region – нетранслируемые области (НТО).

VNTR Variable Number Of Tandem Repeat – Переменное число тандемных повторений (ПЧТП).

WWW World Wide Web – Всемирная паутина.

XML Extensible Markup Language – Расширяемый язык разметки (РЯР).

YAC Yeast Artificial Chromosome – Дрожжевая искусственная хромосома.

ПЦР Полимеразная цепная реакция – Polymerase Chain Reaction (PCR).

Содержание

Навчальне видання

ОГУРЦОВ Олександр Миколайович

ВСТУП ДО БІОІНФОРМАТИКИ

Навчальний посібник
по курсу «Біоіформатика та інформаційна біотехнологія»

для студентів напряму підготовки 051401 «Біотехнологія»,
в тому числі для іноземних студентів

Російською мовою

Відповідальний за випуск *М.Ф. Клещев*
Роботу до видання рекомендувала *М.Г. Зінченко*
В авторській редакції

План 2011 р., поз. 31 / 196-10.

Підп. до друку 20.12.2010 р. Формат 60 × 84 1/16. Папір офісний.
Riso-друк. Гарнітура Таймс. Ум. друк. арк. 12,0. Наклад 300 прим.
Зам. № 19. Ціна договірна

Видавничий центр НТУ «ХПІ».
Свідоцтво про державну реєстрацію ДК № 3657 від 24.12.2009 р.
61002, Харків, вул. Фрунзе, 21

Друкарня НТУ «ХПІ». 61002, Харків, вул. Фрунзе, 21

Учебное пособие содержит материалы по основным вопросам первого раздела курса «Биоинформатика и информационная биотехнология» в соответствии с программой подготовки студентов направления «Биотехнология».

Предназначено для студентов специальностей биотехнологического профиля всех форм обучения.

ISBN 978-966-593-885-9