Министерство образования Республики Беларусь
Учреждение образования «Витебский государственный
университет имени П.М. Машерова»
Кафедра экологии и географии

В.В. Яновская

БИОИНФОРМАТИКА

Курс лекций

Витебск
ВГУ имени П.М. Машерова
2022

УДК 57:004(075.8)
ББК 28с51я73
Я64

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 3 от 03.03.2022.

Автор: доцент кафедры экологии и географии ВГУ имени П.М. Машерова, кандидат биологических наук **В.В. Яновская**

Р е ц е н з е н т :
доцент кафедры гистологии, цитологии и эмбриологии УО «ВГМУ»,
кандидат биологических наук *С.М. Седловская*

Яновская, В.В.

Я64 Биоинформатика : курс лекций / В.В. Яновская. – Витебск : ВГУ имени П.М. Машерова, 2022. – 83 с.

ISBN 978-985-517-950-5.

Курс лекций разработан для магистрантов специальности 1-31 80 04 Биология в соответствии с учебной программой по дисциплине «Биоинформатика» (рег. № УД-24-064), на основе государственного Образовательного стандарта высшего образования 1-31 80 04-2019 Биология и учебного плана учреждения образования «Витебский государственный университет имени П.М. Машерова».

В учебное издание включены материалы курса лекций, контрольные вопросы, список литературы. Рекомендовано для использования преподавателями, студентами.

УДК 57:004(075.8)
ББК 28с51я73

ISBN 978-985-517-950-5

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Биоинформатика – новая наука, возникшая в начале 80-х годов на стыке молекулярной биологии и генетики, математики (статистики и теории вероятности) и информатики. Толчком к этому послужило появление в конце 70-х годов быстрых методов секвенирования последовательностей ДНК. Нарастание объема данных происходило очень быстро и стало ясно, что каждая полученная последовательность не только представляет интерес сама по себе, но и приобретает дополнительный смысл при сравнении с другими. Под биоинформатикой понимают науку, занимающуюся анализом экспериментальных данных молекулярной биологии: секвенированных последовательностей биополимеров, экспериментально определенных пространственных структур биологических макромолекул, данных об экспрессии генов и т.д. Методами биоинформатики являются методы организации информации, компьютерные методы, методы вычислительной математики и статистики.

Курс «Биоинформатика» состоит из четырех разделов: введение в биоинформатику, молекулярная эволюция, методы анализа последовательностей нуклеиновых кислот, методы анализа последовательностей белков. Научную основу курса «Биоинформатика» составляют такие дисциплины как генетика, биофизика, биохимия и молекулярная биология.

Цель изучаемой дисциплины – получение основополагающих сведений о содержании и возможностях информационной биологии (биоинформатики), возможностях приложения методов информационной биологии к решению фундаментальных и прикладных проблем молекулярной биологии.

Задачи: ознакомление с существующими методическими приемами и подходами, используемыми при работе с базами данных биологической направленности, освоение умения прогнозирования основных физико-химических и биологических свойств анализируемых нуклеотидных последовательностей и детерминируемых ими продуктов, а также представление их потенциальных функций. Так как основные данные, которые используются в биоинформатике – молекулярно-биологические, то курс включает в себя краткое изложение необходимых для понимания фактов из молекулярной биологии.

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.

Профессиональные компетенции, формируемые у студентов в результате изучения дисциплины – владение методическими приемами биоинформатики, алгоритмами обработки разных типов молекулярно-биологических данных с навыками программирования, математического и статистического анализа данных.

В результате изучения дисциплины магистр должен: **знать** основные понятия, цели и задачи биоинформационного анализа (состав и организация баз данных по биологическим последовательностям и молекулярным структурам, принципы навигации по биоинформационным ресурсам); основополагающие концепции биоинформатики; способы получения, организации и анализа данных; **уметь** использовать основные подходы и методы биоинформатики для решения конкретных научно-исследовательских задач; **владеть** ключевыми методами, такими как выравнивание последовательностей, филогенетический анализ, дизайн праймеров, аннотация геномной последовательности и т.д.

Раздел I
ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

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 растет и объем 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 посвященного 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.

1.2 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:

– 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;

– развивать программные средства и информационные ресурсы, которые помогают в управлении данными и в их анализе;

– 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 *еще 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.

1.3 Способы представления информации о последовательностях – формы записи Fasta, Genbank, PDB и способы визуализации

1.3.1 Понятие информации

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», подчеркивая невещественность происхождения информации.

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

Наиболее адекватным для биологических применений является следующее определение понятия «информация»:

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

Запомненный выбор еще называют *макроинформацией*. Если информация создается, но не запоминается, то ее называют *микроинформацией*.

Цитата 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.

1.3.2 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 1.1).

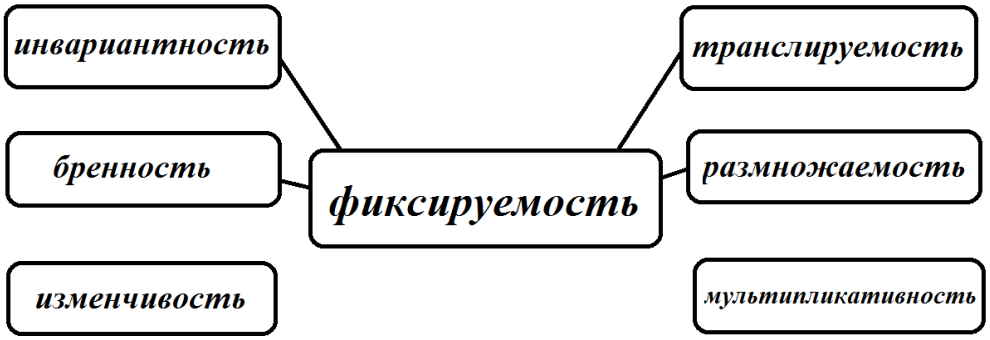


Рисунок 1.1 – Классификация свойств информации относительно ее фиксируемости

Для другой группы определяющим свойством является ее *действенность* (рисунок 1.2). Остальные свойства, входящие в эти группы, можно расценивать как раскрытие, проявление ключевых особенностей в формах, доступных для регистрации.

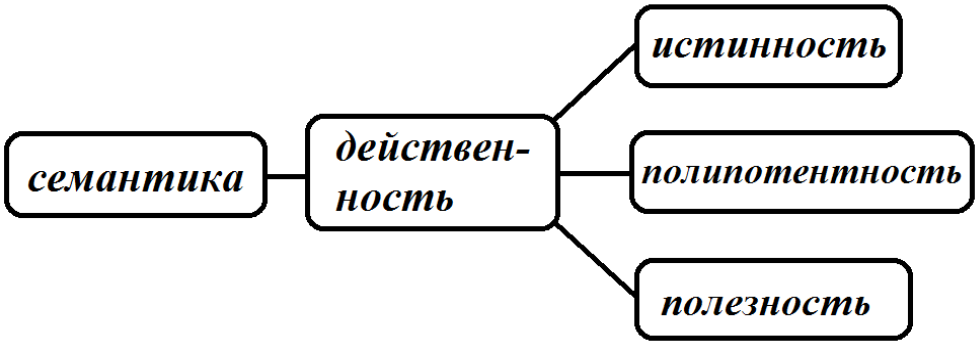


Рисунок 1.2 – Классификация свойств информации относительно ее действенности

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 ее семантике (или содержательности). Например, одно и то же предложение может быть записано и на бумаге, и на электронном носителе.

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

Инвариантность информации по отношению к физической природе ее 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 ее носителя. Свойство бренности позволяет говорить о *сроке жизни* информации, который зависит от состояния ее носителя. Рано или поздно носитель деформируется, и информация исчезает.

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

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

Действенность информации. Вторая группа свойств информации объединяется ключевым свойством – действенностью (рисунок 2). Это свойство выявляется следующим образом: будучи включенной в свою информационную систему, информация может быть использована для построения того или иного *оператора,* который может совершать определенные целенаправленные действия. Оператор, таким образом, выступает в роли посредника, необходимого для *проявления* действенности информации.

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 ее *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 ее 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 еще не произошло.

Истинность информации – свойство, которое выявляется в ходе реализации полезности. Критерий истинности – практика. Из свойства полипотентности информации следует относительность ее 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 ее истинности и полезности, т.е. «гармония объективного и субъективного аспектов информации, отражаемых этими терминами».

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.

С позиций динамической теории информации (для динамических систем) рецепция информации означает *перевод* системы в одно определенное 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 *остается* 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 отдается 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.

Согласно определению информации, информация есть запомненный выбор, т.е. макроинформация. На физическом языке «запомнить», т.е. зафиксировать информацию, означает привести систему в определенное 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.

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 четырех типов нуклеотидов, которые могут быть соединены в любой последовательности. Эти молекулы обладают свойством *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 ее 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. Меллер, 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 ее носителей от информации как таковой. Поэтому-то мы и говорим, что генетическая информация записана в ДНК определенной 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 Меллер 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 ее 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. Меллер, 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.

1.3.3 Базы и банки данных

База данных – это файл специального формата, содержащий информацию, структурированную определенным образом. Базы и банки биологических данных можно отнести к нескольким типам:

– архивные;
– курируемые;
– автоматические;
– производные;
– интегрированные.

Архивные базы данных. Например, GenBank, EMBL, PDB, где любой исследователь может поместить туда свою информацию.

GenBank – база данных генетических последовательностей, основанная в 1982 г. – это аннотированная коллекция всех общедоступных последовательностей ДНК, РНК, белков, снабженных литературными ссылками. Эта база является частью объединения International Nucleotide Sequence Database Collaboration, которое объединяет 3 крупных банка нуклеотидных последовательностей:

1. DDBJ (DNA Data Bank of Japan),
2. EMBL (European Molecular Biology Laboratory)
3. GenBank (National Center for Biotechnology Information).

Эти организации ежедневно обмениваются новой информацией. Большинство журналов требуют посылки новых секвенированных последовательностей в любую из этих 3 баз данных до опубликования статьей. В статьях, посвященных очередной порции последовательностей, должен упоминаться лишь номер последовательности в базе данных GenBank.

Адрес DDBJ: http://www.ddbj.nig.ac.jp/

Адрес GenBank: http://www.ncbi.nlm.nih.gov/Genbank/

EMBL (European Molecular Biology Laboratory) – эта база данных содержит информацию о каждом фрагменте последовательностей, включая литературные ссылки, перекрестные ссылки на документы других баз данных и др.

Адрес EMBL: http://www.ebi.ac.uk/embl/

PDB (Brookhaven Protein DataBank) – содержит данные о коллекции экспериментально определенных трехмерных структур биологических макромолекул (белков и нуклеиновых кислот). С 2002 года в основном депозитарии PDB хранятся структуры, экспериментально определенные с помощью рентгеноструктурного, ядерно–магнитного резонансного и др. методов. Теоретические структуры выделены в отдельную подбазу PDB.

Адрес: http://www.rcsb.org/pdb/

Курируемые базы данных – за содержание записей в таких базах данных отвечают кураторы. Информацию для курируемых баз данных отбирают эксперты из архивных баз. К курируемым базам относятся, например, SwissProt. Эта база данных белковых последовательностей

существует с 1986 года и поддерживается двумя институтами: Swiss Institute of Bioinformatics (SIB) и European Bioinformatics Institute (EBI).

Адрес: http://www.ebi.ac.uk/swissprot/

Автоматические базы данных. В таких базах данных записи генерируются (моделируются) компьютерными программами. Например, **TrEMBL (Translated EMBL)** – автоматическая база предсказаний последовательностей белков. Это формальная трансляция всех кодирующих нуклеотидных последовательностей из банка EMBL. В 2002 году в результате объединения SwissProt, TrEMBL и PIR был создан банк данных **UniProt (Universal Protein Resourse)**. Это основное хранилище белковых последовательностей и их функций. UniProt состоит из трех частей:

– UniProt Knowlegebase – является центральной базой данных и обеспечивает доступ к обширной курируемой информации по белкам, включая их функцию, классификацию и перекрестные информационные ссылки;

– UniProt Archive – UniParc. – отражает хронологию данных определения о всех белковых последовательностях;

– UniProt Reference – UniRef. – содержит базы данных, которые объединяют последовательности в кластеры для ускорения поиска.

Адрес UniProt: http://www.ebi.uniprot.org/index.shtml

Производные базы данных. Получаются в результате компьютерной обработки данных из архивных и курируемых баз данных. Это, например, SCOP, PFAM, GO и др. SCOP (Structural Classification Of Proteins) – база данных по структурной классификации белков.

Адрес: http://scop.protres.ru/

PFAM (Protein families database of alignments and HMMs) – это большая коллекция семейств белков и доменов, построенных на основании экспертной оценки множественных выравниваний. В банке существуют две основные части: PFAMA, содержащая подробно аннотированные белковые семейства, и PFAMB, содержащая различные множественные выравнивания.

Адрес: http://www.sanger.ac.uk/Pfam/

GO (Gene Ontology consortium database). Целью создателей базы было установление контроля за единообразием в описаниях функций, биологических процессов и клеточных компонентов, относящихся к продуктам генов. Унификация описаний в различных базах данных облегчает поиск в них нужного гена. GO – независимая база данных: другие базы данных сотрудничают с ней, помещая ссылки на унифицированные термины GO, либо поддерживают поиск с использованием терминов базы GO, а также стимулируют ее дополнение и уточнение.

Адрес: http://www.geneontology.org/

Интегрированные базы данных. Объединяют информацию из разных баз. Например, введя имя гена, можно найти всю, связанную с ним информацию. К таким базам относится **ENTREZ (Molecular Biology**

DataBase and Retrieval System). Эта интегрированная база данных содержит нуклеотидные и аминокислотные последовательности, которые собираются из крупнейших специализированных хранилищ – баз данных. Основой является **GenBank**, кроме того, информация пополняется из **dbEST, dbSTS, SwissProt, PIR, PDB, PRF, GSDB**. Данные из перечисленных ресурсов поступают в интегрированную базу данных после:

1) присвоения уникального идентификатора последовательности,

2) перевода документов в единый стандарт хранения,

3) проверки данных,

4) проверки всех ссылок по базе данных MedLine,

5) проверки названий организмов по таксономической классификации GenBank Taxonomy.

Адрес ENTREZ: http://www.ncbi.nlm.nih.gov/Database/index.html

Описания многих баз данных по БИ можно найти на русскоязычном сайте, который находится по адресу: http://www.jcbi.ru/index.html

1.3.4 Аналитические программы в биоинформатике

Приведем примеры основных программ сравнения аминокислотных и нуклеотидных последовательностей (рисунок 1.3).

1. ACT – (Artemis Comparison Tool) – геномный анализ;

2. Bio Edit – редактор множественного выравнивания аминокислотных и нуклеотидных последовательностей;

3. Bio Numerics – коммерческий универсальный пакет программ по биоинформатике;

4. BLAST – поиск родственных последовательностей в базе данных аминокислотных и нуклеотидных последовательностей;

5. ClustaIW – множественное выравнивание аминокислотных и нуклеотидных последовательностей;

6. FASTA – набор алгоритмов определения схожести аминокислотных и нуклеотидных последовательностей;

7. Mesquite – программа для сравнительной биологии на языке Java;

8. Muscle – множественное сравнение аминокислотных и нуклеотидных последовательностей. Более быстрая и точная программа в сравнении ClustaIW;

9. Pop Gene – анализ генетического разнообразия популяций;

10. Populations – популяционно-генетический анализ. Примером интегрированного инструмента биолога является также Unipro UGENE. Это свободно распространяемое программное обеспечение для работы молекулярного биолога. Пользовательский интерфейс этого продукта обеспечивает:

– с последовательностями;

– визуализацию хроматограмм;

– использование редактора множественного выравнивания последовательностей;

– просмотр трехмерных моделей PDB и MMDB с поддержкой стереорежима;

– просмотр филогенетических деревьев;

– применение конструктора вычислительных схем, автоматизирующего процесс анализа;

– поддержку сохранения изображений в векторные форматы для удобства публикаций.

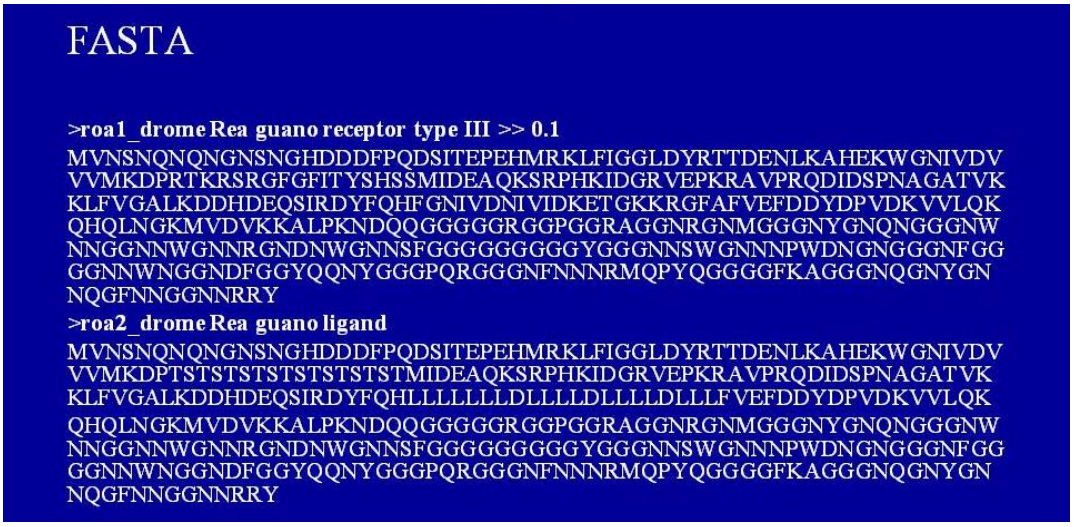


Рисунок 1.3 – Форматы файлов, используемых в биоинформатике

1.3.5 Первичные, вторичные и курируемые базы данных

Одной из базовых задач БИ является хранение и организация доступа к накопленным массивам биологической информации. Реализуют данную задачу с применением технологий баз данных (БД). Базы данных представляют собой информационные модели, содержащие данные об объектах и их свойствах. Они хранят информацию о группах объектов с одинаковыми наборами свойств. Простыми бытовыми примерами БД можно считать любые справочники, энциклопедии, записные книжки и каталоги.

Информация в базах данных хранится в упорядоченном виде, что позволяет обеспечить удобный доступ к нужным фрагментам хранимой информации.

Технически БД можно представить как набор таблиц, каждая их которых предназначена для хранения информации об объектах одного типа. Каждая строка таблицы содержит данные одного объекта и называется записью. При этом столбцы, формирующие строку, называются полями, и каждое поле описывают какую-либо характеристику объекта. Поскольку для каждой записи в БД должна существовать возможность уникальной идентификации, часто выделяют отдельный тип записи для хранения подобной информации, а данное поле называют ключевым. Также следует отметить, что возможно существование полей, содержащих в качестве значения ссылку на объект то-

го же или другого типа, позволяя, таким образом, хранить информацию об иерархических и сетевых связях объектов содержащихся в БД.

По характеру хранимых данных биологические БД можно разделить:

• на *первичные БД*, хранящие результаты молекулярно-биологических исследований. Как правило, это последовательности и структуры биологических полимеров (Genbank, EMBL, DDBJ, SWISS-PROT, TREMBL, PIR, PDB);

• в*торичные*, данные в которых являются результатом обработки первичной биологической информации. Типичными примерами являются БД, хранящие информацию о паттернах, обнаруживаемых в последовательностях, разного рода классификации последовательностей и структур (PROSITE, Pfam, BLOCKS, PRINTS, DSSP, SCOP);

• **составные (композитные) БД**, агрегируют информацию из первых двух видов, предоставляя расширенные по сравнению с отдельными БД возможности по поиску и навигации в данных (NRDB, OWL, GO).

По механизму наполнения базы данных можно разделить:

• на архивные базы данных, фактически являются хранилищем файлов определенного формата, предоставляемых учеными. Как правило, это первичные базы данных наподобие PDB;

• автоматические базы данных, представляющие результат работы какого-либо метода. Часто по предыдущей классификации их можно отнести ко вторичным (DSSP);

• курируемые базы данных, наполнение которых контролируется группой/лабораторией/исследовательским центром, их поддерживающим. Типичный пример − SWISS-PROT.

Поскольку БИ ориентирована на автоматическую обработку данных, основу большинства первичных и вторичных биологических баз данных составляют файлы определенного формата. В каждом подобном файле хранится информация об одном основном объекте данной БД, например данные о пространственной структуре одного комплекса в случае БД PDB. Обычно пользователь редко работает с самим файлом, поскольку веб-интерфейс сайта БД предоставляет более удобное для человека представление информации об объектах в виде различного рода сводных таблиц, последовательностей символов, рисунков и ссылок на другие сайты, содержащие дополнительную связанную информацию. Однако всегда следует помнить, что, как правило, данные файлы доступны для скачивания (при необходимости).

1.3.6 Автоматическое аннотирование последовательности

Автоматическое аннотирование последовательности – это процесс, в котором идентифицируются гены, их регуляторные области и функции генов, также определяют гены, которые не кодируют белки (в частности, гены рРНК, тРНК и малой ядерной РНК); обнаруживают и характеризуют

мобильные генетические элементы и семейства повторов, которые могут также присутствовать в геноме.

После определения нуклеотидной последовательности встает следующая задача по ее аннотации, которая заключается в идентификации всех генов и кодируемых белков, мобильных элементов и семейств повторов, которые могут присутствовать в геноме.

Гены, кодирующие белки, обнаруживаются при анализе нуклеотидной последовательности самим исследователем или при помощи компьютерных программ. Гены, кодирующие белки, содержат так называемую **открытую рамку считывания,** которая начинается с **инициирующего кодона** АТГ и заканчивается одним из трех **терминирующих кодонов –** ТАА, ТАГ или ТГА. Сканирование последовательности ДНК для обнаружения открытой рамки считывания, ограниченной АТГ с одной стороны и стоп-кодоном с другой, является одной из *стратегий поиска генов*. Однако этот метод высоко эффективен только для аннотации бактериальных геномов. В случае же геномов эукариот продуктивность метода резко снижается, поскольку большинство эукариотических генов состоят из **экзонов** (кодирующих участков гена) и **интронов** (некодирующих участков гена), и программа часто интерпретирует экзоны, как отдельные гены, т.к. стоп-кодоны часто встречаются в интронах.

Следует отметить, что последние версии программ настроены на поиск специфических черт открытых рамок: интрон-экзонных сочленений, 3'полиА-сигналов и **преимущественных кодонов**. Например, аланин может кодироваться четырьмя кодонами, но в геноме человека кодон ГЦЦ встречается в 41% аланиновых кодонов, а ГЦГ только в 11%. Наиболее часто встречающиеся кодоны присутствуют в экзонах, но не встречаются в интронах и пространствах между генами.

После обнаружения предполагаемых открытых рамок считывания для определения гена проводят поиск гомологичных последовательностей среди расшифрованных генов других организмов в базах данных (например, в Genbank).

1.3.7 Идентификаторы записей в базах данных

Для каждой записи в БД должна существовать возможность уникальной идентификации. В качестве уникального имени обычно используется идентификатор или инвентарный помер. Подобная двойственность является следствием того, что на ранних этапах становления биоинформатики, когда число последовательностей было невелико, имена последовательностям старались давать в удобочитаемой форме, закладывая в аббревиатуру указание на биологическую функцию последовательности. Так, в идентификаторах БД EMBL и Genbank первые две (три) буквы указывают на биологический вид организма, а оставшиеся – на функцию. Однако с увеличением объема БД и возрастанием скорости добавления последовательностей

подобная схема именования перестала удовлетворять потребности научного сообщества из-за отсутствия возможности автоматической генерации имен. На смену (в дополнение) к идентификаторам пришли инвентарные номера − уникальные символьные (буквы и цифры) последовательности. Следует иметь в виду, что три наиболее известные базы данных EMBL, GenBank и SwissProt используют общую схему нумерации последовательностей, т.е. единый инвентарный номер однозначно идентифицирует последовательность в этих трех базах данных.

Основные базы данных последовательностей биологических полимеров. База данных Genbank содержит аннотации последовательностей ДНК различных организмов. Каждая запись включает ряд полей, список которых несколько отличается для прокариотических и эукариотических последовательностей. Так, для записей, характеризующих прокариотические гены, свойственны следующие описания:

• LOCUS − название локуса (произвольное имя), длина нуклеотидной последовательности, тип молекулы (ДНК) и ее топология (линейная, кольцевая);

• DEFINITION − содержит короткое описание гена;

• ACCESSION − номера (идентификаторы) данного объекта в других БД;

• VERSION − перечислены синонимы и предыдущие идентификаторы;

• KEYWORDS − список терминов, характеризующих запись;

• SOURCE − общее название организма, являющегося источником данной последовательности;

• ORGANISM − полная таксономическая идентификация организма-источника;

• REFERENCE − ссылки на статьи, связанные с выделением и определением функций последовательности;

• COMMENT − комментарии, не подходящие по формату другим полям.

Секция, описывающая открытую рамку считывания гена:

• координаты стартового и стоп-кодона;

• тип таблицы кодонов, используемой для трансляции;

• translation – декодированная аминокислотная последовательность Вторая важная БД, о которой стоит упомянуть, – база белковых последовательностей SWISS-Prot (wvvw.expast.org/sprot). Данная БД, в отличие БД Genbank, ориентирована на белковые последовательности, в том числе последовательности, «транскрибированные in silico». Также база SWISS-Prot содержит подробные аннотации известных последовательностей и тесно интегрирована с другими БД. Например, если для последовательности из SWISS-Prot доступна структурная информация, то данные об этой последовательности будут содержать и ее PDB-идентификатор. Формат записей в SWISS-Prot состоит из следующих полей:

• ID/AC (accession number) − название записи и инвентарный номер. Иногда в данном поле могут присутствовать несколько различных номеров;

• DT − даты создания/обновления информации о записи;

• DE − поле описания (description) перечисляет все известные имена данного белка;

• GN (gene) − название гена (генов), кодирующих данный продукт;

• OS/OC/OX − содержат название организма, таксономическую классификацию и уникальный таксономический идентификатор организма, являющегося источником данной последовательности. Секция ссылок (RN/RP/RX/RA/RT/RL) включает все литературные ссылки, использованные для аннотации данной записи;

• CC − блок комментариев. Состоит из текста, разделенного на различные «темы» и описывающие: функцию белка, его внутриклеточную локализацию, посттрансляционные модификации, возможные связи с различными заболеваниями и т.д.;

• поле DR содержит кросс-ссылки на идентификаторы данного белка в других Бд (например, PDB);

• KW − поле, включающее ключевые слова (keywords), характеризующие данную запись.

• FT (features) − самое важное поле содержит список доменов и важных сайтов последовательности с указанием номеров (интервалов) аминокислотных остатков: описания посттрансляционных модификаций, вариантов последовательностей (известные замены остатков), доменную структуру, повторы, элементы вторичной структуры и т.д.

• SQ поле содержит саму аминокислотную последовательность белка.

Контрольные вопросы

1. Дайте определение понятия биоинформатика. Определите цель, задачи, предмет биоинформатики.
2. Какими свойствами обладает информация. Дайте им характеристику.
3. Где хранятся биоинформационные данные?
4. Базы данных и аналитические программы, их характеристика.
5. Что такое аннотирование последовательности и иденфикаторы записей в базах данных.
6. В каких видах деятельности реализуется предмет биоинформатики?
7. Перечислите медицинские применения биоинформатики.
8. Как записывается последовательность белка в формате FASTA?
9. Для чего предназначена программа BLAST?

Раздел II
МОЛЕКУЛЯРНАЯ ЭВОЛЮЦИЯ

2.1 Парное и множественное выравнивание. Алгоритм и программы выравнивания

Выравнивание последовательностей – это метод, основанный на размещении 2-х или более последовательностей мономеров ДНК, РНК или белков друг под другом таким образом, чтобы легко увидеть сходные участки в этих последовательностях.

Парное выравнивание используется для нахождения сходных участков 2-х последовательностей, различают (рисунок 2.1, 2.2):

– глобальное выравнивание предполагает, что последовательности гомологичны по всей длине, в глобальное выравнивание включаются обе входные последовательности целиком.

– локальное выравнивание применяется, если последовательности содержат как родственные (гомологичные), так и неродственные участки, результат локального выравнивания выбор участка в каждой из последовательностей и выравнивание между этими участками.

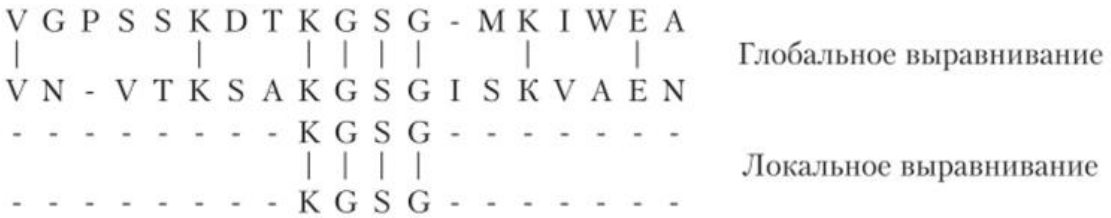


Рисунок 2.1 – Порядок глобального и локального выравнивание

Рисунок 2.2 – Праное выравниание

Парное выравнивание последовательностей – это наиболее простой случай множественного выравнивания последовательностей, т.е. способ расположения нескольких последовательностей друг под другом путем внесения в них пропусков, чтобы одинаковые или близкие по своим свойствам мономеры, формировали столбцы данного выравнивания.

Множественное выравнивание последовательностей (англ. multiple sequence alignment, MSA) – выравнивание трех и более биологических последовательностей, обычно белков, ДНК или РНК. В большинстве случаев предполагается, что входной набор последовательностей имеет эволюционную

связь. Используя множественное выравнивание, можно оценить эволюционное происхождение последовательностей, проведя филогенетический анализ (рисунок 2.3).

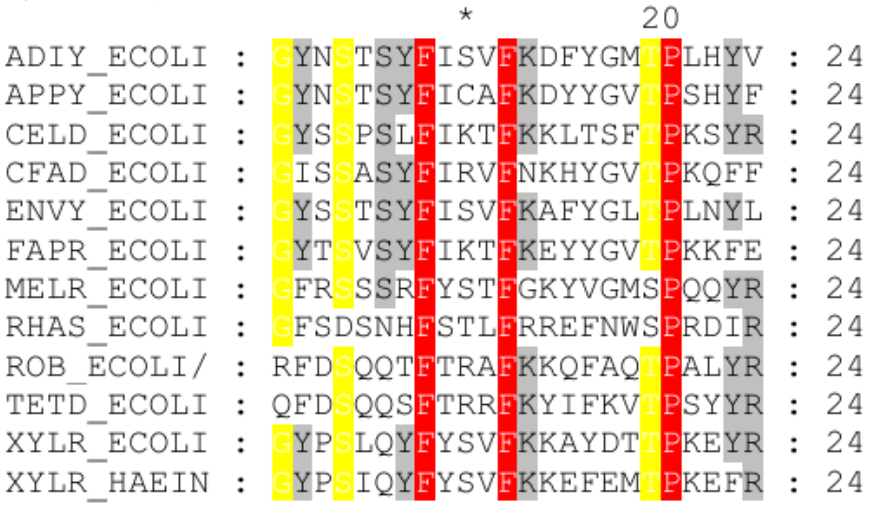


Рисунок 2.3 – Множественное выравнивание

Алгоритм Нидлмана-Вунша. Процент идентичности между двумя пептидными или нуклеотидными последовательностями является функцией количества аминокислот или нуклеотидных остатков, которые идентичны в двух последовательностях. Алгоритм Нидлмана-Вунша для глобального выравнивания последовательностей:

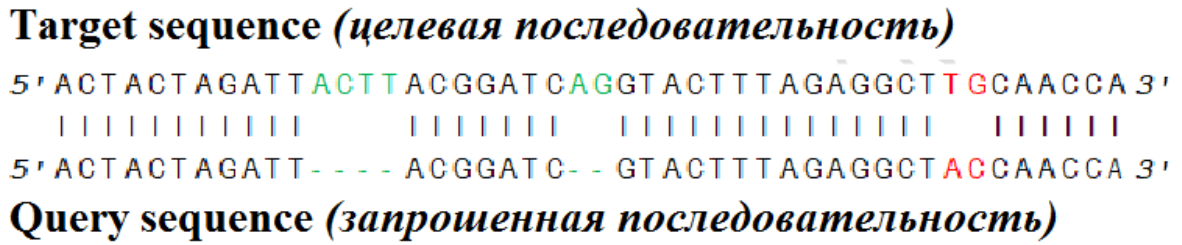


Рисунок 2.4 – Глобальное выравнивание последовательностей
по алгоритму Нидлмана-Вунша

В указанной последовательности есть совпадения (match), они выделены черным цветом, красным цветом выделены участки в которых нет совподений (mismatch), зеленый – пробелы в последовательности (gap). Пробелы и не совпадения штрафуются (например, им присваивается значение -1). За каждый участок последовательности присваиваются веса (score):

Match +1 +5

Mismatch -1 -1

Gap -1 -2

Для заполнения каждой ячейки нужно оценить 3 веса (score), для первой клетки Y - E:

1) слева направо: gap (-2)

2) сверху вниз: gap (-2)
3) по диагонали mismatch (-1)
ИТОГО: выбираем максимальное значение -1
Для второй клетки E – E, есть совпадение поэтому:
- по диагонали mismatch (1)
- слева направо: gap (-2)
- сверху вниз: gap (-2)

Итого: выбираем максимальное значение -1.

В верхнем левом углу выставляется ноль, от которого производится заполнение таблицы выравнивания по 3-м направлениям: вправо, вниз и по диагонали. Движение по диагонали подразумевает, что мы продвинулись на один шаг. Движение вправо или вниз подразумевает, что мы делаем gap. Идентичные остатки определяются как остатки, которые являются одинаковыми в двух последовательностях в данной позиции выравнивания.

Процент идентичности последовательности рассчитывается из оптимального выравнивания путем взятия числа остатков, идентичного между двумя последовательностями, деления его на общее количество остатков в самой короткой последовательность и умножение на 100.

Оптимальное выравнивание – это выравнивание, при котором процент идентичности является максимально возможным. Разрывы могут быть введены в одну или обе последовательности в одном или нескольких положениях выравнивания, чтобы получить оптимальное выравнивание учитываются как неидентичные остатки для расчета процента идентичности последовательности.

2.2 Эволюция молекул и организмов. Критерии сравнения нуклеотидных и белковых последовательностей

Жизнь – это способ существования открытых коллоидных систем, обладающих свойствами самовоспроизведения, регуляции и обновления на основе преобразования потоков веществ, энергии и информации путем взаимодействия белков и нуклеиновых кислот. Решающую роль в превращении неживого вещества в живое играют белки. Они способны образовывать коллоидные гидрофильные комплексы, сливаться друг с другом и образовывать коацерваты – открытые системами, обладают упорядоченностью, способностью к самообновлению и поглощают вещества из окружающей среды.

Возникновение коацерватов привело к естественному отбору (движущий фактор биологической эволюции). Включение в состав коацерватов ионов металлов привело к образованию ферментов. В результате включения в состав коацерватов нуклеиновой кислоты и ферментов сформировались предбиологические системы, т.е. смеси ДНК и белка (ДНК способна мутировать, а белки – ускорять химические реакции). В результате произошел пере-

ход от химической к биологической эволюции. На границе между коацерватами и внешней средой появилась клеточная мембрана. С образованием мембраны появились протобионты – первые примитивные клетки.

В ходе эволюции наиболее вероятной последовательностью появления живых организмов является: анаэробные гетеротрофы → фотоавтотрофы аэробные гетеротрофы → автотрофы. Первые организмы – гетеротрофы (прокариоты), окаменелые остатки и следы жизнедеятельности обнаружены в осадочных породах возрастом около 3,5 млрд. лет. Автотрофы возникли 3 млрд. лет назад (анаэробные бактерии, осуществляющие одностадийный фотосинтез).

Цианобактерии первые организмы, осуществившие 2-х стадийный фотосинтез с выделением кислорода. Постепенно атмосфера насытилась достаточным количеством кислорода (прекратилась химическая эволюция), появилась возможность кислородного типа обмена, что привело к появлению аэробов. Образование озонового экрана способствовало выходу организмов из водной среды на сушу.

2,5 млрд. лет назад появились протисты, а около 1,5 млрд. лет назад возникли многоклеточные организмы, которые усложнялись и сформировали типы животных и отделы растений.

Молекулярная эволюция – наука, изучающая изменения генетических макромолекул (ДНК, РНК, белков) в процессе эволюции, закономерности и механизмы этих изменений, а также реконструирующая эволюционную историю генов и организмов.

Объекты исследования молекулярной эволюции:

1. Последовательности НК как носителей генетической информации.
2. Последовательности белков.
3. Структура белков.
4. Геномы организмов.

Основными задачами молекулярной эволюции являются выявление закономерностей эволюции генетических макромолекул и реконструкция эволюционной истории генов и организмов. Молекулярная эволюция взаимосвязана с такими областями науки, как:

1) палеонтология (датировка эволюционных событий);
2) генетика (организация и передача наследственной информации);
3) молекулярная биология (строение генетических макромолекул);
4) эволюция (эволюционные закономерности);
5) биофизика (механизмы функционирования генетических макромолекул);
6) математика (построение моделей эволюции);
7) информатика (обработка и анализ данных);
8) биохимия (обмен нуклеиновых кислот и белков).

Разделами молекулярной эволюции как науки являются:

1. Эволюция макромолекул – изучает типы и скорости изменений, происходящих в генетическом материале (ДНК), а также созданных на его основе белков, и механизмов, ответственных за эти изменения.

2. Молекулярная филогения – изучает эволюционную историю макромолекул и организмов, получаемую на основе молекулярных данных. Эволюция макромолекул и молекулярная филогения тесно взаимосвязаны, и прогресс в одном из этих разделов способствует исследованиям в другом. Знание филогении нужно для определения последовательности изменений в изучаемых молекулах, а знание способов и темпов изменений изучаемой молекулы необходимо для восстановления эволюционной истории группы организмов.

3. Пребиотическая эволюция («происхождение жизни»), развитие этого раздела ограничивается тем, что в настоящее время неизвестны законы, направляющие процесс переноса информации в пребиотических системах.

2.3 Гомология последовательностей. Ортологи и паралоги

При сравнительном описании любой формы или типа подобия используется термин «гомология» (от греч. ομολογειν – «согласующийся»). В биологии термин «гомология» используют для описания какого-либо подобия между органами, признаками, генами и геномами. Например, считают гомологичными крыло птицы и конечность крокодила, так как предполагается, что они представляют собой парные придатки, имеющие общий план строения, несмотря на их значительные структурные и функциональные различия.

Ричард Оуэном (Richard Owen, 1804–1892) предложил различать термины гомологию и аналогию для определения принципа, по которому можно ранжировать виды в естественной системе. Ученый исходил из того, что факта того, что конечности всех тетрапод имеют одинаковый план строения и общий план строения конечностей можно проследить, несмотря на различие их функций, таких, как ходьба, лазанье, плаванье, рытье или полет (рисунок 2.5).

Р. Оуэн описал и каталогизировал виды путем идентификации подобных частей у разных видов и их ранжированием по паттернам общих частей тела (т. е. гомологичным признакам), невзирая на их функцию. В результате возникла система, в которой гомологичные признаки являлись общими для совокупного набора видов («мутовки видов»). Р. Оуэн полагал, что гомологичные признаки идентичны из-за своего происхождения при сохранении общего плана строения тела, или архетипа. Архетип может быть понят как абстрактный или систематизирующий принцип живой природы. Согласно интерпретации Р. Оуэна, каждый вид есть частная реализация архетипа.

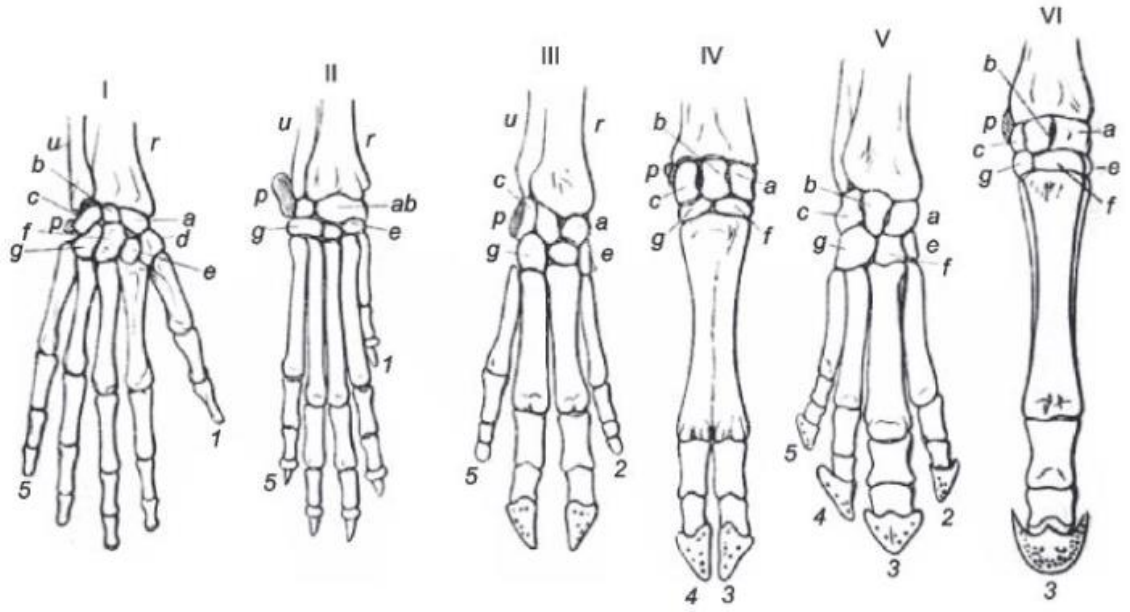


Рисунок 2.5 – Схема строения скелета с указанием гомологичных костей (из: Gegenbaur, 1870).
I – человек, II – собака, III – свинья, IV – корова, V – тапир, VI – лошадь.
r – radius; u – ulna; a – scaphoid; B – lunare; c – triquetrum; d – trapezium; e – trapezoid; f – capitatum; g – hamatum; p – pisiforme.

Из систематизирующего принципа архетипа Р. Оуэна следует, что между последовательностями ДНК, РНК, белков организмов близких рангов возможно установление степени родства путем их сравнения и оценки различий эволюционных или случайных отклонений. В БИ для определения значимости совпадений или различий введены свои определения для терминов сходство (подобие) и гомология: сходство – это наличие или измерение сходства или различия, независимо от источника сходства, гомология означает, что последовательности или организмы, в которых они обнаружены, являются потомками общего предка.

О подобии последовательностей можно судить, проведя процедуру их выравнивания, а о гомологии организмов (или органов) – на основании наблюдаемого подобия. Здесь гомологии – это предположение, которое возникает из наблюдения подобия. Гомология между ДНК, РНК или белками обычно определяется по сходству их нуклеотидных или аминокислотных последовательностей. Вспомним, что выравнивания последовательностей используются, чтобы указать, какие участки каждой последовательности гомологичны. Значительное сходство является доказательством того, что две последовательности связаны эволюционными изменениями от общей предковой последовательности.

Причины гомологии: ортология, паралогия, аналогия

Два сегмента ДНК могут иметь общее происхождение из-за трех явлений:

– ортологии – события видообразования,

– паралогии – событие дублирования,

– ксенологи – горизонтального (или латерального) переноса генов.

1. Термин «ортолог» придумал в 1970 году молекулярный эволюционист Уолтер Фитч. Гомологические последовательности являются ортологичными, если предполагается, что они произошли от одной и той же предковой последовательности, разделенной событием видообразования.

Ортологи – это гены у разных видов, которые произошли в результате вертикального спуска от одного гена последнего общего предка. Например, регуляторный белок гриппа растений есть и у растения Arabidopsis, и у хламидомонады Chlamydomonas. Вариант Chlamydomonas более сложен: он дважды пересекает мембрану, а не один раз, содержит дополнительные домены и подвергается альтернативному сплайсингу1. Однако он может полностью заменить гораздо более простой белок Arabidopsis, если перенести его из водорослей в геном растения с помощью генной инженерии. Значительное сходство последовательностей и общие функциональные домены указывают на то, что эти два гена являются ортологичными генами, унаследованными от общего предка (рисунок 2.6).

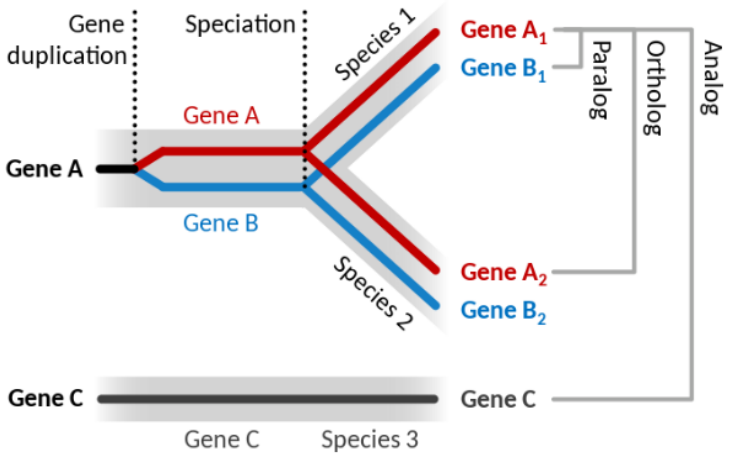


Рисунок 2.6 – Явления орто-, пара-, и аналогии

Примечание к рисунку: Вверху: наследственный ген дублируется, чтобы произвести два паралога (гены A и B). Событие видообразования приводит к появлению ортологов у двух дочерних видов. Внизу: у отдельного вида неродственный ген выполняет аналогичную функцию (Ген C), но имеет отдельное эволюционное происхождение и является аналогом.

Учитывая, что точное происхождение генов у разных организмов трудно установить из-за дупликации генов и событий перестройки генома, наиболее убедительные доказательства того, что два похожих гена являются ортологами, обычно обнаруживаются путем проведения филогенетического анализа происхождения генов. Ортологи часто, но не всегда, выполняют одну

и ту же функцию. Ортологические последовательности дают полезную информацию для таксономической классификации и филогенетических исследований организмов.

Паттерн генетической дивергенции может быть использован для отслеживания родства организмов. Два очень тесно связанных организма, вероятно, будут иметь очень похожие последовательности ДНК между двумя ортологами. Напротив, организм, который далее эволюционно отделен от другого организма, вероятно, будет демонстрировать большее расхождение в последовательности изучаемых ортологов.

2. Паралоги – это гены, которые связаны между собой посредством событий дупликации в последнем общем предке (Last common ancestor LCA) сравниваемых видов. Они возникают в результате мутации дублированных генов во время отдельных событий видообразования. Когда потомки от LCA имеют общие мутировавшие гомологи исходных дублированных генов, эти гены считаются паралогами.

Например, в LCA один ген (ген A) может быть продублирован, чтобы создать отдельный похожий ген (ген B), эти два гена будут продолжать передаваться последующим поколениям. Во время видообразования одна среда будет способствовать мутации в гене A (ген A1), создавая новый вид с генами A1 и B. Затем в отдельном событии видообразования одна среда будет благоприятствовать мутации в гене B (ген B1), приводящей к возникновению новый вид с генами A и B1 (рисунок 2.7).

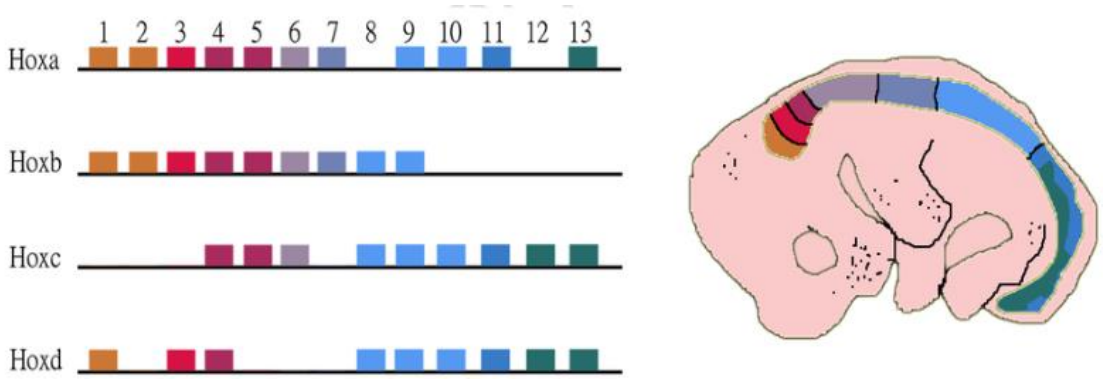


Рисунок 2.7 − Паралогия гена Hox

Гены потомков A1 и B1 паралогичны друг другу, потому что они являются гомологами, которые связаны посредством события дупликации у последнего общего предка двух видов. Дополнительные классификации паралогов включают:

– аллопаралоги (внешние паралоги),

– симпаралоги (внутренние паралоги).

Аллопаралоги – это паралоги, которые произошли от дупликаций генов, предшествовавших данному событию видообразования, они возникли

в результате событий дублирования, которые произошли в LCA сравниваемых организмов.

Симпаралоги – это паралоги, возникшие в результате дублирования генов паралогов в последующих событиях видообразования. Если потомок с генами A1 и B претерпел другое событие видообразования, в котором дуплицировался ген A1, у нового вида были бы гены B, A1a и A1b – гены A1a и A1b являются симпаралогами.

Пример, гены Hox позвоночных организованы в наборы паралогов. Hox гены – это подмножество родственных гомеобоксных генов, задающих области плана тела в качестве зародыша вдоль головки хвоста оси животных. Каждый Hox-кластер (HoxA, HoxB и т.д.) находится на отдельной хромосоме.

Например, кластер HoxA человека находится на хромосоме 7, показанный на рисунке 2.3 кластер мыши HoxA имеет 11 паралоговых генов (2 отсутствуют).

Паралогичные гены могут формировать структуру целых геномов и, таким образом, в значительной степени объяснять эволюцию генома. Примеры включают гены Homeobox (Hox) у животных – эти гены претерпели не только дупликации генов в хромосомах, но и дупликации всего генома. В результате гены Hox у большинства позвоночных сгруппированы по множеству хромосом.

Другой пример – гены глобина, которые кодируют миоглобин и гемоглобин и считаются древними паралогами. Точно так же четыре известных класса гемоглобинов (гемоглобин A, гемоглобин A2, гемоглобин B и гемоглобин F) являются паралогами друг друга. Хотя каждый из этих белков выполняет одну и ту же основную функцию переноса кислорода, они уже немного разошлись по функциям: гемоглобин плода (гемоглобин F) имеет более высокое сродство к кислороду, чем гемоглобин взрослого человека.

3. Ксенология – возникновение гомологичных ДНК-последовательностей в геномах разных видов при «горизонтальном» (ненаследственном) переносе генов между организмами. Горизонтальный перенос происходит при физическом контакте клеток, обменивающихся генетическим материалом, т.е. в паразитарных, симбиотических, ассоциативных системах. Ксенологичные гены (ксенологи) обнаруживаются у филогенетически отдаленных, но территориально близких групп клеток или организмов.

В качестве носителей ксенологичной ДНК выступают ретровирусы, захватывающие фрагменты оттранслированной в РНК ДНК клетки-хозяина одного вида и встраивающих эти последовательности в геном клеток-хозяев другого вида: плазмиды при конъюгации, бактериофаги при трансдукции, содержащаяся в среде свободная ДНК при трансформации.

2.4 Горизонтальный перенос генов. Филогенетическое дерево и методы его построения (UPGMA, NEIGHBOR-JOINING, MINIMAL EVOLUTION)

Генетический перенос – это передача генетической информации от одного организма другому. Это происходит при помощи двух механизмов: вертикального переноса генов и горизонтального переноса генов. *Вертикальный перенос генов* происходит, когда генетическая информация передается от одного поколения следующему, что происходит гораздо чаще, чем горизонтальный перенос генов. И половое, и бесполое размножение являются формами вертикального переноса генов, когда один или несколько организмов передают часть генома или весь свой геном потомству. Кроме того, вертикальный перенос генов происходит как у прокариотических, так и у эукариотических видов.

Горизонтальный перенос генов происходит, когда генетическая информация передается представителю того же поколения, и чаще всего встречается у прокариотических видов. В то время как среди эукариот межвидовой горизонтальный перенос генов чрезвычайно редок, у прокариот он часто встречается. Горизонтальный перенос генов между различными видами – важный источник генетического разнообразия у прокариот.

Большинство прокариотических видов размножаются бесполым путем. Хотя это позволяет быстрее производить потомство, полученные потомки обладают ограниченным генетическим разнообразием. Горизонтальный перенос генов, таким образом, играет жизненно важную роль в обеспечении генетического разнообразия прокариот. Посредством горизонтального переноса генов прокариоты могут делиться небольшой частью своего генома с другими организмами, конспецифичными (принадлежащими тому же виду) или гетероспецифичными (принадлежащими другому виду), в одном поколении. Многие ученые полагают, что горизонтальный перенос генов и мутации являются наиболее значительными источниками генетической изменчивости прокариот. Таким образом, горизонтальный перенос генов обеспечивает некий исходный материал, на который действует естественный отбор. Ярким примером этого является появление устойчивых к антибиотикам штаммов бактерий. Гены, придающие устойчивость к антибиотику, могут передаваться между различными видами и штаммами бактерий, что дает бактериям-реципиентам избирательное преимущество, например, устойчивые к пенициллину штаммы *Neisseria gonorrhoeae*, вызывающие гонорею. Более того, по некоторым оценкам, по меньшей мере 18% генома *E. coli* было приобретено посредством горизонтального переноса генов в течение миллионов лет эволюции.

Филогенетика – наука, изучающая процессы образования биоразнообразия.

Кладистика – направление филогенетической систематики. Характерные особенности кладистической практики состоят в использовании так

называемого кладистического анализа (строгой схемы аргументации при реконструкции родственных отношений между таксонами), строгом понимании монофилии и требовании взаимно-однозначного соответствия между реконструированной филогенией и иерархической классификацией. Кладистический анализ – основа большинства принятых в настоящее время биологических классификаций, построенных с учетом родственных отношений между живыми организмами. Кладистика относится к числу трех ведущих таксономических школ, доминирующих в современной биологической систематике. Ей противостоят фенетика, основанная на количественной оценке так называемого общего сходства, и эволюционная таксономия, которая, подобно кладистике, при построении системы опирается на эволюционную близость (то есть общность происхождения), однако не требует строгого соответствия системы и филогении (в частности, это выражается в признании права на существование в системе парафилетических групп).

Филогенетическое дерево (**эволюционное дерево**, **дерево жизни**) – дерево, отражающее эволюционные взаимосвязи между различными видами или другими сущностями, имеющими общего предка.

Вершины филогенетического дерева делятся на три класса: листья, узлы и (максимум один) корень. Листья – это конечные вершины, то есть те, в которые входят ровно по одному ребру; каждый лист отображает некоторый вид живых организмов (или иной объект, подверженный эволюции, например, домен белка). Каждый узел представляет эволюционное событие: разделение предкового вида на два или более, которые в дальнейшем эволюционировали независимо. Корень представляет общего предка всех рассматриваемых объектов. Ребра филогенетического дерева принято называть «ветвями» (рисунок 2.4).

Идея «дерева» появилась в ранних взглядах на жизнь, как на процесс развития от простых форм к сложным. Современные эволюционные биологи продолжают использовать деревья для иллюстрации эволюции, так как они наглядно показывают взаимосвязи между живыми организмами (рисунок 2.8).

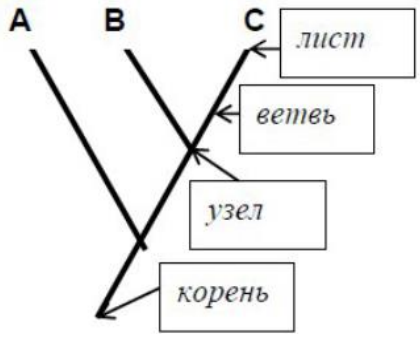


Рисунок 2.8 – Филогенетическое дерево

Примечание: листья – рассматриваемые объекты (A, B, C); узлы – точки схождения ветвей (узел, указанный стрелкой, объединяет объекты B и C, а нижележащий узел уже объединяет группу B+C и объект A); ветви – линии, соединяющие листья с узлами и узлы друг с другом; корень – узел, объединяющий все рассматриваемые объекты в одну группу.

Контрольные вопросы

1. Что такое генетический код?
2. Дайте определение выравнивание последовательностей.
3. Охарактеризуйте парное и множественное выравнивание. Какие алгоритмы и программы существуют для выравнивания.
4. Что такое глобальное и локальное совпадение при выравнивании последовательности?
5. В чем заключается поиск мотивов совпадения при выравнивании последовательности?
6. В чем заключается сущность алгоритма Нидлмана-Вунша?
7. Какие выделяют критерии сравнения нуклеотидных и белковых последовательностей?
8. Дайте характеристику ортологам и паралогам.
9. Какие органы называют гомологичными?
10. Что такое горизонтальный перенос генов?
11. Какие методы построения филогенетического дерева существуют.

Раздел III
МЕТОДЫ АНАЛИЗА ПОСЛЕДОВАТЕЛЬНОСТЕЙ НУКЛЕИНОВЫХ КИСЛОТ

3.1 Гены и регуляция их экспрессии

Ген – это участок ДНК, кодирующий всего один белок или РНК, кроме непосредственно кодирующей части, он также включает в себя регуляторные и иные структурные элементы, имеющие разное строение у прокариот и эукариот.

Экспрессия генов – процесс, в ходе которого наследственная информация от гена (последовательности нуклеотидов ДНК) преобразуется в функциональный продукт – РНК или белок. Некоторые этапы эксперессии генов могут регулироваться: это транскрипция, трансляция, спайсинг РНК и стадия посттрансляционных модификации белков. Процесс активации экспрессии генов короткими двуцепочечными РНК называется **активацией РНК.**

Регуляция экспрессии генов позволяет клеткам контролировать основой дифференцировки клеток, морфогенеза и адаптации. Экспрессия генов является субстратом для эволюционных изменений, так как контроль над временем, местом и количественными характеристиками экспрессии одного гена может иметь влияние на функции других генов в целом организме.

Как осуществляется регуляция экспрессии генов?

Экспрессия генов – это реализация заложенной в них информации, то есть синтез РНК и белков. Другими словами, под экспрессией генов понимают их активность.

В клетках живых организмов экспрессия генов регулируется: одни гены могут быть реализованы, другие – нет. Причем регуляция может осуществляться на разных этапах: может выполняться или нет транскрипция, из пре-мРНК в результате альтернативного сплайсинга могут образовываться разные мРНК, может блокироваться трансляция и др.

У эукариот, обладающих отграниченным от цитоплазмы ядерным содержимым и более сложным геномом, регуляция экспрессии генов намного разнообразнее и сложнее, чем у прокариот.

У прокариот пока молекула РНК синтезируется на участке ДНК, она тут же может транслироваться (начиная с уже синтезированного конца). Поэтому у них регуляция экспрессии (активности) генов осуществляется почти исключительно на уровне ДНК, так как в РНК часто невозможно внести какие-нибудь изменения до ее трансляции.

В 1961 г. Жакобом и Моно была предложена **модель оперона** как системы регуляции генов у бактерий. **Оперон состоит из промотора, оператора, структурных генов оперона (их может быть разное количество) и терминатора.** В области промотора прикрепляется фермент РНК-полимераза. В области оператора присоединяется белок-репрессор, который кодируется отдельно отстоящим от оперона геном-регулятором (может быть сцеплен со своим опероном, а может находиться на расстоянии).

Если белок-репрессор соединяется с оператором, то транскрипция всех структурных генов оперона становится невозможной, так как РНК-полимераза не может перемещаться по цепи ДНК.

В свою очередь активность белка-репрессора может блокироваться определенным для него низкомолекулярным соединением – индуктором (тем или иным питательным веществом бактерий). В результате взаимодействия с индуктором белок-репрессор видоизменяется и уже не может присоединиться к оператору своего оперона. В этом случае гены оперона экспрессируются (т.е. на них идет синтез).

Бывает обратная ситуация, когда индуктор активирует белок-репрессор.

Таким образом, в зависимости от того, какие индукторы находятся в цитоплазме, у прокариот экспрессируются те или иные генные группы.

Вышеописанный механизм экспрессии генов относится к **негативной регуляции**, так как гены транскрибируются, если они не выключены репрессором. И наоборот: не транскрибируются, если выключены.

Кроме негативной регуляции у бактерий существует также **позитивная.** В этом случае вместо белка-репрессора действие оказывает белок-активатор. На эти белки также действуют индукторы, активируя или инактивируя их.

Также у прокариот были выявлены опероны, которые актируются двумя регуляторными белками, соединенными друг с другом.

У многоклеточных организмов в клетках разных тканей экспрессируются разные гены, т. е. для **эукариот** характерна *дифференциальная экспрессия*.

У эукариот, также как и у прокариот, существуют регуляторные белки с похожим механизмом действия. При этом для эукариот не характерна регуляция по типу оперона. Цистроны (транскрибируемые участки) эукариот обычно содержат по одному гену (это не касается геномов хлоропластов и митохондрий).

Кроме регуляторных белков, взаимодействующих с ДНК, у эукариот существуют и другие способы регуляции экспрессии генов.

Конденсация и деконденсация хроматина. Это наиболее универсальный метод регуляции транскрипции. Когда нужно экспрессировать определенные гены, хроматин в этом месте деконденсируется.

Альтернативные промоторы. У гена может быть несколько промоторов, каждый из которых начинает транскрипцию с разных его экзонов в зависимости от типа клетки. В конечном итоге будут синтезированы разные белки.

Метилирование и деметилирование ДНК. Метилирование ДНК происходит в регуляторных областях гена. Метилируется цитозин в последовательности ЦГ, после чего ген инактивируется. При деметилировании активность гена восстанавливается. Процесс регулируется ферментом метилтрансферазой.

Гормональная регуляция. При гормональной регуляции гены активируются в ответ на внешний химический сигнал (поступление в клетку определенного гормона). Этот гормон запускает те гены, которые имеют специфические последовательности нуклеотидов в регуляторных областях.

Геномный импринтинг. Это малоизученный способ регуляции экспрессии генов у эукариот. Он возможен только у диплоидных организмов и выражается в том, что активность генов зависит, от какого из родителей они были получены. Выключение генов осуществляется путем метилирования ДНК.

Альтернативный сплайсинг. Это регуляция на уровне процессинга. При альтернативном сплайсинге порядок сшивки экзонов может быть различным. Отсюда следует, что на основе одной и той же нуклеотидной последовательности ДНК могут быть синтезированы разные белки. Хотя их отличие друг от друга будет в основном заключаться лишь в разных сочетаниях одних и тех же аминокислот.

Тканеспецифическое редактирование РНК также протекает на уровне процессинга. Выражается в замене отдельных нуклеотидов в РНК в определенных тканях организма.

Кроме того, у эукариот иРНК часто не подвергается процессингу вообще (а распадается) или подвергается с задержкой. Это также можно рассматривать как способ регуляции экспрессии генов.

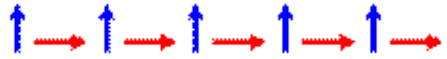
Регуляция стабильности иРНК. У эукариот существует регуляция и на уровне трансляции, когда готовые иРНК не «допускаются» к рибосомам или разрушаются. Другие же иРНК могут дополнительно стабилизироваться для многократного использования.

Посттрансляционная модификация белка. Чтобы молекула полипептида превратилась в активную молекулу белка, в ней должны произойти различные модификации определенных аминокислот, должны быть сформированы вторичная, третичная и возможно четверичная структуры. На этом этапе также можно повлиять на реализацию генетической информации, например, не дав молекуле сформироваться.

Риборегуляторы. Были обнаружены РНК, выполняющие регуляторные функции путем ослабления работы отдельных генов.

Для высокоорганизованных животных отмечается существование надклеточного уровня регуляции экспрессии генов.

Генная экспрессия – это совокупность молекулярных механизмов реализации наследственной информации, благодаря которому, ген проявляет свой потенциал в конкретном фенотипическом признаке организма. Все этапы экспрессии генов протекают с использованием энергии и обслуживаются десятками разнообразных ферментов. Процесс экспрессии гена состоит из нескольких этапов:



Ген → Про-мРНК → мРНК → Полипептид → Белок → Признак

транскрипция → процессинг → трансляциия → модификация → экспрессия

а) на основе гена ДНК синтезируется про-мРНК. Первый этап экспрессии называется «транскрипцией»;

б) крупная молекула про-мРНК подвергается «процессингу», в результате этого значительно уменьшается в размерах. Образуется «зрелая» мРНК, считывание информации с которой упрощается. Биологический смысл процессинга – облегчение доступа к генетической информации;

в) мРНК при участии тРНК «выбирает» необходимые аминокислоты, которые связываются на рибосоме в строго определенную последовательность полипептида. Процесс переноса информации с мРНК на полипептид называется «трансляцией»;

г) синтезированный полипептид подвергается «модификации» и превращается в активный белок;

д) функционируя, белок делает свой вклад в морфологический или функциональный признак (фенотип) клетки или организма. Это процесс называется «экспрессией».

В процессе транскрипции участвует не только смысловая часть гена, но и другие регуляторные и структурные части. Образуемая про-мРНК содержит все элементы, характерные для гена ДНК. Процессинг существенно модифицирует про-мРНК, которая превращается в мРНК и содержит намного меньше структурно-функциональных элементов. На основе мРНК

трансляция создает молекулы совершенно другой природы – полипептиды, ничего не имеющие общего с нуклеиновыми кислотами и обладающими совершенно другими свойствами и организацией. Модификация полипептидов приводит к еще одному природному явлению – появлению сложной пространственной организации молекулы белка. Происходит переход линейной информации ДНК и РНК в пространственную организацию протеина, которая, в свою очередь, является основой специфического пространственного взаимодействия молекул в живом организме, что и лежит в основе жизни и всех жизненных явлений. В данном случае процесс модификации обеспечивает пространственную организацию – объединение четырех субъединиц гемоглобина в единый комплекс. В результате всех этапов экспрессии проявляется признак – способность к транспорту газов (O_2 и CO_2).

Оперон – это последовательность специальных, функциональных сегментов ДНК, а также структурных генов, которые кодируют и регулируют синтез определенной группы белков одной метаболической цепи, например, ферментов гликолиза. *Оперон (регулируемая единица транскрипции)* состоит из следующих структурных частей (специальных последовательностей нуклеотидов) (рисунок 3.1):

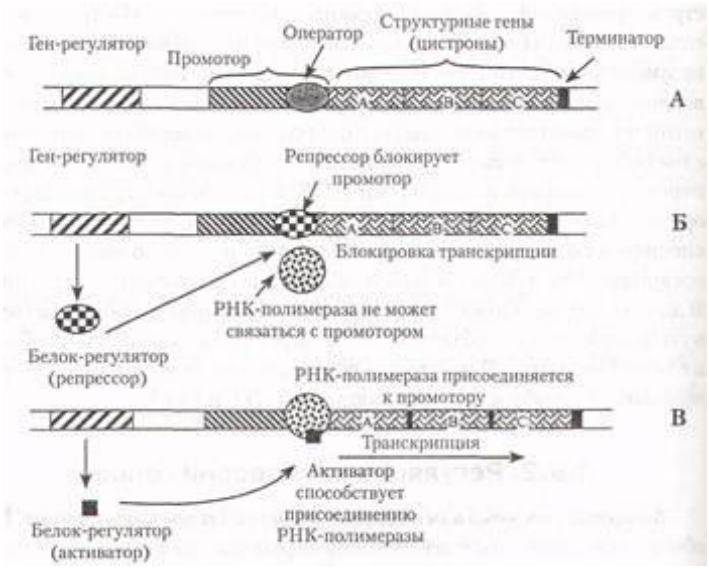


Рисунок 3.1 – Структурные части оперона

1. Ген-регулятор, контролирующий образование белка-регулятора.

2. Промотор – участок ДНК, к которому присоединяется РНК-полимераза и начинается транскрипция.

3. Оператор – участок промотора, связывающий белок-регулятор.

4. Структурные гены (цистроны) – участки ДНК, кодирующие мРНК конкретных белков.

5. Терминаторный участок ДНК несет сигнал об остановке транскрипции.

3.2 Анализ последовательностей ДНК

Несмотря на научный прогресс, ДНК-анализ остается технологически сложной процедурой, ведь речь идет об исследовании микроскопических структур, состоящих из миллионов «букв», последовательность которых следует установить. Это называется секвенированием ДНК.

Для того чтобы «прочитать» молекулу, ее для начала нужно выделить, затем многократно скопировать, а после – «нарезать» на небольшие кусочки, удобные для анализа. Азотистые основания, составляющие «алфавит» генетического кода, при этом окрашиваются особым флуоресцентным красителем, благодаря которому в ходе дальнейшего анализа их можно будет распознать при просвечивании лазером.

Некоторые современные методы секвенирования ДНК напоминают компьютерную томографию, в ходе которой одна-единственная нить ДНК проходит через нанопору, а компьютер фиксирует изменения ионного тока в единицу времени, что позволяет распознать каждую «букву» генетического кода.

Технологии продолжают совершенствоваться, что позволяет ускорить и удешевить процесс ДНК-анализа. Это достигается за счет одномоментного «чтения» сразу нескольких участков ДНК, разработки нового программного обеспечения и усложнения конструкции приборов, предназначенных для автоматического секвенирования, − секвенаторов.

Так, с момента появления первых секвенаторов в 90-е годы минувшего столетия ученым удалось снизить стоимость ДНК-анализа в 13 раз.

Наибольшую известность получили следующие методики секвенирования ДНК:

– метод Сэнгера (метод терминации цепи);

– пиросеквенирование (секвенирование путем синтеза);

– секвенирование на основе лигирования;

– секвенирование ДНК одиночных молекул;

– нанопоровое секвенирование (тот самый «ДНК-томограф»).

Что показывает анализ ДНК? Возможности ДНК-анализа выходят за пределы подтверждения уже установленного другими методами диагноза − с его помощью можно узнать о заболеваниях, которые еще никак себя не проявили, но при стечении определенных обстоятельств могут серьезно пошатнуть здоровье пациента.

С помощью анализа ДНК определяется:

– наследственная предрасположенность к конкретным патологиям, которые уже встречались в семье (например, раку или психическим заболеваниям);

– общий «генетический анамнез» человека, который желает знать наверняка, какие заболевания могут возникнуть у него в будущем;

– причина неясных симптомов в отсутствии возможности поставить диагноз иным путем (особенно актуально для детей с редкой генетической патологией);

– индивидуальная непереносимость определенных лекарственных препаратов;

– степень генетического родства с предполагаемыми членами семьи;

– вероятность осложнений во время беременности;

– склонность к алкоголизму или наркомании (на основании выявления генов, ответственных за синтез ферментов, способных расщеплять алкоголь и другие соединения);

– риски при наличии серьезных физических нагрузок (важно для профессиональных спортсменов);

– возможные причины бесплодия и т.д.

Благодаря ДНК-анализу будущие родители могут оценить вероятность рождения ребенка с наследственной патологией, а в случае, если заболевание уже выявлено у малыша, − с первых месяцев его жизни выработать оптимальный план лечения, который позволит избежать осложнений.

Сроки и стоимость ДНК-анализа пропорциональны объему исследования: чем больше информации вы хотите получить, тем больший объем ДНК предстоит «прочитать» специалистам, сопоставив полученные данные с имеющимися сведениями о функциях обнаруженных генов.

Срок проведения анализа составляет в среднем 3–4 недели − в зависимости от того, где проводится тест (образцы могут быть отправлены в другую лабораторию, где есть соответствующее оборудование), какие реактивы потребуются для его проведения и каков объем предполагаемой работы.

Наиболее широко используемый метод для ДНК-анализа – секвенирование по Сангеру.

В основе метода секвенирования ДНК, разработанного Сэнгером и соавт., называемого также методом секвенирования путем терминации цепи, лежал принцип ферментативного построения комплементарной цепи ДНК по существующей одноцепочечной матрице при происходящем в разных местах цепи ДНК ингибировании ее дальнейшего роста.

Первый метод секвенирования ДНК, предложенный Ф. Сэнгером и А. Коулсоном в 1975 г, основан на ферментативных реакциях и носит название плюс-минус -метод. Данный подход предполагает выделение одноцепочечного фрагмента ДНК, соответствующего исследуемому участку генома. Этот фрагмент используют затем в реакции полимеразного копирования в качестве матрицы, а в качестве праймера − синтетические олигонуклеотиды или природные субфрагменты, получаемые после гидролиза определенными рестриктазами.

В настоящее время существует множество вариантов метода Сэнгера. Главное, этот метод удалось полностью автоматизировать. Так, например, при секвенировании ДНК по Сэнгеру на 5-конец праймера вводят

флуоресцентные метки, причем для каждого из четырех анализируемых нуклеотидов используются флуоресцирующие агенты с различными спектральными характеристиками. После электрофоретического разделения гель сканируется при четырех различных длинах волн и полученная информация сразу обрабатывается на ЭВМ. При этом все биохимические операции также проводятся роботом.

Хотя сам принцип специфической терминации, положенный в основу первоначального метода секвенирования ДНК с помощью дидезокситерминаторов, остался неизменным, само секвенирование ДНК по Сэнгеру все же стало в значительной степени другим.

Наиболее широко сейчас применяется метод ферментативного секвенирования, или метод секвенирования путем терминации (остановки синтеза) цепи, предложенный Ф. Сэнгером в 1977 г.

Секвенирование ДНК по Сэнгеру. Это наиболее употребительный метод секвенирования ДНК, поскольку по сравнению с химическим методом он позволяет анализировать более крупные фрагменты ДНК с меньшей вероятностью ошибки. Секвенируемую ДНК сначала клонируют в одноцепочечный бактериофаговый вектор. Чаще всего используют фаг M13 и его производные. Одноцепочечная ДНК взаимодействует как субстрат с ДНК-полимеразой, которая точно копирует первую цепь. Однако если нормальный дезокси-нуклеозидфосфат заменить его аналогом-дидезоксинуклеозидфосфатом, то дальнейшее функционирование полимеразы прекращается и рост синтезируемой цепи останавливается. Для инициирования этого процесса используют химически синтезированный универсальный нуклеотидный праймер. Реакцию проводят в присутствии четырех аналогов dNTP, один из которых содержит метку P. Исходно имеется четыре реакционные смеси, в каждой из которых понижена концентрация одного из аналогов меченого нуклеотида. Поэтому в результате реакции получаются смеси меченных радионуклидами фрагментов ДНК, один конец у которых одинаков, но при этом они имеют разную длину и разные основания на другом конце. После инкубации ДНК в каждой смеси превращается в одноцепочечную форму. Затем смеси подвергают электрофорезу в полиакриламидном геле. Длину и последовательность фрагментов ДНК можно считывать непосредственно с гелевого слоя. Этот метод позволяет секвенировать в одной серии реакции от 250 до 500 пар оснований. Полученные данные часто анализируют с помощью компьютера.

Ключевым моментом ферментативного метода секвенирования ДНК, разработанного Сэнгером и соавт., является образование специфически терминированных меченых фрагментов вновь синтезированной ДНК. Такая терминация построения комплементарной цепи ДНК происходит при включении ДНК-полимеразой в растущую цепь ДНК-модифицированных аналогов природных.

В ферментативном методе секвенирования ДНК по Сэнгеру каждая дорожка соответствует только одному конкретному типу оснований и поэтому для правильного прочтения радиоавтографа секвенирующего геля необходимо следить, чтобы какая-либо полоса содержалась только в одной дорожке, поскольку в противном случае точное установление данного нуклеотида будет затруднено или скорее просто невозможно. В методе Сэнгера нет необходимости сопоставления наличия полос пуриновых или пиримидиновых оснований в одной или двух дорожках, то можно считать, что чтение радиоавтографа секвенирующего геля при этом будет проще.

Однако, существует ряд ограничений, из-за которых метод секвенирования ДНК гибридизацией так и не стал основным. Главными препятствиями этому служат серьезные проблемы при секвенировании повторяющихся элементов генома и чтение относительного малого числа нуклеотидов в ходе одного эксперимента.

3.3 Компьютерные программы, используемые для анализа секвенированных последовательностей генов

Появление быстрых методов секвенирования ДНК ферментативным построением новой цепи ДНК в условиях терминации по Сэнгеру привело к резкому увеличению числа секвенированных фрагментов ДНК. Определяемые в ходе секвенирования нуклеотидные последовательности, в виде так называемых ДНКовых текстов, повлекли за собой разработку специализированных компьютерных программ по их анализу, поскольку обработка таких больших массивов данных без помощи компьютеров стала просто невозможна. Однако первые появившиеся компьютерные программы обращения с нуклеотидными последовательностями и анализа ДНК характеризовались минимальным набором сервисных функций. Стремительный рост числа разнообразных программ, рассчитанных на проведение тех или иных операций с нуклеотидными последовательностями ДНК, даже потребовал выделения отдельных номеров журнала «Nucleic Acids Research», целиком посвященных данной проблеме.

Многие компьютерные программы тех лет представляли собой небольшие программки для решения конкретных задач, вроде поиска сайтов рестрикционных эндонуклеаз, определения размеров получающихся фрагментов после расщепления ими молекул ДНК или определения молекулярной массы таких фрагментов, их нуклеотидного состава. С целью некоторого упорядочения информации обо всех этих программах и лучшей ориентации исследователей был подготовлен специальный указатель, вобравший в себя максимально возможное число известных к тому времени компьютерных программ и дающий краткое описание их возможностей. Однако становилась очевидной насущная потребность создания так называемых пакетов прикладных программ, которые бы позволяли проводить

целый необходимых операций по всевозможному анализу секвенированных фрагментов ДНК, начиная от занесения последовательности нуклеотидов в компьютер до выявления особенностей кодируемых ими белков.

Другим аспектом компьютерного анализа секвенированных молекул ДНК стал вопрос хранения полученных данных и необходимость обеспечения широкого доступа ученых к уже известным нуклеотидным последовательностям. Это привело к образованию специализированных белковых данных, сначала одного, потом нескольких и уже затем многочисленных. В настоящее время, кроме трех основных, так называемых первичных банков данных (GebBank, EMBL, DDBJ), главной целью ставящих сбор нуклеотидных последовательностей, существует еще множество баз данных, преследующих какую-либо цель.

Чтобы выявить какие-то особенности или характерные черты исследуемого гена или фрагмента ДНК, необходимо провести анализ его нуклеотидной последовательности, ставшей известной в результате секвенирования. Причем, зачастую требуется всесторонний анализ, который более удобно осуществить с помощью интегрированного пакета специализированных программ. В настоящее время имеется широкий выбор различных пакетов таких программ, отличающихся требованиями, предъявляемыми ими к самой компьютерной технике, к операционным системам. В силу особенностей компьютерного парка нашей страны, представленного подавляющим количеством IBM-совместимых компьютеров, программы, написанные для других типов машин, здесь, за редкими исключениями, упоминаться не будут.

Еще один уже упоминавшийся выше модуль MapDraw из пакета программ Lasergene рассчитан на поиск в секвенированной последовательности сайтов различных рестрикционных эндонуклеаз, поиск открытых рамок считывания и их трансляцию в белковые продукты. Построение как линейных, так и кольцевых карт сопровождается возможностью добавления большого числа поясняющих символов различных элементов. Важной чертой этого модуля является создание рисунков, пригодных для публикации. Широкие возможности анализа белков заключены в модуле Protean. Так, данная программа позволяет выявлять в исследуемых белках участки с а-спиральной структурой и Р-складками и прочие элементы их структурной организации, рассчитывать гидрофильные и гидрофобные области белков, предсказывать физико-химические и электрофоретические свойства анализируемых с помощью компьютера белков.

Важным элементом анализа секвенированных последовательностей ДНК является их сравнение друг с другом или так называемое множественное выравнивание. Ранее большинство компьютерных программ позволяло одновременно сравнивать друг с другом только две последовательности, что снижало ценность получаемых результатов. В настоящее время уже многие программы анализа ДНК и белков рассчитаны на одновременный анализ

большого числа родственных последовательностей. Такая же возможность реализована в модуле Meg Align пакета программ Lasegene. Причем данный пакет позволяет проводить множественное выравнивание как последовательностей ДНК, так и белков. Повышению достоверности проведенного анализа способствует возможность ручного редактирования и окончательного доведения таких данных, поскольку случается так, что компьютер в некоторых случаях дает определенные систематические ошибки. Затем на основе уже выравненных последовательностей можно реконструировать филогенетические деревья, с вычисленным процентом сходства и генетическими расстояниями. Полученные результаты возможно сохранить в виде специальных файлов, что весьма удобно.

3.4 Нуклеиновый состав (изохоры, GC-острова) ДНК

Дезоксирибонуклеиновая кислота (ДНК) относится к нуклеиновым кислотам. **Нуклеиновые кислоты** – это класс нерегулярных биополимеров, мономерами которых являются нуклеотиды. **Нуклеотиды** состоят из **азотистого основания**, соединенного с пятиуглеродным углеводом (пентозой) – **дезоксирибозой** (в случае ДНК) или **рибозой** (в случае РНК), который соединяется с остатком фосфорной кислоты (H_2PO_3–). **Азотистые основания** бывают двух типов: пиримидиновые основания – урацил (только в РНК), цитозин и тимин, пуриновые основания – аденин и гуанин (рисунок 3.2).

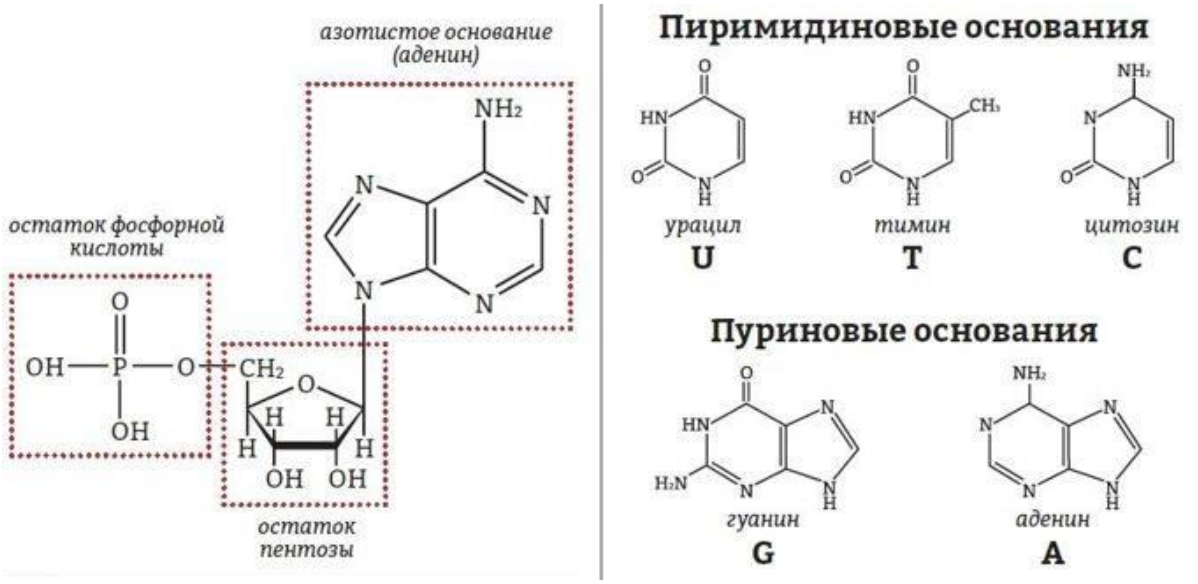


Рисунок 3.2 − Структура нуклеотидов (слева), расположение нуклеотида в ДНК (снизу) и типы азотистых оснований (справа): пиримидиновые и пуриновые

Атомы углерода в молекуле пентозы нумеруются числами от 1 до 5. Фосфат соединяется с третьим и пятым атомами углерода. Так нуклеинотиды

соединяются в цепь нуклеиновой кислоты. Таким образом, мы можем выделить 3’ и 5’-концы цепи ДНК (рисунок 3.3):

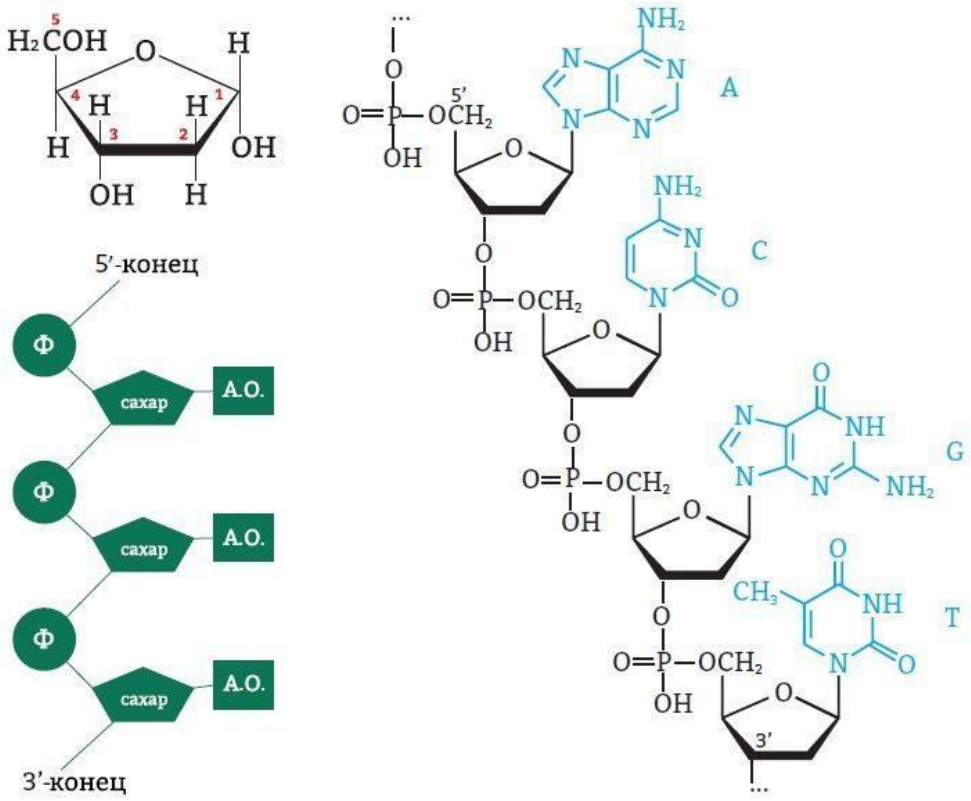


Рисунок 3.3 − Выделение 3’ и 5’-концов цепи ДНК

Две цепи ДНК образуют **двойную спираль**. Эти цепи в спирали сориентированы в противоположных направлениях. В разных цепях ДНК азотистые основания соединены между собой с помощью **водородных связей.** Аденин всегда соединяется с тимином, а цитозин – с гуанином. Это называется **правилом комплементарности**.

Правило комплементарности: A-T G-C

Например, если нам дана цепь ДНК, имеющая последовательность

3’– ATGTCCTAGCTGCTCG – 5’,

то вторая ей цепь будет комплементарна и направлена в противоположном направлении – от 5’-конца к 3’-концу:

5’– TACAGGATCGACGAGC– 3’.

GC-состав (**гуанин-цитозиновый состав**, **ГЦ-состав**) – доля гуанина (G) и цитозина (C) среди всех остатков нуклеотидов рассматриваемой нуклеотидной последовательности. GC-состав может быть определен как для фрагмента молекулы ДНК или РНК, так и для всей молекулы или даже всего генома.

Пара GC соединена тремя водородными связями, тогда как пара AT (аденин − тимин) −двумя. Поэтому ДНК с высоким содержанием GC более устойчива к денатурации в растворе, чем ДНК с низким. Кроме водородных

связей, на стабильность вторичной структуры ДНК и РНК влияют гидрофобные или стэкинг-взаимодействия между соединениями нуклеотидами, не зависящие от последовательности оснований нуклеиновых кислот.

При проведении ПЦР GC-состав праймера используется для предсказания температуры плавления этого праймера и температуры отжига. Высокий GC-состав праймера позволяет использовать его при высоких температурах отжига.

GC-состав обычно представляется в процентном отношении (**доля G+C** или **доля GC**) для одной из цепи ДНК или РНК. Процентный GC-состав вычисляется как: **GC=(G+C)/L*100,** где G+C − суммарное количество гуанинов и цитозинов, а L – длина цепи ДНК или РНК в нуклеотидах: A+T+G+C.

Композиционная гетерогенность

Неоднородность распределения по геному ГЦ- и АТ-обогащенных районов, другими словами, композиционная гетерогенность ДНК − одна из наиболее важных характеристик молекулярной организации геномов.

Более сорока лет назад Бернарди и сотрудники при исследовании генома мыши *Mus musculus* обнаружили, что комплекс ДНК и серебра может быть разделен с помощью равновесного центрифугирования в градиенте плотности Cs_2SO_4 (сульфат цезия) по частоте сайтов на молекуле ДНК, связавших серебро. Это открытие позволило с высокой точностью разделять ДНК на фракции. Дальнейшее изучение этих фракций ДНК привело к открытию четкой композиционной гетерогенности ДНК. Композиционно гомогенные сегменты ДНК, принадлежащие к небольшому числу семейств, различающихся по плавучей плотности, были названы **изохорами**. Фракционирование ДНК по плавучей плотности при центрифугировании фрагментов ДНК в градиенте Cs_2SO_4 (или сахарозы) было выявлено у большого числа видов животных. Это отражает гетерогенность ДНК по нуклеотидному составу: АТ-богатые последовательности обладают большей плавучей плотностью, чем ГЦ-богатые. Относительные количества ДНК в семействах изохор формируют так называемый композиционный изохорный паттерн генома (также он называется геномным фенотипом), т.е. характерный «рисунок» из изохор, являющийся специфичным для каждого отряда или семейства.

У человека были выявлены два «легких» семейства изохор: L1 (1,698 г/см3) и L2 (1,700 г/см3), и три «тяжелых»: H1 (1,704 г/см3), H2 (1,708 г/см3) и H3 (1,712 г/см3). У человека семейства L1 и L2 составляют свыше 62% всего генома; H1 – 22%, H2 – 9%, семейство H3 составляет около 3% генома.

Наибольшее значение с точки зрения структурно-функциональной организации генома имеет вопрос о связи различных семейств изохор с кодирующими последовательностями ДНК и особенностями их экспрессии.

В первых исследованиях этого вопроса было показано, что большинство из 40 взятых в анализ генов человека расположены в ГЦ-богатых семействах. Впоследствии локализация in silico (путем компьютерного анализа) более 14000 генов человека привела авторов к тому же самому выводу, позднее подтвержденному на еще больших выборках кодирующих последовательностей. На рисунке 3.4 представлена диаграмма, демонстрирующая распределение плотности генов в каждом из семейств изохор у человека (рисунок 3.4).

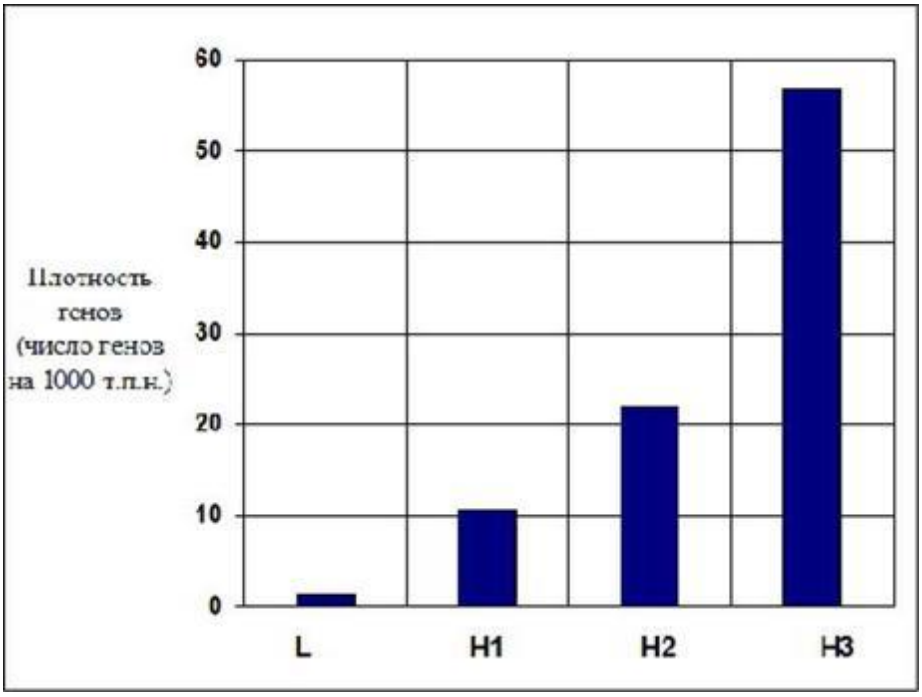


Рисунок 3.4 – Плотность генов в различных семействах изохор у человека

Выявление связи временных и межтканевых различий уровня экспрессии генов с ГЦ-уровнем, иными словами, распределения тканеспецифических генов и генов домашнего хозяйства относительно изохор, является не менее важным аспектом характеристики функционального значения композиционной гетерогенности геномной ДНК. Обобщая данные по структуре хроматина и распределению генов и изохор, Дж. Бернарди в 1993 г. писал: «Скорее всего, наибольший уровень транскрипции встречается в семействе изохор H3, поскольку там концентрация генов, прежде всего генов домашнего хозяйства, является наибольшей». Гипотеза о высокой транскрипционной активности генов, локализованных в семействе H3, подтверждалась и исследованиями нуклеотидного контекста стартовых АУГ-кодонов. Привлечение дополнительного критерия – процентного содержания гуанина или цитозина в третьем положении кодона (ГЦ3-уровень) – помимо молярного отношения ГЦ и АТ (ГЦ-уровень), и статистический анализ последовательностей ДНК из геномных баз данных

человека и шпорцевой лягушки (*Xenopus laevis*) позволили установить, что **гены домашнего хозяйства**:

1) преимущественно локализованы в ГЦ-богатых семействах изохор;

2) не составляют большинства генов в ГЦ-богатых семействах изохор;

3) являются не менее ГЦ3-обогащенными, чем тканеспецифические гены.

Повышенное содержание кодирующих последовательностей, более высокий уровень рекомбинации, большое число ГЦ-обогащенных коротких интерспесных повторов (SINE) в тяжелых изохорах, в то время как в легких локализовано значительно меньше генов, ниже уровень рекомбинации и находятся почти исключительно ГЦ-обедненные длинные интерсперсные повторы (LINE), дает основание говорить об **изохорах как структурно-функциональных единицах организации генома.** Границы между тяжелыми и легкими изохорами на примере детально исследованного в отношении композиционного состава на молекулярном уровне кластера генов гистосовместимости (MHC) человека являются более чем композиционными структурами – временные границы репликации в фазе S клеточного цикла практически точно соответствуют физическим границам локализации изохор.

3.5 Статистика ДНК как характеристика генома

В основе статистического анализа ДНК лежат методы математической оценки сходства и различия анализируемых последовательностей. При этом, как при проведении любого типа статистического анализа, для получения обоснованных выводов требуется некая совокупность оцениваемых параметров, то есть объемная выборка. В данном случае выборка представляет собой набор последовательностей одного и того же гена (или белка), принадлежащих разным организмам. Обычно в выборку включают последовательности не только одного гена, но и одного ранга, т.е. последовательности разных особей одного вида или последовательности разных видов.

Формирование такой выборки может быть выполнено разными способами. Новые последовательности исследователи получают в результате специфичной амплификации (в процессе полимеразной цепной реакции или ПЦР) целевого участка генома интересующего их объекта и последующей расшифровкой (секвенированием) продукта ПЦР. Дополнительно, для проведения широких сравнений, последовательности могут быть получены из одной из публичных баз данных. Чаще всего с этой целью используется база данных GenBank NCBI, которая содержит наибольшее число депонированных последовательностей.

Для работы с последовательностями ДНК используют специализированное программное обеспечение, позволяющее просматривать последовательности, сравнивать их, проводить разнообразные статистические расчеты. Количество таких программ значительно, и большинство из них предо-

ставляют исследователю сходный набор возможностей, различаясь только качествами интерфейса и количеством дополнительных опций. Чрезвычайно часто используемой является программа MEGA (Molecular Evolutionary Ge-netics Analysis), которая, благодаря ее авторам, доступна в интернете для свободной и бесплатной загрузки.

Статистический анализ последовательностей или, **«статистика ДНК»**, это сравнительно новый подход к исследованию геномов, возникший на стыке генетики, математики и информатики. Проведение статистического анализа последовательностей ДНК невозможно без использования компьютерной техники и специального программного обеспечения, созданного специально для работы с большими объемами генетических текстов.

Анализ нуклеотидной последовательности ДНК с использованием методов математической статистики предоставляет исследователю колоссальный объем информации, касающейся частот встречаемости конкретных нуклеотидов и кодонов, скорости и особенностей нуклеотидных замещений в гене, сходства и различия «рисунка» нуклеотидных заменив последовательностях ДНК разных организмов. Однако, как и при использовании любого аналитического метода, статистический анализ ДНК требует тщательного выбора применяемых подходов, а также внимательной трактовки полученных результатов.

Основной задачей статистического анализа последовательностей ДНК традиционно считается оценка уровня сходства двух сравниваемых последовательностей с тем, чтобы сделать заключение о степени их эволюционной близости, т.е. родства. Сравнению могут подвергаться как идентичные гены разных организмов, так и гены, сходные по кодируемому продукту, внутри одного или нескольких геномов. В большинстве случаев наблюдаемое сходство последовательностей какого-либо участка генома или геномов говорит об общности их происхождения (гомологии). Благодаря сравнениям такого рода удается выявить эволюционное родство организмов даже при наличии значительных различий в морфологии и образе жизни, а сделать предположения о происхождении генов и их семейств, в том числе и при изменении функции продукта этих генов.

3.6 Регуляторные участки (промоторы, терминаторы)

Транскрипция – это синтез РНК на матрице ДНК. У прокариот синтез всех трех видов РНК катализируется одним сложным белковым комплексом – РНК-полимеразой. Синтез мРНК начинается с обнаружения РНК-полимеразой особого участка в молекуле ДНК, который указывает место начала транскрипции − **промотора.** После присоединения к промотору РНК-полимераза раскручивает прилежащий виток спирали ДНК. Две цепи ДНК в этом месте расходятся, и на одной из них фермент осуществляет синтез мРНК. Сборка рибонуклеотидов в цепь происходит с соблюдением

их комплементарности нуклеотидам ДНК, а также антипараллельно по отношению к матричной цепи ДНК. РНК-полимераза способна собирать полинуклеотид лишь от 5'-конца к 3'-концу, матрицей для транскрипции может служить только одна из двух цепей ДНК, а именно та, которая обращена к ферменту своим 3'-концом (3' → 5'). Такую цепь называют кодогенной.

Терминатор − это участок, где прекращается дальнейший рост цепи РНК и происходит ее освобождение от матрицы ДНК. РНК-полимераза также отделяется от ДНК, которая восстанавливает свою двухцепочечную структуру.

Фрагмент молекулы ДНК, включающий промотор, транскрибируемую последовательность и терминатор, образует единицу транскрипции − транскриптон.

Оперонная регуляция (т.е. регуляция на уровне транскрипции) – основной механизм регуляции активности генов у прокариот и бактериофагов.

Оперон – участок генетического материала, транскрипция которого осуществляется на одну молекулу иРНК под контролем белка-репрессора (рисунок 3.5).

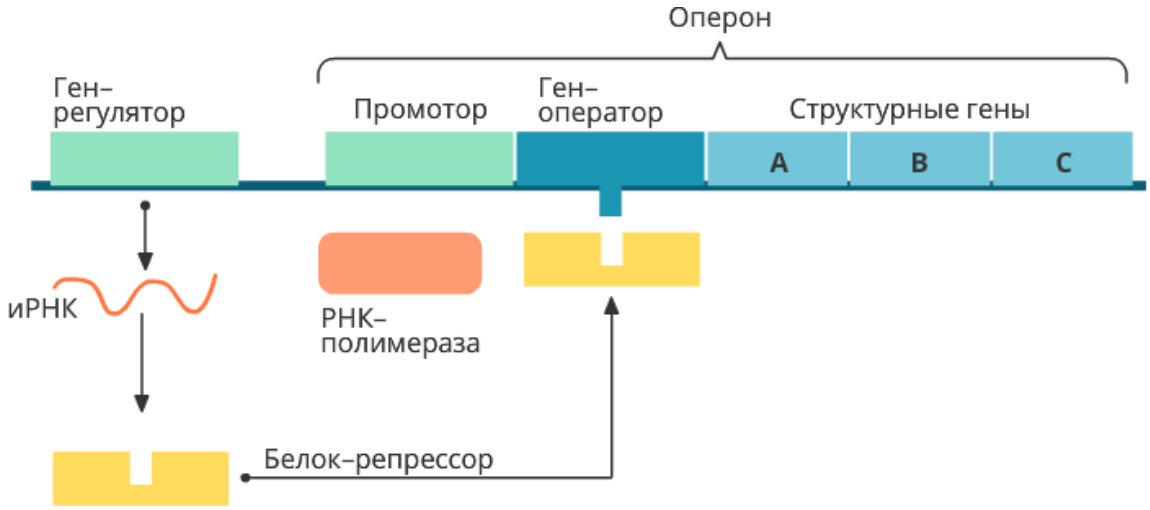


Рисунок 3.5 – Структура оперона

В соответствии с биохимическими критериями **промотор** представляет собой последовательность нуклеотидов, обеспечивающую базальный (но не максимальный) уровень транскрипции соответствующего транскриптона. Он является той минимальной последовательностью, которая специфически распознается холоферментом РНК-полимеразы среди случайных последовательностей нуклеотидов.

Бактериальный промотор содержит две канонические последовательности: в области -35 и в области -10.

Промотор эукариотических генов, узнающийся РНК-полимеразой промотор содержит два базовых регуляторных элемента: ТАТА-

последовательность (положение -25) и специфическую нуклеотидную последовательность, обогащенную пиримидинами в положении -75.

Основной элемент промотора – место связывания РНК-полимеразы, которое она занимает перед началом синтеза РНК. В состав промоторов могут входить также участки связывания белков-регуляторов.

У эукариот регуляторные элементы собраны в регуляторные регионы. Основной регуляторный элемент эукариот – это **коровый (базальный) промотор.** Он обеспечивает сборку базального транскрипционного комплекса (из основных факторов транскрипции и РНК-полимеразы) и инициацию транскрипции на базальном (исходном, базовом) уровне. Часто этот уровень так низок, что приводит к синтезу лишь единичных молекул РНК и, в дальнейшем, белков.

3.7 Подбор праймеров для ПЦР

Праймеры являются чрезвычайно важной составляющей любого процесса амплификации, включая ПЦР. Для того чтобы химически синтезированные олигонуклеотидные праймеры (или иным способом приготовленные) могли служить затравочными молекулами для ферментативного синтеза комплементарных цепей ДНК, необходим их отжиг на подходящей одноцепочечной нуклеиновой кислоте с образованием одно/двухцепочечного стартового комплекса с наличием способного к удлинению по одноцепочечной ДНК/РНК-матрице спаренного 3’-конца такого праймера, несущего на нем гидроксильную группу. Отжиг праймеров на матрице может происходить в широком диапазоне температур, который зависит от длины конкретного праймера, а также от его химического состава и в этой связи можно напомнить, что A-T-пары образуют между собой две водородные связи, а G-C-пары – три водородные связи, что вносит существенный вклад в температурные оптимумы при молекулярной гибридизации нуклеиновых кислот с разными GC- и AT-составами.

Как имеется множество вариаций ПЦР, решающих часто весьма разные задачи, так существуют и довольно сильно отличающиеся принципы подбора олигонуклеотидных праймеров. Например, в целях диагностики необходимо выявлять наличие подходящего фрагмента ДНК или РНК (по сути – любого, но специфичного именно для выявляемого организма) и поэтому в таких случаях праймеры подбираются, исходя из задач обнаружения выбранного участка с максимальной достоверностью и экономией (времени, денежных и прочих ресурсов), тем более, если такие анализы должны носить массовый характер. При проведении научных исследований часто возникают задачи по клонированию конкретного гена, у которого есть начало и конец и в этом случае праймеры для его амплификации подбираются не с целью достичь максимальной эффективности процесса, а так чтобы было удобно вести дальнейшие эксперименты, но при этом праймеры могут оказаться далеко не самыми оптимальными по целому ряду параметров. Наконец, с помощью ПЦР изучают полиморфизм ДНК разных организмов – ви-

дов, сортов/пород, отдельных особей или индивидов. В таких случаях праймеры подбираются с таким расчетом, чтобы вариабельный участок оказывался между ними, что также накладывает определенные ограничения. Во многих подходах для генотипирования организмов вообще используются праймеры с произвольными последовательностями, не заботясь о присутствии для них мест отжига в геномах, поскольку они выбираются весьма короткими (часто от 10 до 12 нуклеотидов) и отжигаться на множестве мест теоретически должны. Подробные обзоры таких способов амплификации неопределенных фрагментов ДНК были сделаны нами не так давно и поэтому здесь на праймерах для таких вариантов ПЦР останавливаться не будем.

Относительно взаимного расположения мест отжига праймеров, определяющих размер(ы) ампликонов, которые на самом деле диктуются многими привходящими обстоятельствами, то этот вопрос имеет больше отношения к эффективности процесса амплификации. В данной статье будут изложены главные принципы подбора праймеров, различные основные требования, предъявляемые к ним с учетом решаемых в каждом конкретном случае своих задач. Значительное внимание будет уделено различным модификациям праймеров, включая вырожденные праймеры, тогда как специальных праймеров для различных вариаций ПЦР коснемся довольно кратко. Есть также масса и других задач, решаемых с помощью ПЦР, под которые подбираются специализированные праймеры. Однако ни в одной даже очень большой статье невозможно рассмотреть все многообразие праймерных систем, применяемых в ПЦР. Что касается праймеров, используемых в различных способах изотермической амплификации нуклеиновых кислот, имеющих свои особенности, то они также останутся за пределами рассмотрения.

Требования, предъявляемые к праймерам:

1. Размер праймера должен быть 16–25 нуклеотидов. Меньше 16-ти – слабая связь с целью.

2. Разница в температуре плавления праймеров – не более 6 градусов.

3. Ц+Г должно быть 50-60%.

4. Для улучшения качества отжига рекомендуется подбирать праймеры так, чтобы последние несколько нуклеотидов 3’-конца праймера содержали GC-основания.

5. Проверить сбалансированность PCR по нуклеотидам и праймерам.

6. Отсутствие внутренней вторичной структуры (праймеры не должны быть само- и взаимнокомплиментарными).

7. Отсутствие комплиментарности между 3’-концами (чтобы не образовывалось праймер-димеров).

8. Оптимальная концентрация праймеров подбирается эмперически, но не должна быть более 50 пикомолей по пробирку – иначе начнется неспецифический отжиг праймеров.

9. Упрощенный расчет оптимальной температуры отжига праймера:

$Tm = [(F+N)*2^0C] + [(G+C)*4^0C]$ (если суммарная длина олигонуклеотида не превышает 20 оснований);

Tm = 22+1.46([2*(G+C)] + (F+A)) (если суммарная длина олигонуклеотида составляет 20-30 оснований).

10. Область отжига праймеров должна находиться вне зон мутаций, делеций или инсерций в пределах видов или иной, взятой в качестве критерия при выборке праймеров, специфичности. При попадании на такую зону, отжига праймеров происходить не будет, и как следствие – ложноотрицательный результат.

3.8 Анализ частоты использования кодонов

Определение нуклеотидного состава нуклеиновых кислот и аминокислотного состава белков преимущественно проводится с помощью специальных компьютерных программ (например, MEGA). При анализе нуклеотидного состава ДНК и РНК наиболее часто анализируются следующие показатели:

1. **Частота транзиций**

Транзиция − это мутация, обусловленная заменой одного пуринового основания на другое (А ↔ Г) или одного пиримидинового на другое (У, Т↔Ц).

Транзиции − простые замены (не происходит изменения ориентации пурин-пиримидин в мутантном сайте двухцепочечной молекулы ДНК).

Частота транзиций (*P*) вычисляется по формуле: $P=n_p/L$, где n_p – число наблюдаемых транзиций; L – общее число нуклеотидных сайтов, по которым сравниваются последовательности.

2. **Частота трансверсий**

Трансверсия − это мутация, обусловленная заменой пуринового основания на пиримидиновое и наоборот (А, Г ↔У, Т, Ц).

Трансверсии − сложные или перекрестные замены (происходит изменение ориентации пурин–пиримидин в мутантном сайте двухцепочечной молекулы ДНК). Частота трансверсий (*Q*) определяется по формуле: $Q=n_q/L$

где n_Q − число наблюдаемых трансверсий; L − общее число нуклеотидных сайтов, по которым сравниваются последовательности.

3. **Соотношение наблюдаемых трансверсий и транзиций (*q*)**, определяемое по формуле: $q=n_q/n_{p/}$

4. **Общая ГЦ-насыщенность** − это общее содержание гуанина и цитозина. Повышение ГЦ-насыщенности свидетельствует об увеличении термодинамической стабильности ДНК за счет образования большего количества водородных связей (трех между гуанином и цитозином вместо двух между аденином и тимином).

5. **ГЦ-насыщенность отдельных положений кодона** − ГЦ1, ГЦ2 и ГЦ3-насыщенность. Наибольшее значение имеет определение ГЦ3-насыщенности как маркера мутационного давления.

6. **Зависимость ГЦ-насыщенности отдельных положений кодона от общей ГЦ-насыщенности**. Эта зависимость, изученная С. Кумаром и М. Неем, наиболее часто имеет вид, показанный на (рисунок 3.6).

Данный график показывает:

1. С увеличением общей ГЦ-насыщенности наблюдается различной степени линейный рост содержания гуанина и цитозина во всех положениях нуклеотида в кодоне.

2. Наибольший рост характерен для значений ГЦ3 (наклон тренда −1,61), что свидетельствует о выраженном влиянии на замены в данном положении кодона мутационного давления и слабом влиянии отрицательного отбора. Это объясняется тем, что 72% замен по третьему положению не приводят к изменению кодируемой аминокислоты.

3. Меньший наклон тренда по значениям ГЦ1, равный 0,99, связан с меньшей долей синонимичных замен (5%).

4. Несмотря на то, что все замены по второму положению несинонимичны, наклон тренда по значениям ГЦ2 является положительным (0,39). Объяснить этот факт можно тем, что часть замен аминокислот является нейтральной, т.е. не приводит к изменению функции белка. В этом случае отрицательный отбор не оказывает никакого влияния на данную мутацию, и она закрепляется, отражая основное направление нуклеотидных замен.

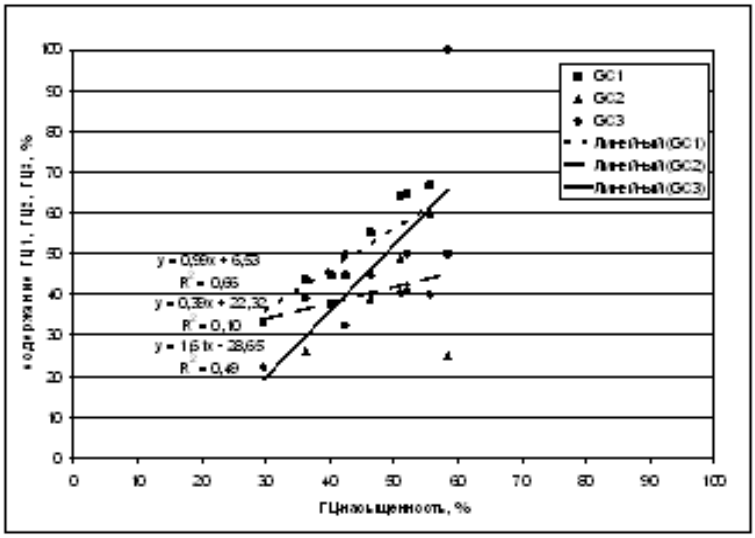


Рисунок 3.6 – Классическая зависимость содержания гуанина и цитозина в отдельных положениях кодона от общей ГЦ-насыщенности на примере экзонов гена, кодирующего алкогольдегидрогеназу класса 3 человека

Таким образом, можно утверждать, что в экзонах изучаемого гена АДГ «разрешены» преимущественно замены в третьем положении кодона, в меньшей степени − в первом, а в наименьшей степени – во втором, что связано с вероятностью синонимичных и несинонимичных замен. Выраженное предпочтение синонимичных замен свидетельствует о структурно-функциональных ограничениях, налагаемых на молекулу, а получение такого графика может выступать в качестве одного из критериев ее эволюционной сформированности.

При анализе аминокислотного состава белков чаще анализируются:

1. В связи с ГЦ-насыщенностью − *содержание* аминокислот *GARP* (кодируемых ГЦ-богатыми кодонами) и *FYMINK* (кодируемых ГЦ-бедными кодонами, таблица 3.1).
2. В связи с зарядом аминокислот – *содержание положительнозаряженных, отрицательнозаряженных и нейтральных аминокислот*.
3. В связи с химической структурой − *содержание ароматических, серосодержащих, гидроксильных аминокислот и др*.

Таблица 3.1 – Аминокислоты группы GARP и FYMINK и соответствующие им кодоны мРНК

Аминокислоты группы GARP			Аминокислоты группы FYMINK		
G	глицин	ГГУ			
		ГГЦ			
		ГГА			
		ГГГ	F	фенилаланин	УУУ
A	аланин	ГЦУ			УУЦ
		ГЦЦ	Y	тирозин	УАУ
		ГЦА			УАЦ
		ГЦГ	M	метионин	АУГ
R	аргинин	ЦГУ	I	изолейцин	АУУ
		ЦГЦ			АУЦ
		ЦГА			АУУ
		ЦГГ	N	аспарагин	ААУ
		АГА			ААЦ
		АГГ	K	лизин	ААА
P	пролин	ЦЦУ			ААГ
		ЦЦЦ			
		ЦЦА			
		ЦЦГ			

Сайт − это участок небольшого размера молекулы нуклеиновой кислоты или белка, обычно равный одному нуклеотиду или одному аминокислотному остатку соответственно.

Константные сайты − это нуклеотидные или аминокислотные сайты выравненных последовательностей, в которых не наблюдаются замены.

Вариабельные сайты − это нуклеотидные или аминокислотные сайты выравненных последовательностей, в которых наблюдаются замены.

Синглетонные сайты − это вариабельные сайты не менее пяти выравненных последовательностей, в которых преобладает частота одного нуклеотида или аминокислотного остатка. Пример: сайты № 192, 207 на (рисунок 3.6).

Парсимоничные сайты − это вариабельные сайты не менее пяти выравненных последовательностей, в которых как минимум два нуклеотида или аминокислотных остатка встречаются хотя бы два раза. Пример: сайты №191, 192 на рисунке 3.7.

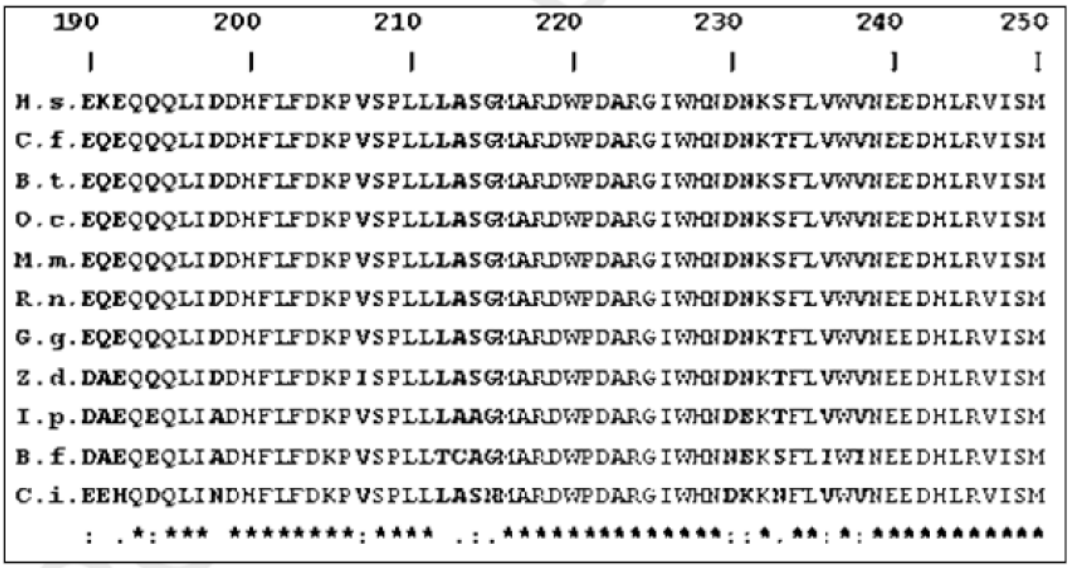


Рисунок 3.7 – Выравненные участки каталитических доменов
М-изоферментов креатинкиназ позвоночных
и соответствующие участки креатинкиназ ланцетника и оболочника

Существуют две системы классификации сайтов нуклеиновых кислот в связи с их положением в кодоне − система вырожденности и синонимичности/несинонимичности (таблица 3.2).

Таблица 3.2 – Соотношение систем вырожденности и синонимичности/несинонимичности сайтов

Система классификации сайтов	
По вырожденности	**По синонимичночности/ несинонимичности**
невырожденный	абсолютно несинонимичный
двухкратновырожденный	несинонимичный на 2/3 синонимичный на 1/3
трехкратновырожденный	несинонимичный на 1/3 синонимичный на 2/3
четырехкратновырожденный	абсолютно синонимичный

Нолькратно вырожденный сайт (невырожденный сайт) – это нуклеотидный сайт, в котором любая возможная замена является несинонимичной. Пример: в первом положении кодона ЦУУ (Лей) возможны три несинонимичные замены:

1. ЦУУ→УУУ (Фен).
2. ЦУУ→АУУ (Иле).
3. ЦУУ→ГУУ (Вал).

Вывод: первый сайт кодона ЦУУ является невырожденным.

Двухкратно вырожденный сайт − это нуклеотидный сайт, в котором одна из трех возможных замен является синонимичной. Пример: в третьем положении кодона УУУ (Фен) возможны две несинонимичные и одна синонимичная замена:

1. УУУ →УУЦ (Фен).
2. УУУ →УУА (Лей).
3. УУУ →УУГ (Лей).

Вывод: третий сайт кодона УУУ является двухкратно вырожденным.

Трехкратно вырожденный сайт − это нуклеотидный сайт, в котором две замены из трех возможных синонимичны. Пример: в третьем положении кодона АУУ (Иле) возможны одна несинонимичная и две синонимичные замены:

1. АУУ→АУЦ (Иле).
2. АУУ→АУА (Иле).
3. АУУ→АУГ (Мет).

Вывод: третий сайт кодона АУУ является трехкратно вырожденным.

Четырехкратно вырожденный сайт − это нуклеотидный сайт, в котором возможны лишь синонимичные замены. Пример: в третьем положении кодона ЦУУ (Лей) возможны три синонимичные замены:

1. ЦУУ→ЦУЦ (Лей).
2. ЦУУ→ЦУА (Лей).
3. ЦУУ→ЦУГ (Лей).

Вывод: третий сайт кодона ЦУУ является четырехкратно вырожденным.

Несинонимичный сайт − это сайт, в котором возможна несинонимичная замена.

Синонимичный сайт − это сайт, в котором возможна синонимичная замена.

Если все замены в данном сайте несинонимичны, сайт называется абсолютно несинонимичным. Если все замены в данном сайте синонимичны, сайт называется абсолютно синонимичным. Если одна из замен в данном сайте несинонимична, сайт называется несинонимичным на ⅓ (соответственно, синонимичным на ⅔). Если две замены в данном сайте несинонимичны, сайт называется несинонимичным на ⅔ (соответственно, синонимичным на ⅓).

Кодон (триплет) − это наименьшая функциональная единица гена, состоящая из трех рядом расположенных нуклеотидов, кодирующая одну аминокислоту.

Серия кодонов − это группа синонимичных кодонов.

Синонимичные (эквивалентные, изоакцепторные) кодоны – это кодоны, кодирующих одну и ту же аминокислоту.

Двухкратно вырожденная серия кодонов − это серия кодонов, состоящая из двух синонимичных триплетов.

Четырехкратно вырожденная серия кодонов − это серия кодонов, состоящая из четырех триплетов.

Шестикратно вырожденная серия кодонов − это серия кодонов, состоящая из шести триплетов.

Претерминальные кодоны − это кодоны, которые могут стать терминальными в результате замены одного нуклеотида и, следовательно, прервать синтез пептидной цепочки.

ГЦ3-кодоны − это кодоны, содержащие в третьем положении гуанин или цитозин, исключая невырожденные кодоны (АУГ− метионин, УГГ− триптофан) и стоп-кодоны.

Стратегия кодирования белка определяется картиной использования кодонов в соответствующих ему мРНК и ДНК. Многообразие возможных стратегий кодирования связано с вырожденностью генетического кода (в среднем на каждую из 20 аминокислот приходится три синонимичных кодона). Одним из наиболее часто анализируемых при изучении стратегии кодирования показателей является картина использования синонимичных кодонов.

Методики изучения стратегии кодирования белков:

1. Сравнения количества (доли) кодонов в изучаемых последовательностях нуклеиновых кислот.

2. Сравнения показателей относительного использования синонимичных кодонов (RSCU − relative synonymous codon’s usage), которые производятся по формуле: $\mathbf{RSCU_К = (RSCU_{СК} * n_К)/n_{СК}}$

где $RSCU_К$ − показатель относительного использования кодона А,

$RSCU_{СК}$ − количество кодонов в серии,

$n_К$ − частота использования кодона А,

$n_{СК}$ − частота использования всех кодонов серии.

Рассчитаем в качестве примера RSCU для кодона УГУ в мРНК, кодирующей НАДН-дегидрогеназу 6 человека (таблица 3.3). $RSCU_{СК}$ = 2 (серия представлена двумя кодонами − УГУ и УГЦ), $n_К$ = 1 (кодон УГУ используется в данной мРНК 1 раз), $n_{СК}$ = 1 (кодон УГУ используется 1 раз, кодон УГЦ не используется), тогда: $RSCU_{УГУ}$= (2*1)/2=2.

3. Сравнения суммарных показателей RSCU серии кодонов, соответствующих одной и той же аминокислоте (рисунок 3.8).

4. Суммарные показатели RSCU соответствующих аргинину ГЦЗ-кодонов мРНК, кодирующих алкогольдегидрогеназы классов 1В и С человека, значительно отличаются (4,2 и 1,2 соответственно), а соответствующих цистеину − сходны (0,8 и 0,82).

Таблица 3.3 – Показатели RSCU и количество кодонов УГУ и УГЦ в мРНК, кодирующих НАДН-дегидрогеназу 6 человека, трихинеллы и цианорабдитис

Организм/показатель	Кодон УГУ	Кодон УГЦ
Человек	n=1 RSCU=2,0	n=0 RSCU=0
Трихинелла	n=0 RSCU=0	n=2 RSCU=2,0
Цианорабдитис	n=3 RSCU=2,0	n=0 RSCU=0

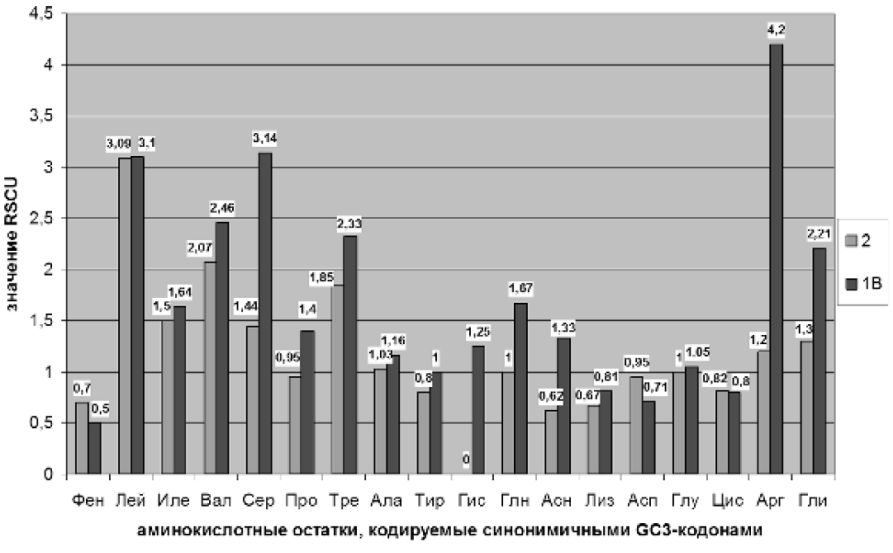


Рисунок 3.8 – Показатель относительного использования синонимичных ГЦ3-кодонов для алкогольдегидрогеназ классов 1В и 2 человека

4. Вычисление дистанции Дж. МакАйнерни (*Djk*):

$$D_{jk} = \sum_{i-1}^{n} \frac{abs(RSCU_{ji} - RSCU_{ki})}{n},$$

Контрольные вопросы

1. Дайте определение понятиям «ген», «экспрессия генов», «активация РНК».
2. Как осуществляется регуляция экспрессии генов у прокариот и эукариот?
3. Какие способы регуляции экспрессии генов существуют у эукариот?
4. Что такое «генная экспрессия» и в чем ее сущность?
5. Что такое оперон и какие структурные части оперона выделяют?
6. В чем заключается анализ последовательностей ДНК?
7. Перечислите методики секвенирования ДНК.
8. В чем медицинское предназначение анализа ДНК?
9. Какие компьютерные программы используют для анализа секвенированных последовательностей генов?
10. Опишите изохоры и GC-острова.
11. Дайте определение «промоторы» и «терминаторы».
12. Как осуществляется подбор праймеров для ПЦР?
13. Какие показатели используют при анализе нуклеотидного состава?
14. Что анализируют при анализе аминокислотного состава?

Раздел IV
МЕТОДЫ АНАЛИЗА ПОСЛЕДОВАТЕЛЬНОСТЕЙ БЕЛКОВ

4.1 Фолдинг и транспорт белков у про- и эукариот

4.1.1 Фолдинг, белки шапероны

Кроме последовательности аминокислот полипептида (первичной структуры), крайне важна трехмерная структура белка, которая формируется в процессе фолдинга (от англ. folding - «сворачивание»).

Фолдинг – это процесс формирования «правильной» третичной структуры из полипептидной цепочки.

Универсальным механизмом, обеспечивающим быстрый и безошибочный фолдинг, является участие в сворачивании белковых цепей **шаперонов** – специализированных белков, обнаруженных во всех органеллах всех организмов от бактерий до приматов. Шапероны бывают двух типов.

1. **Молекулярные шапероны**, которые связываются с белковой нитью, предотвращая ее агрегацию или деградацию.

2. **Шаперонины**, которые обеспечивают фолдинг белков.

Молекулярные шапероны состоят из белков **Hsp70** и их гомологов:

– **Hsp70** в цитозоле и матриксе митохондрий,

– **BiP** в эндоплазматическом ретикулуме,

– **DnaK** в бактериях.

Белки Hsp70 называют белками теплового шока (heat shock proteins), поскольку они активно синтезируются клеткой при нагревании. Связанный с АТФ Hsp70 (Hsc70 это постоянно синтезируемый гомолог Hsp70) присоединяется к гидрофильному участку белковой цепи. Гидролиз АТФ→АДФ разрешает фолдинг белковой цепи. Замена АДФ на АТФ приводит к диссоциации Hsp70 со свернутой белковой цепи (рисунок 4.1 вверху).

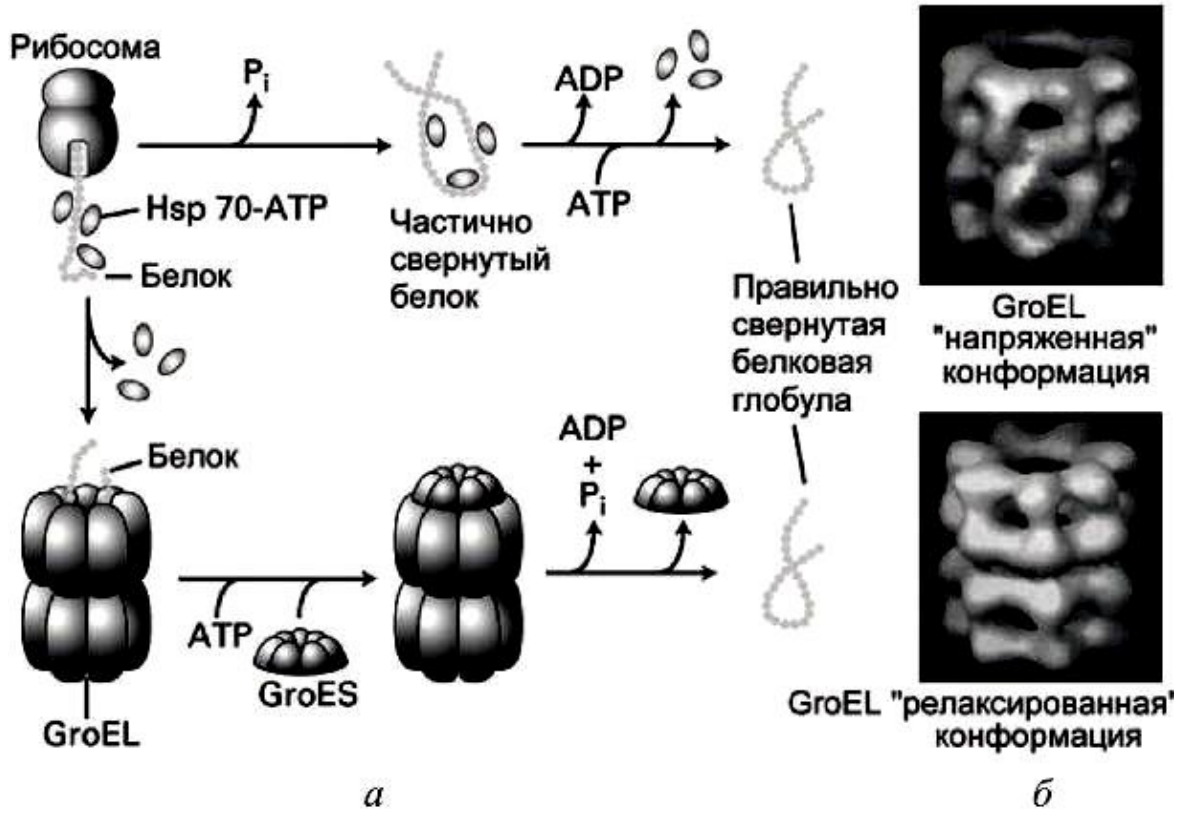


Рисунок 4.1 – Фолдинг белков с участием шаперонов и шаперонинов:
а – схема фолдинга; б – конформации шаперонина

Шаперонины – большие цилиндрические макромолекулярные комплексы, состоящие из двух олигомерных колец (по восемь мономеров в шаперонине эукариот TriC и по семь мономеров в шаперонинах бактерий, митохондрий и хлоропластов GroEL) – обеспечивают изоляцию белковой нити на время фолдинга (рисунок 4.1 внизу). Связывание GroEL с АТФ высвобождает свернутый белок. Функционирование GroEL осуществляет-

ся с участием ко-шаперонина GroES, который закрывает (кэпирует) полость шаперона на время фолдинга.

После правильного сворачивания, крышка открывается и готовая функциональная молекула белка покидает шаперонин.

Шапероны и шаперонины могут исправлять поврежденную структуру белка. Если исправить не удается, белок должен быть разрушен в протеасоме – это тоже сложный белок, образующий полость.

Результатом нарушения фолдинга белков являются конформационные болезни. К ним относятся:

- Прионовые заболевания;
- Болезнь Альцгеймера;
- Синдром Марфана;
- Куриная слепота;
- Злокачественные опухоли (нарушение фолдинга p53).

Прионы – это инфекционные белки. Вызывают тяжелые заболевания центральной нервной системы у человека и ряда высших животных («медленные инфекции»):

✓ у людей – куру («смеющаяся смерть»), болезнь Крюцельда-Якоба, семейная фатальная бессонница;

✓ у коров, норок, оленей, козы – бешенство (губчатая энцефалопатия);

✓ у овец – почесуха (scrapy).

4.1.2 Секреция белков у прокариот: Sec-аппарат и сигнальный пептид

Большинство бактерий способны транспортировать синтезируемые белки в окружающую среду. Для этого бактериальные клетки используют различные системы секреции в зависимости от строения и конечной локализации белка. Все эти системы должны специфически распознавать свои субстраты и облегчать секрецию без нарушения целостности клеточной оболочки. Однако для достижения такой цели эти системы используют существенно разные механизмы и отличаются друг от друга по своей сложности.

Грамотрицательные бактерии имеют две мембраны – цитоплазматическую мембрану (IM) и внешнюю мембрану (OM), разделенные периплазматическим пространством. Такая организация клеточной стенки делает процесс секреции топологически сложным. В грамотрицательных бактериях биологические молекулы, выделяемые во внешнюю среду, должны пересечь два гидрофобных барьера. Для обеспечения транспорта через клеточную стенку в указанных микроорганизмах функционируют как минимум шесть специализированных систем секреции (т.н. системы секреции типов 1–6).

Системы секреции можно классифицировать по тому, за один или два этапа происходит транспортировка белковой молекулы. Один класс систем секреции, включающий типы 1, 3 и 4, обеспечивает одноэтапную секрецию

из цитоплазмы сразу во внеклеточную среду. Второй класс систем секреции (типов 2, 5 и 6) предполагает экспорт в два этапа (из цитоплазмы в периплазматическое пространство, и из периплазмы в среду), причем в последовательных этапах транслокации работают разные молекулярные механизмы. Первый этап – транспорт в периплазму – происходит при участии одного из трех разных путей: Sec, Tat или SRP. Основные особенности этих путей – Sec (SecB-зависимый путь), Tat (твин-аргинин зависимый путь, twin-arginine translocation) и SRP (signal recognition particle) – схематически показаны на рисунке 4.2.

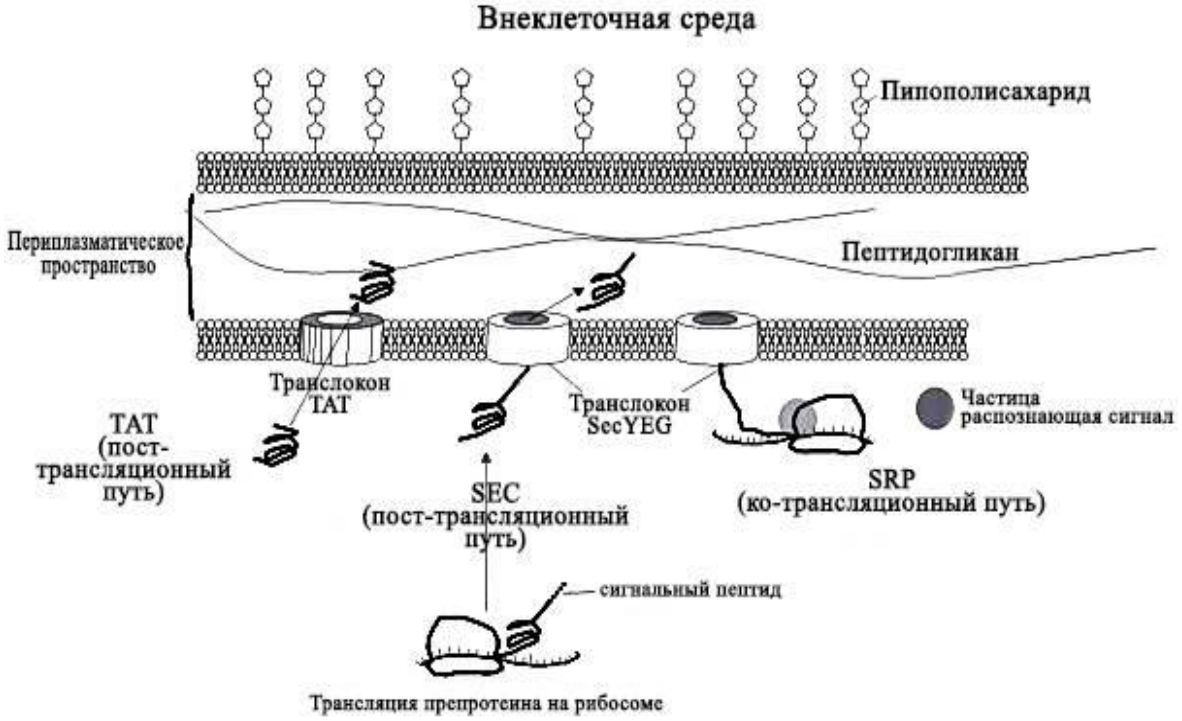


Рисунок 4.2 – Пути транслокации через плазматическую мембрану

Все три перечисленных пути используют разные механизмы и приспособлены для транслокации разных белков-мишеней. Транслокация по пути Sec или Tat является пост-трансляционным процессом, в то время как путь SRP является ко-трансляционным, т.е. осуществляет экспорт полипептидной цепи, которая продолжает синтезироваться на рибосоме. Белки-мишени пути Sec для транслокации должны быть развернуты полностью до состояния вытянутой полипептидной цепи (т.е. теряют третичную структуру в ходе транслокации), эти белки подвергаются рефолдингу в периплазме. Белки-мишени пути Tat транслоцируются в свернутом состоянии, не теряя третичной структуры, которую они приобретают в цитоплазме. С использованием пути Tat возможно экспортировать функционально активные белки, имеющие сложную третичную структуру.

Sec-аппарат. Наибольшая доля секретируемых и мембранных белков у прокариот используется для транслокации через путь Sec, в связи с чем исторически путь Sec получил название основного секреторного пути (general secretory pathway, GSP). Природные белки-мишени пути Sec, как правило, локализуются в периплазматическом пространстве или интегрируются во внешней мембране. Секреторный **аппарат Sec** гомологичен транслокону в эндоплазматическом ретикулуме высших эукариот и транслокону Sec 61 у дрожжей. Центральную роль в транслоконе Sec играет комплекс из трех белков SecYEG, который образует канал в цитоплазмотической мембране. В состав транслоказного комплекса также входят другие белки: мембранная АТРаза SecA, цитоплазматический шаперон SecB и акцессорные белки SecD и SecF, YidC и YajC. Белки-мишени пути Sec синтезируются в цитоплазме в виде предшественников, содержащих на N-конце короткий (15-30 аминокислотных остатков) пептид, называемый **сигнальным пептидом**. SecA распознает сигнальный пептид и в кооперации с шапероном SecB, транспортирует препротеин к транслокону SecYEG (рисунок 4.3).

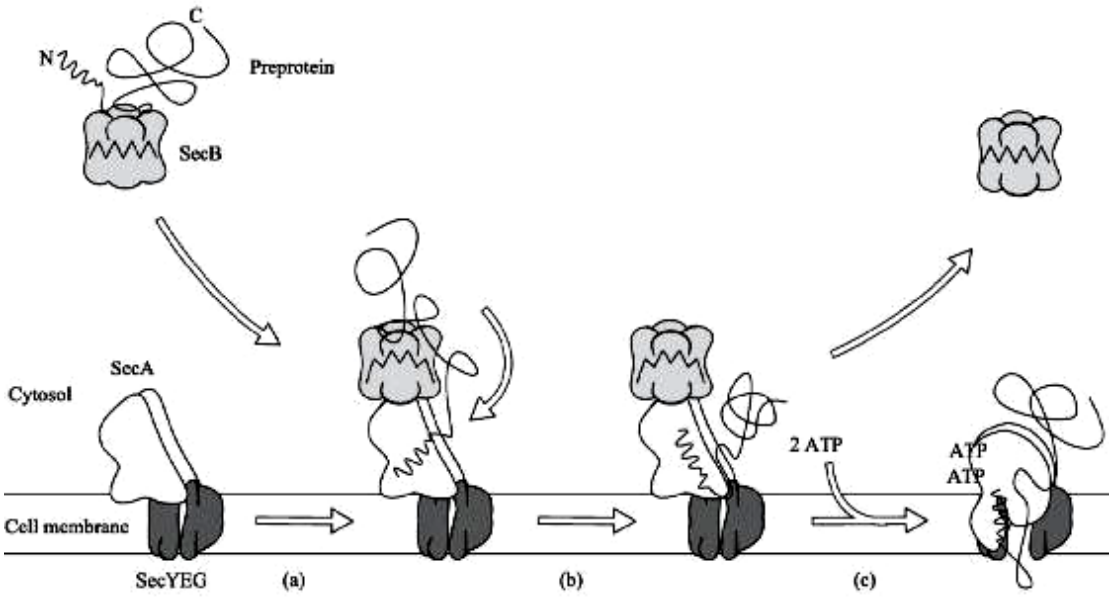


Рисунок 4.3 – Sec-аппарат

В ходе транслокации белка-мишени, как только С-конец сигнального пептида достигает поверхности цитоплазмотической мембраны, сигнальный пептид отщепляется при участии сигнальных пептидаз (пептидазы I (LepB) или пептидазы II (LSPA).

Сигнальные пептиды имеют сильно различающиеся аминокислотные последовательности, но демонстрируют и общие структурные характеристики. Типичный сигнальный пептид имеет на N-конце короткий участок, т.н. N-домен, длиной от 2 до 10 а.о., богатый положительно заряженными аминокислотами; в центре сигнального пептида находится участок (длиной

10–20 а.о.), богатый аминокислотными остатками с гидрофобными боковыми цепями (Н-домен); на С-конце сигнального пептида находится С-домен, менее гидрофобный, чем Н-домен, и содержащий сайт узнавания сигнальной пептидазы. Во время транспорта белков из цитоплазмы сигнальный пептид отщепляется сигнальной пептидазой с высвобождением зрелого белка.

Вторичная структура экспортируемого белка играет важную роль в том, насколько эффективно будет отщепляться сигнальный пептид. Получение эффективной транслокации требует подбора сигнального пептида к выбранному белку-мишени.

4.1.3 Распределение белков по компартментам клетки эукариот

Мембранные структуры клетки активно участвуют во внутриклеточном транспорте белков. Существует три основных мехнизма, с помощью которых клетка решает эту задачу (рисунок 4.4).

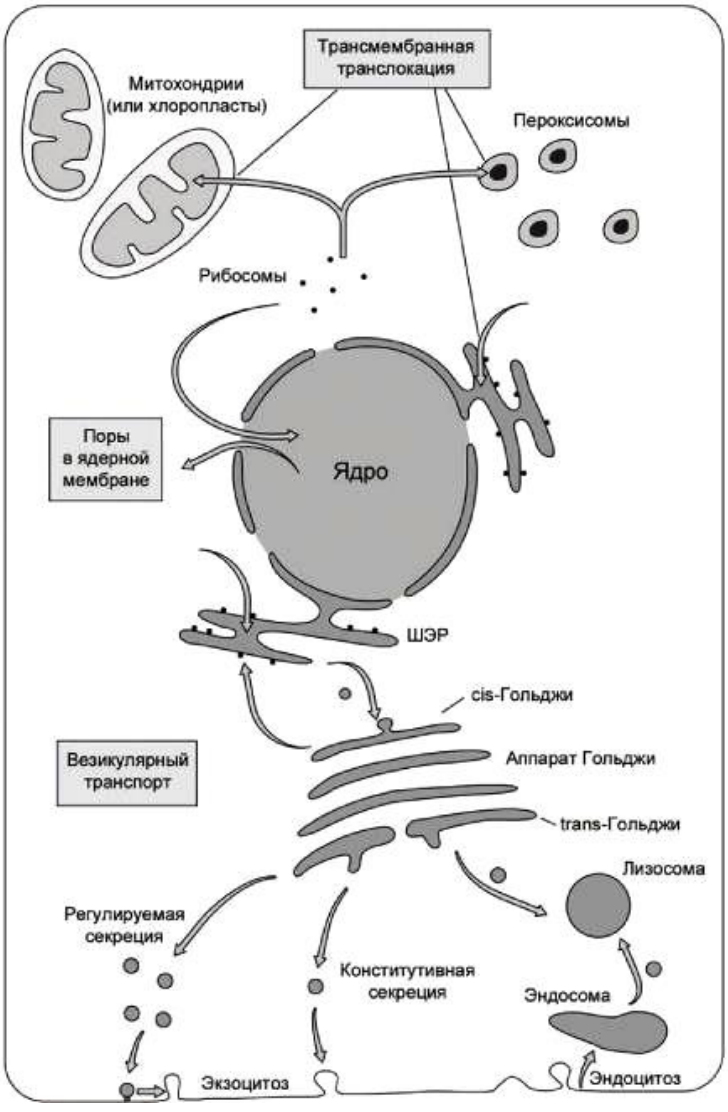


Рисунок 4.4 – Механизмы внутриклеточного транспорта белков в эукариотической клетке

1. Белок после синтеза и фолдинга в неизменном виде доставляется в нужную органеллу через специализированные поры в мембранах. Такой вид доставки называется управляемый транспорт (gated transport).

2. Трансмембранная транслокация белков в ходе которой полипептид сначала денатурируется, затем полипептидная цепочка протягивается через одну или несколько мембран, а затем снова происходит фолдинг функционального белка.

3. Везикулярное движение (vesicular trafficking) белков, в ходе которого от мембраны отпочковывается везикула, в состав которой входят транспортируемые вещества. С помощью управляемого транспорта, например, в ядро клетки через ядерные поры доставляются все вещества. Трансмембранная транслокация обеспечивает доставку синтезированных в цитозоле белков в пероксисомы, митохондрии и хлоропласты. Везикулярное движение обеспечивает доставку веществ в лизосомы и секрецию веществ из клетки.

Распределение белков обеспечивается наличием в их структуре специальных сортировочных сигналов (sorting signals). Сразу после окончания синтеза белка рибосомой в цитозоле, когда он является просто полипептидом, сортировочными сигналами являются последовательности аминокислот на концах белковой цепи, которые называют нацеливающими.

Для белков, которые синтезируются на мембранах шероховатого эндоплазматического ретикулума, дополнительные сортировочные сигналы (такие, как сахара или фосфатные группы) могут быть добавлены с помощью специализированных ферментов в цистернах аппарата Гольджи в ходе посттрансляционной модификации белков. Такие сигналы обычно представляют собой специфические лиганды, которые распознаются рецепторными белками, а эти рецепторные белки с присоединенными к ним транспортируемыми белками, в свою очередь, присоединяются к мембранным транслокационным комплексам соответствующего компартмента. Нацеливающие последовательности белка, которые представляют собой цепочку из 3–80 аминокислот, тоже распознаются специализированными рецепторами, которые доставляют данный белок к соответствующим транслокационным комплексам. После доставки в нужный компартмент нацеливающие последовательности обычно отрезаются от белковой цепи специализированными ферментами.

Одними из наиболее изученных нацеливающих последовательностей являются сигнальные пептиды (или сигнальные последовательности) – цепочки из 5–15 преимущественно гидрофобных аминокислот. Наличие такой сигнальной последовательности в синтезируемом белке вынуждает рибосому присоединиться к эндоплазматическому ретикулуму, и синтезируемая белковая цепь направляется не в цитозоль, а в люмен ретикулума. Другой пример нацеливающей последовательности – сигнал импортирования для белков, которые должны быть перемещены из цитозоля в матрикс митохондрии. Этот сигнал представляет собой цепочку из 20–80 аминокислот,

формирующих полярную α-спираль, у которой положительно заряженные аминокислоты выстроены с одной стороны спирали, а гидрофобные аминокислоты – с другой. Для транспорта белков в ядро клетки определена последовательность из пяти положительно заряженных аминокислот. Для переноса белков в пероксисому служит пероксимальная нацеливающая последовательность Ser-Lys-Lys-COOH – С-концевой трипептид. Существуют также сортировочные сигналы, которые не способствуют перемещению белка, а наоборот, служат сигналом о том, что белок уже доставлен к месту назначения и никуда далее его не следует перемещать. Например, белки с так называемой KDEL-последователь-остью Lys-Asp-Glu-Leu-COOH на С-конце остаются в эндоплазматическом ретикулуме и не должны удаляться из него везикулярным транспортом.

Иллюстрацией вышесказанного может быть кальций-связывающий белок гладкого эндоплазматического ретикулума калретикулин (calcium-binding protein of the endoplasmatic reticulum – calreticulin). Первые 17 аминокислот на N-конце калретикулина являются сигнальной последовательностью, которая инициирует транслокацию белка в люмен эндоплазматического ретикулума, а последние 4 аминокислоты – последовательность KDEL – не позволяет белку уйти из ретикулума. Между этими сортировочными сигналами располагается первичная структура функционального белка.

4.1.4 Деградация белков

Активность ферментов зависит от их концентрации, а значит, определяется балансом процессов синтеза и деградации белков в клетке. Диапазон времени жизни белков – от нескольких минут (например, белки обеспечивающие митоз, митотические циклины) до времени жизни всего организма (например, белки в хрусталике глазного яблока). Эукариотические клетки имеют несколько внутриклеточных протеолитических механизмов деградации (утилизации) в основном следующих трех типов белков:

1) неверно свернутые (misfolded) или денатурированные (развернутые) белки, 2) «нормальные» белки, чья концентрация должна быть снижена, 3) внеклеточные белки, захваченные клеткой в результате, например, эндоцитоза или фагоцитоза.

Главный внутриклеточный механизм деградации – это ферментативный протеолиз в лизосомах, чья «кислая» среда заполнена гидролитическими ферментами (рисунок 4.5).

Деградация в лизосомах предназначена главным образом для протеолиза чужих белков, попадающих в клетку в результате: 1) эндоцитоза, 2) фагоцитоза, 3) протеолиза нефункциональных органелл клетки, которые отторгаются и «перевариваются» клеткой (аутофагия).

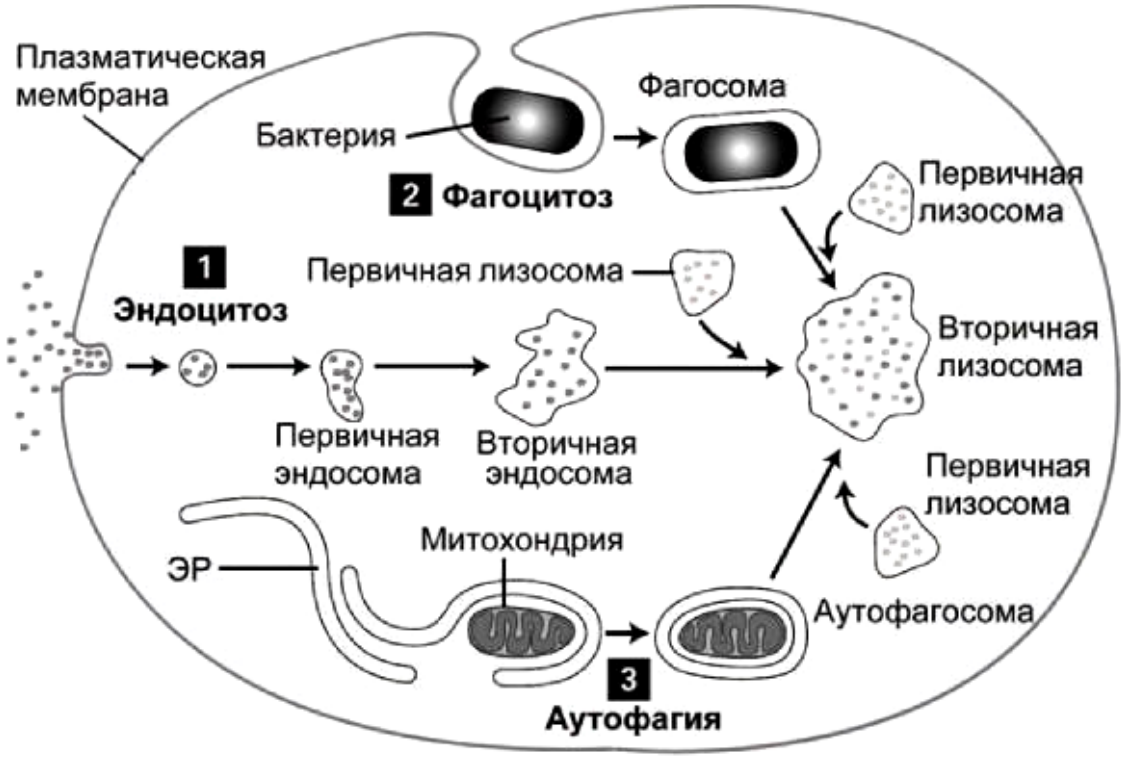


Рисунок 4.5 – Протеолиз в лизосомах

Убиквитирование. Другой, отличный от лизосомального, цитозольный механизм деградации белков реализуется путем модификации лизина в белке добавлением **убиквитина (или юбикитина, ubiquitin)**, полипептида из 76 аминокислотных остатков. Убикитированные белки деградируют затем в протеосомах. Убиквитирование осуществляется в три этапа тремя ферментами.

1. АТФ-зависимая активация убиквитин-активирующего фермента E1 (ubiquitin-activating enzyme) посредством присоединения к нему молекулы убиквитина.

2. Перенос молекулы убиквитина на цистеин убиквитинсопряженного фермента E2 (ubiquitin-conjugating enzyme).

3. Формирование пептидной связи между молекулой убиквитина, находящейся на ферменте E2, и лизином на белке, который должен деградировать, реакция катализируется ферментом убиквитин лигаза E3 (ubiquitin ligase).

Эти три этапа повторяются много раз, новые молекулы убиквитина присоединяются к уже присоединенным, в результате чего полиубиквитиновая цепь распознается протеосомой. Протеосома имеет полое цилиндрическое ядро, закрытое с двух сторон кэпирующими белками. Внутри протеосомы убиквитированные белки разрезаются на короткие (7–8 остатков) пептиды множественными протеазами, расположенными на стенках протеосомы, а молекулы убиквитина отделяются от белков.

Убиквитиновая деградация белков используется в двух случаях.

1. Удаление цитозольных белков, чья концентрация должна быть снижена. Например, циклин должен присутствовать в клетке только на определенном этапе клеточного цикла. Фосфорилирование циклина изменяет его конформацию, и внутренняя, исходно погруженная внутрь белковой глобулы, последовательность rg-X-X-Leu-Gly-X-Ile-Gly-Asp/Asn (где X – любая аминокислота) становится доступной ферментам юбиквитинизации.

2. Протеолиз белковых молекул неправильно свернутых (misfolding) в эндоплазматическом ретикулуме.

Мисфолдинг, так же как и в предыдущем случае, делает доступными те гидрофильные части белковой нити, которые спрятаны внутрь глобулы при правильном фолдинге. Эти белки переносятся в цитозоль, где распознаются ферментами юбиквитинизации.

В обоих случаях белки содержат последовательности аминокислот, распознаваемые ферментами убиквитинизации.

4.2 Мотивы и домены

В ключевое отличие между мотивом и доменом заключается в том, что мотив не является независимо стабильным, в то время как домен является независимо стабильным.

Белки − важные биологические макромолекулы, присутствующие в нашем организме. С другой стороны, генетический код гена определяет аминокислотную последовательность белка. Более того, белки имеют первичную, вторичную и третичную структуры. Первичная структура - это аминокислотная последовательность полипептидной цепи. Когда полипептидные цепи складываются друг с другом, образуется вторичная структура белка. Альфа-спирали, бета-листы и супервторичные структуры относятся к вторичным структурам. Определенные группы супервторичных элементов известны как белковые мотивы. Третичная структура белка относится к его трехмерной структуре, которая определяет функцию белка. Домен − это складчатая часть белковой молекулы, которая является глобулярной и выполняет дискретную функцию. Это фундаментальная функциональная и трехмерная структура белка.

Мотив – это определенная группа супервторичных элементов белков, таких как альфа-спирали и бета-структуры. Это своего рода узоры, присутствующие в разных белках. Мотивы описывают схемы складывания вторичных структурных элементов и их взаимодействия. Эти схемы складывания стабилизируются с помощью аналогичных связей, которые присутствуют в третичных структурах. Однако они не такие сложные, как третичные структуры.

Это простые комбинации вторичных структур белков. Мотив сам по себе нестабилен. Более того, мотивы объясняют структуру белка, но не

предсказывают функцию белка. Примерами белковых мотивов являются бета-альфа-бета-мотив, греческий ключевой мотив, бета-бочка, бета-меандровый мотив и т.д.

Домен – это фундаментальная, функциональная и трехмерная единица белка. Он выполняет определенную функцию. Один белок может иметь несколько различных доменов. Каждый домен – это независимая единица. Это шаровидная структура. Он отвечает за конкретную функцию или взаимодействие. Домены могут использоваться для предположения о функции не охарактеризованного белка. При анализе белка это важно учитывать, поскольку домены являются функциональными единицами белка (рисунок 4.6).

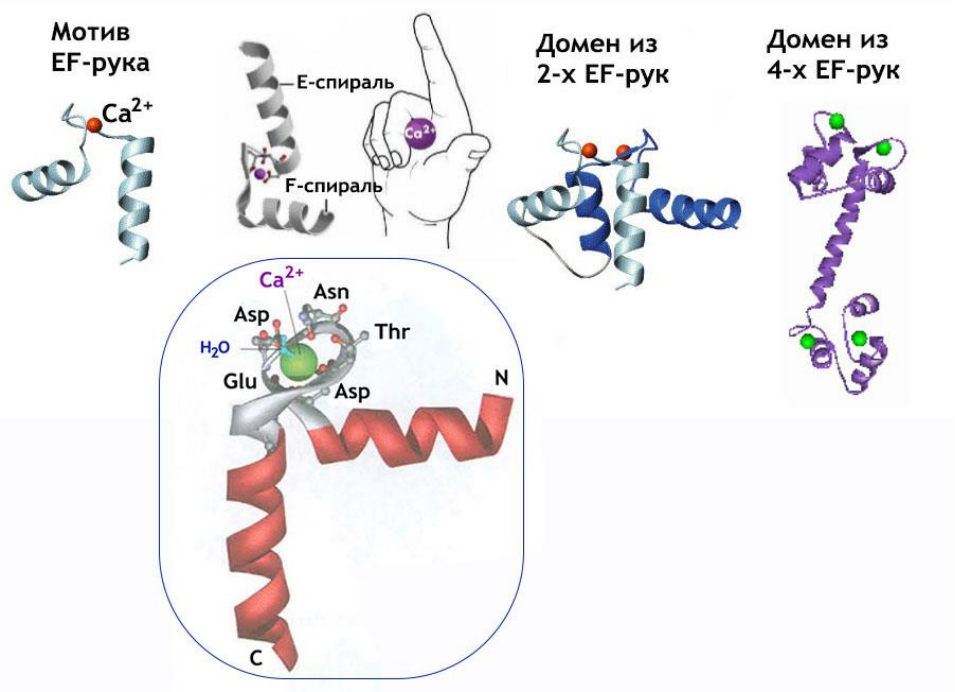


Рисунок 4.6 – Домен

Домены − это очень стабильные и компактные структуры. Их легко отличить от других регионов. Пируваткиназа имеет три различных домена, как показано на рисунке 4.6. Длина домена может варьироваться, и средняя длина составляет 100 аминокислот. Каждый домен содержит гидрофобное ядро, построенное из вторичных структурных единиц. Несколько доменов вместе составляют третичную структуру белка.

В чем сходство между мотивом и предметной областью?

- Мотив и домен представляют собой единицы, присутствующие в самих белковых молекулах.
- Они полезны при классификации семейств белков.

В чем разница между мотивом и доменом?

Мотив − это определенная группа супервторичных элементов белков, таких как альфа-спирали и бета-структуры, а домен – это функциональная

единица белка. Кроме того, мотив является вторичной структурой, а домен отвечает за третичную структуру белка. Более того, домен – это самостоятельная единица, в отличие от мотива. Кроме того, домен отображает функцию белка, а мотив − нет. Это основные различия между мотивом и доменом. В приведенной ниже инфографике в табличной форме представлена разница между мотивом и доменом.

Мотив – это расположение вторичных структур белковой молекулы. Сам по себе он обычно нестабилен, в отличие от домена. Домен – это независимо стабильная структура белка. Следовательно, это может быть часть или целая молекула белка. Это трехмерная фундаментальная функциональная единица белка. Более того, у него есть функция, и это самостоятельная единица. Мотив может быть частью домена. Но домен не может быть частью мотива. В этом разница между мотивом и доменом.

4.3 Сворачивание белков, предсказание структуры белка, предсказание функции и клеточной локализации белков

Сворачивание белков (фолдинг) – это процесс сворачивания полипептидной цепи в правильную пространственную структуру. Для обеспечения фолдинга используется группа вспомогательных белков под названием шапероны.

Предсказание структуры белка – направление молекулярного моделирования, предсказание по аминокислотной последовательности трехмерной структуры белка (вторичной, третичной или четвертичной). Данная задача является одной из самых важных целей биоинформатики и теоретической химии. Данные, полученные при помощи предсказания, применяются в медицине (например, в фармацевтике) и биотехнологии при создании новых ферментов.

Цели предсказания:

– предсказание вторичной структуры проще чем, предсказание третичной;

– аккуратное предсказание может упростить предсказание третичной структуры;

– на основе вторичной структуры можно предположить функцию белка;

– классификация белков;

– предсказание изменения структуры при функционировании белка.

Основные методы:

• статистические: Chou-Fasman method, GOR I-IV

• методы ближайших соседей: NNSSP, SSPAL

• нейронные сети: PHD, Psi-Pred, J-Pred

• HMM.

Основные предположения:

• последовательность содержит достаточно информации для предсказания;

• боковые группы определяют структуру;

• окно в 13-17 остатков достаточно для предсказания;

• основы для выбора размера окна:

• α-спирали 5-40 остатков

• β-тяжи 5-10 остатков.

Алгоритмы предсказания вторичной структуры – это набор методов предсказания локальной вторичной структуры белков, основанных только на знании об их аминокислотной последовательности. Для белков предсказание состоит в соотнесении отдельных участков аминокислотной последовательности с наиболее вероятными классами вторичных структур, таких, как α-спирали, β-тяжи или петли. Точность предсказания определяется, как соотношение количества аминокислот, для которых предсказанный структурный класс совпал со структурным классом, определенным для этой аминокислоты алгоритмом DSSP (или похожим алгоритмом, к примеру, алгоритмом STRIDE), к общему числу аминокислот в последовательности. Эти алгоритмы производят разметку аминокислотной последовательности белка в соответствии с принадлежностью аминокислот к одному из классов вторичной структуры, различающихся специфическими паттернами водородных связей и наборами двугранных углов. Для DSSP это 8 классов, которые можно объединить в три группы: 3 класса спиралей (α-спираль, π-спираль и 3-спираль), два класса β-структур (изолированные β-мостики и β-листы) и три вида петли (повороты, изгибы и неклассифицированные элементы, отвечающие характеристикам петли). Чаще всего для оценки качества структуры используют упрощенную классификацию, в которой классы внутри этих трех групп считаются тождественными. Алгоритмы предсказания вторичной структуры белка можно условно разделить на группы, основываясь на принципах, лежащих в их основе. Эти группы включают в себя статистические методы, методы ближайших соседей, методы, использующие нейронные сети, методы опорных векторов и методы, основанные на скрытых марковских моделях.

Ниже рассмотрим некоторые из этих алгоритмов.

Статистический метод Чоу-Фасмана основан на расчете оценки вероятности принадлежности определенной аминокислоты к определенному классу вторичной структуры в базах данных. Предсказание делается относительно трех классов вторичных структур: петли, β-листа и поворота. Цель алгоритма − найти участок из определенного для каждого класса вторичной структуры количества идущих подряд аминокислот, для каждой из которых оценка вероятности принадлежности к этому классу вторичной структуры больше заданного значения. На выход такие алгоритмы выдают предсказанные таким образом участки для каждого из трех основных классов вторичных структур, картированные на последовательность.

Первый этап метода ближайших соседей (алгоритм NNSSP) заключается в поиске гомологичной последовательности, для которой известна трехмерная структура. Учитывая локальные структурные особенности определенного аминокислотного остатка в трехмерной структуре гомологичной последовательности, такие, как доступность для растворителя, полярность и вторичная структура, каждому аминокислотному остатку присваивается «класс окружения». Оценка вероятности принадлежности аминокислоты в центре исследуемого сегмента длиной n аминокислот к определенному классу вторичной структуры рассчитывается как логарифм частоты нахождения этой аминокислоты в окружении, к которому относится большинство ее соседей, в базах данных.

Один из алгоритмов, использующих нейронные сети, PSIPRED, включает в себя четыре основных этапа: генерация позиционной весовой матрицы с помощью PSI-BLAST, первичное предсказание вторичной структуры и дальнейшая фильтрация предсказаний. Второй и третий этапы задействуют две нейросети. Для определения принадлежности аминокислоты к определенному классу вторичной структуры на вход первой нейронной сети подается фрагмент позиционной весовой матрицы размером 33x21, соответствующий фрагменту исходной последовательности в 33 аминокислоты с аминокислотой интереса по центру. Эта сеть имеет два скрытых слоя и три выходных узла, соответствующих трем предсказываемым классам вторичной структуры. Вторая нейронная сеть используется для фильтрации предсказаний первой сети и также обладает тремя выходного узлами для каждого класса вторичной структуры в центральной позиции исследуемого окна. На выход алгоритм выдает разметку аминокислотной последовательности по элементам вторичной структуры.

Помимо вышеописанного, классические алгоритмы с использованием скрытых марковских моделей, такие как алгоритм прямого-обратного хода, алгоритм Витерби и алгоритм Баума-Велша, могут быть оптимизированы для соотнесения аминокислотной последовательности с классами вторичных структур.

Наилучшие современные методы определения вторичной структуры белка достигают около 80% точности. Точность ныне существующих методов предсказания вторичных структур оценивается такими еженедельно обновляющимися ресурсами, как LiveBench и EVA.

Предсказание функции белка − определение биологической роли белка и значения в контексте клетки. Предсказание функций проводится для плохо изученных белков или для гипотетических белков, предсказанных на основе данных геномных последовательностей. Источником информации для предсказания могут служить гомология нуклеотидных последовательностей, профили экспрессии генов, доменная структура белков, интеллектуальный анализ текстов публикаций, филогенетические и фенотипические профили, белок-белковые взаимодействия.

Функция белка − очень широкий термин: роли белков варьируются от катализа биохимических реакций до передачи сигнала и клеточного транспорта, и один белок может играть определенную роль в нескольких клеточных процессах. В целом, функцию можно рассматривать как «все, что происходит с белком или с его помощью». Проект «Генная Онтология» предложил полезную классификацию функций, в основе которого лежит список (словарь) четко сформулированных терминов, разделенных на три основные категории – *молекулярные функции*, *биологические процессы* и *клеточные компоненты*. Из этой базы данных можно по названию белка или его идентификационному номеру найти присвоенные ему термины «Генной Онтологии» или аннотации, сделанные на основе расчетных или экспериментальных данных.

Несмотря на то что на сегодняшний день для экспериментального доказательства функций белка используются такие современные методы, как анализ микрочипов, РНК-интерференция и двугибридный анализ, технологии секвенирования продвинулись настолько, что темпы экспериментально доказательной характеристики открытых белков сильно отстают от темпов открытия новых последовательностей. Поэтому аннотирование новых белковых последовательностей будет в основном осуществляться путем предсказания на основе вычислительных методов, так как таким образом можно осуществлять характеристику последовательностей гораздо быстрее и одновременно по нескольким генам/белкам. Первые методики предсказания функций были основаны на сходстве гомологичных белков с известными функциями (так называемое *предсказание функций, основанное на гомологии*). Дальнейшее развитие методов привело к появлению *предсказаний на основе геномного контекста* и *на основе структуры белковой молекулы*, что позволило расширить спектр получаемых данных и комбинировать методики, основанные на разных типах данных, для получения наиболее полной картины роли белка. Ценность и производительность вычислительного предсказания функции генов подчеркивает тот факт, что по состоянию на 2010 год 98% аннотаций Генной Онтологии были сделаны на основе автоматического извлечения из других баз аннотаций и только 0,6% − на основе экспериментальных данных.

Методы, основанные на гомологии

Белки, имеющие сходные последовательности, как правило, являются гомологичными и, стало быть, имеют сходную функцию. Поэтому в недавно секвенированных геномах белки обычно аннотируют по аналогии с последовательностями схожих белков из других геномов. Однако не всегда близкородственные белки выполняют одну и ту же функцию, например, дрожжевые белки Gal1 и Gal3 являются паралогами с 73% и 92% сходства, приобретшие в ходе эволюции очень разные функции: так, Gal1 является галактокиназой, а Gal3 − индуктором транскрипции. К сожалению, нет четкого порога степени сходства по последовательности для безопасного предсказания функций; многие белки с одинаковой функцией имеют едва

обнаруживаемые сходства, тогда как встречаются очень схожие по последовательности, но совершенно разные по функциям (рисунок 4.7).

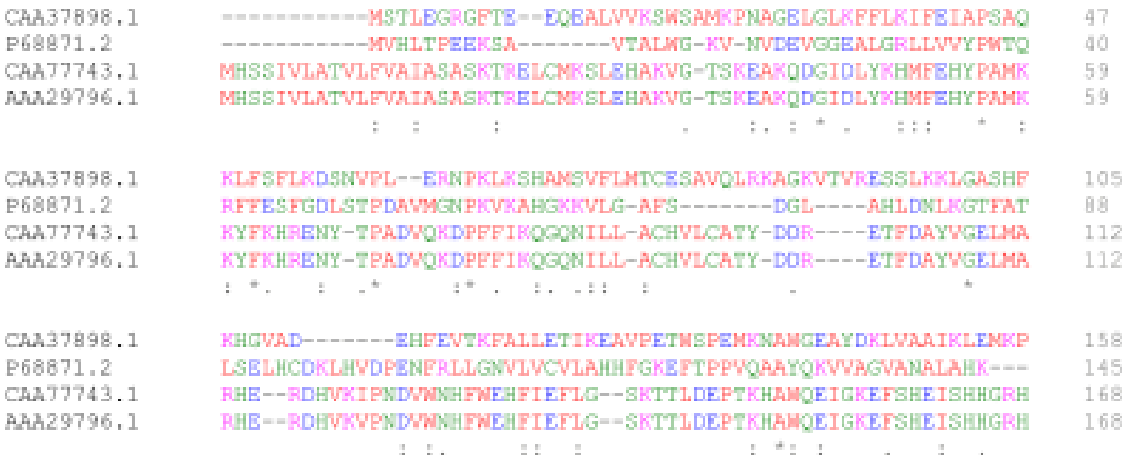


Рисунок 4.7 – Часть множественного выравнивания белковых последовательностей гемоглобина из четырых разных организмов. Белки, сходные по последовательности, могут иметь также и сходную функцию

Методы, основанные на мотивах последовательностей

Развитие таких баз данных белковых доменов, как Pfam позволяет находить в искомой последовательности уже известные домены для предположения возможных функций. В ресурсе dcGO содержатся аннотации как к отдельным доменам, так и супра-доменам (т.е. комбинациям из двух или более последовательно расположенных доменов), что позволяет сделать предсказание более приближенным к реальности. Также, внутри самих белковых доменах содержатся более короткие характерные последовательности, связанные с определенными функциями (так называемые мотивы), наличие которых в искомом белке можно определить поиском в базах данных мотивов, таких как PROSITE. Мотивы также могут быть использованы для предсказания внутриклеточной локализации белка: наличие особых коротких сигнальных пептидов предопределяет, в какие органеллы белок будет транспортирован после синтеза, и было разработано множество ресурсов для определения таких сигнальных последовательностей, например, SignalP, который обновлялся несколько раз по мере развития методов. Таким образом, некоторые особенности функции белков можно предсказать без сравнения с полноразмерными гомологичными последовательностями.

Методы, основанные на структуре белка

Поскольку 3D-структура белка, как правило, является более консервативной, чем белковая последовательность, сходство структур может указывать на сходство и функций белков. Было разработано много программ для поиска похожих укладок внутри базы данных белковых структур (Protein Data Bank), например, FATCAT, CE, DeepAlign. В случае, когда для искомой белковой последовательности нет решенной структуры, сначала составляют

вероятную трехмерную модель последовательности, на основе которой в дальнейшем делается предсказание функции белка; так работает, например, сервер по предсказанию функции белка RaptorX. Во многих случаях вместо структуры всего белка, поиск ведется по структурам отдельных мотивов, содержащим, например, сайт связывания лиганда или активный сайт фермента. Для аннотации последних в новых белковых последовательностях была разработана база данных Catalytic Site Atlas (рисунок 4.8).

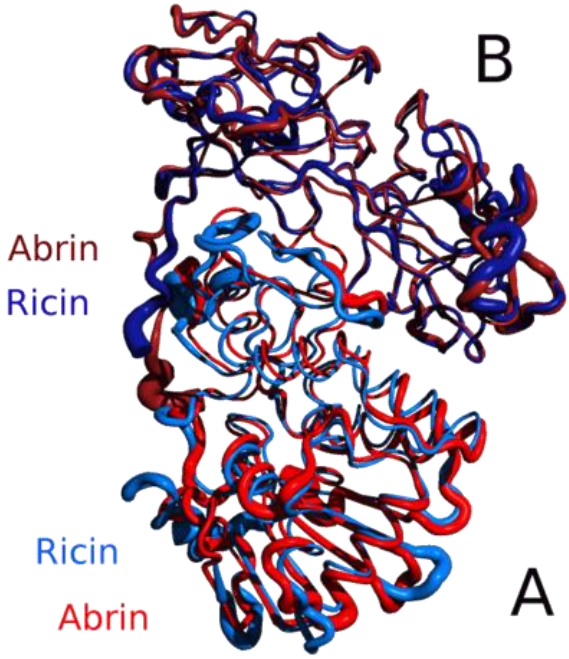


Рисунок 4.8 – Пространственное выравнивание двух беловых токсинов рицина и абрина

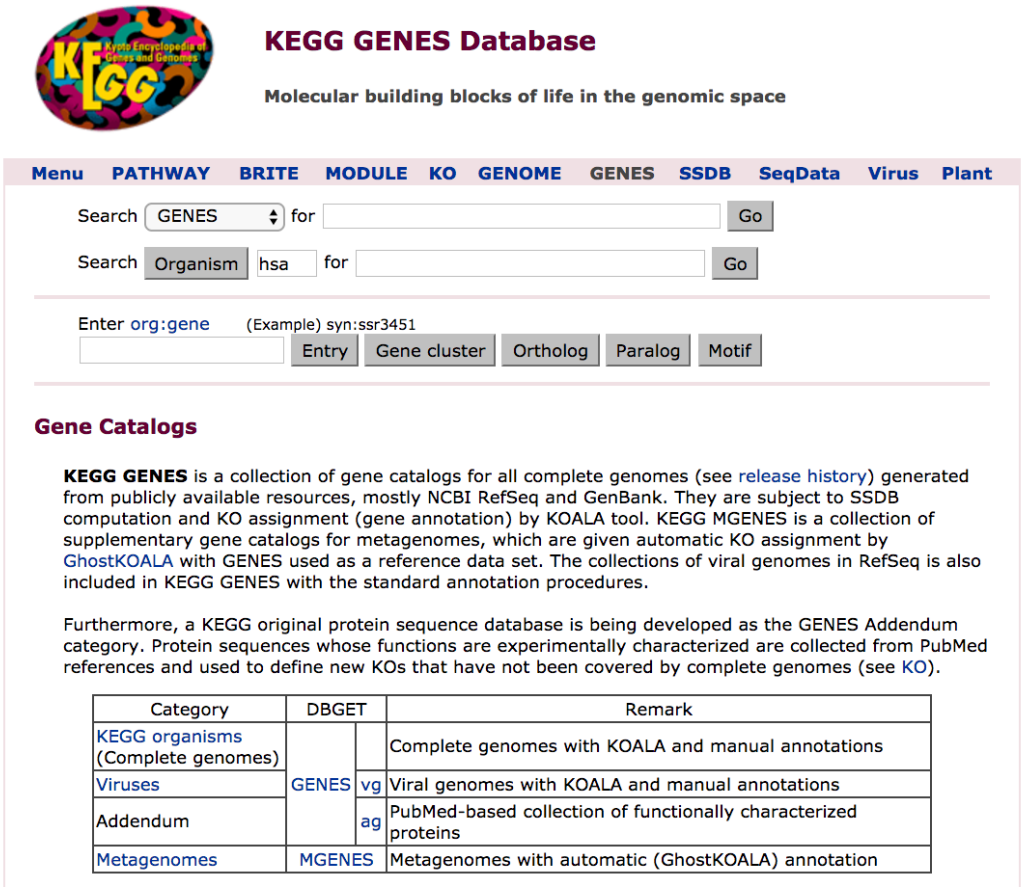
Методы, основанные на геномном контексте

Многие из недавно появившихся методов прогнозирования основаны не на сравнении последовательностей или структуры, как описанные ранее, а на корреляции между новыми генами/белками и уже аннотированными: для каждого гена составляется филогенетический профиль (по наличию или отсутствию в различных геномах), которые затем сравнивают для установления функциональных связей (предполагается, что гены с одинаковыми профилями функциональны связаны друг с другом). В то время, как методы на основе гомологии часто используются для установления молекулярных функций, предсказание на основе геномного контекста может быть использовано для предположения биологического процесса, в котором участвует белок. Например, белки, участвующие в одном и том же пути передачи сигнала, имеют общий для всех видов геномный контекст.

4.4 Энциклопедия KEGG и ее использование

KEGG (Kyoto Encyclopedia of Genes and Genomes) − веб-ресурс, предоставляющий доступ к ряду биологических баз данных и инструментам для анализа биологических и медицинских данных, созданный

в 1995 году в рамках проекта «Геном человека». С момента создания интегрированная база данных KEGG значительно расширилась и на данный момент насчитывает шестнадцать баз данных, для удобства поиска разделенных на четыре категории: системная информация, геномная информация, химическая информация и информация, связанная непосредственно со здоровьем человека. Также KEGG предоставляет ряд инструментов для удобной работы с базами данных и анализа хранящейся в них информации.



Веб-ресурс KEGG был создан в 1995 году в Японии при поддержке Kanehisa Laboratories. Базы данных KEGG непрерывно обновляются и дополняются. Главной целью проекта KEGG является интеграция полученной геномной информации, данных о биологических и химических процессах, происходящих в живых организмах, знаний о человеческих болезнях и открытых лекарствах в единое целое для понимания высокоуровневой организации различных биологических систем, таких как клетка, организм или целая экосистема.

Доступ к данным, хранящимся на KEGG, осуществляется с помощью веб-сайта KEGG. Главная страница сайта содержит список ссылок на основные базы данных KEGG, вторичные базы данных, созданные для удобного поиска, и различные инструменты для анализа биологических и медицинских данных. Представленные ссылки указывают на страницы с по-

дробным описанием каждой базы данных/инструмента и с интерфейсом поиска/работы. По ссылке KEGG2 располагается страница с полным перечнем всех баз данных и программных средств ресурса KEGG, в том числе те, которые доступны на сайте GenomeNet. Поиск данных на сайте KEGG можно осуществлять разными способами: непосредственно в основных базах данных таблица 2, по субъектам таблица 3 и по организмам. Поиск по субъектам и по организмам осуществляется с помощью интерфейсов, специально созданных для упрощения работы с базами данных.

KEGG BRITE – это тотальное структурированное формализованное описание объектов и явлений биологии, отраженных в базах KEGG. До 2005 года BRITE существовал как отдельная база данных, впоследствии включенная в проект KEGG. База данных KEGG BRITE отражает онтологию - иерархическую классификацию биологических сущностей, к числу которых относятся гены, белки, организмы, патологии, лекарственные препараты, химические соединения и т. п.

KEGG MODULE –это коллекция оформленных вручную функциональных единиц, называемых модулями KEGG, которые используются для аннотации и биологической интерпретации отсеквенированных геномов. В этой базе лежат метаболические схемы с высоким разрешением, изображающие функциональные подпути, характерные для определенных таксонов, и молекулярные комплексы, встречающиеся в этих процессах. Представлены 4 типа модулей.

KEGG GENOME – это коллекция организмов KEG с полногеномной последовательностью, каждый из которых идентифицирован трех или четырехбуквенным кодом, и некоторых вирусов, имеющих отношение к болезням.

Контрольные вопросы

1. В чем различие протекания процессов трансляции и фолдинга белка?
2. Охарактеризуйте белки шапероны.
3. Какую информацию получают при предсказании вторичной структуры белка?
4. В чем разница между мотивом и доменом?
5. Какие цели предсказания структуры белка?
6. Методы предсказания структуры белка, основанные на гомологии.
7. Методы предсказания структуры белка, основанные на мотивах последовательностей.
8. За счет чего анализ генома данного пациента позволяет проводить индивидуально специфическое лечение?
9. Какие белки называются мишенями? Как анализ геномов позволяет проводить поиск мишеней?
10. В каких случаях проводится секвенирование биологических последовательностей?
11. Какие задачи решает геномика и протеомика?
12. Как и для чего используют энциклопедию KEGG?

СПИСОК ЛИТЕРАТУРЫ

Основная

1. Биоинформатика: учебно-методический комплекс / Гом. гос. ун-т имени Ф.Скорины; сост. Дроздов Д.Н., Зятьков С.А., Гончаренко Г.Г. – Гомель: ГГУ имени Ф.Скорины, 2020. – 86 с.
2. Бизяев, Н.С. Пособие по филогенетике / Н.С. Бизяев – Киров, 2016. – 44 с.
3. Бутвиловский, В.Э. Молекулярная эволюция: материалы к факультативному курсу: курс лекций / В.Э. Бутвиловский, А.В. Бутвиловский, Е.А. Черноус. – 2-е изд. доп. – БГМУ, 2012. – 96 с.
4. Васильева, Н.Ю. Биоинформатика. Множественное выравнивание. Филогенетические деревья: методическое пособие / Н.Ю. Васильева. – Одесса: «Одесский национальный университет имени И.И. Мечникова», 2014. – 70 с.
5. Жарикова, А.А. RHK: биология и биоинформатика / А.А. Жарикова, А.А. Миронов // Молекулярная биология. – 2016. – Т. 50, № 1. – С. 80–88. – Библиогр.: с. 86–88.
6. Льюин, Б. Гены = Genes: монография / ред. Г.П. Георгиев; пер. А.Л. Гинцбург [и др.]. – Москва: Мир, 1987. – 544 с.: ил. – Библиогр. в конце глав. – Указ. предм., латин. назв.: с. 529–538.
7. Молекулярная биология клетки: пособие / сост.: М.С. Морозик [и др.]; М-во образования РБ, УО «Мозырский гос. пед. ун-т им. И.П. Шамякина». – Мозырь: УО «МГПУ им. И.П. Шамякина», 2010. – 76 с.: ил. – Библиогр.: с. 75.
8. Огурцов, А.Н. 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.

Дополнительная

9. Бельчакова Н.Л. Большой практикум по биоинженерии и биоинформатике: учеб.-метод. пособие: в 3 ч. / Н.Л. Белькова. – Иркутск: Изд-во ИГУ, 2013. – Ч. 2: Нуклеиновые кислоты. – 2014. – 155 с.
10. Бородовский М. Задачи и решения по анализу биологических последовательностей / М. Бородовский, С. Екишева. – М. – Ижевск: РХД, 2008. – 440 с.
11. Воронова, Н.В. Основы статистического анализа ДНК: учеб. материалы / Н.В. Воронова, М.М. Воробьева. – Минск: БГУ, 2015. – 17 с.
12. Гельфанд, М.С. Что может биоинформатика? / М.С. Гельфанд // Химия и жизнь – XXI век. – 2009. – № 9. – С. 10–15.
13. Геном, клонирование, происхождение человека / под общ. ред. Л.И. Корочкина. – Фрязино: Век 2, 2004. – 222 с.: ил. – (Наука для всех). – Библиогр. в конце глав. – Словарь: с. 213–218.

14. Глазко, В.И. Введение в ДНК-технологии / В.И. Глазко, И.М. Дунин, Г.В. Глазко, Л.А. Калашникова. – М.: ФГНУ «Росинформмагротех», 2001. – 434 с.

15. Дромашко, С.Е. Электронные системы для экологических исследований и образования // Экологический вестник. – 2009. – № 2. – С. 129–134.

16. Дурбин Р. Анализ биологических последовательностей / Р. Дурбин, Ш. Эдди, А. Крог, Г. Митчисон. – М. – Ижевск: РХД, 2006. – 480 с.

17. Заводник, И.Б. Актуальные проблемы биологии: протеомика, геномика, биоинформатика, метаболомика / И.Б. Заводник, Н.В. Супрун // Біялогія: праблемы выкладання. – 2011. – № 5. – С. 3–8.

18. Игнасимуту С. Основы биоинформатики / С. Игнасимуту; пер. с англ. А.А. Чумичкин. – Ижевск: Регулярная и хаотическая динамика: Ин-т компьютер. исслед., 2007. – 316 с.

19. Каменская М.А. Информационная биология / М.А. Каменская. – М.: Академия, 2006. – 361 с.

20. Картавцев, Ю.Ф. Молекулярная эволюция и популяционная генетика / Ю.Ф. Картавцев. – Владивосток, 2009.

21. Кунин, Е.В. Логика случая. О природе и происхождении биологической эволюции / Пер. с англ. – М.: ЗАО Издательство Центрполиграф, 2014 – 527 с.

22. Компьютеры и суперкомпьютеры в биологии / под ред. В.Д. Лахно, М.Н. Устинин. – Москва-Ижевск: Институт компьютерных исследований, 2002. – 528 с.

23. Леск А. Введение в биоинфрматику: пер. с англ. / А.М. Леск; ред.: А.А. Мировнов, В.К. Швядаса. – М.: Бином. Лаборатория знаний, 2009. – 318 с.

24. Лудченко, А.А. Основы научных исследований: учеб. пособие / А.А. Лудченко, Я.А. Лудченко, Т.А. Примак; под ред. А.А. Лудченко. – 2-е изд., стер. – К.: О-во «Знания», КОО, 2001. – 113 с.

25. Лукашов, В.В. Молекулярная эволюция и филогенетический анализ / В. В. Лукашов. – М.: Бином, 2009.

26. Математические методы для анализа последовательностей ДНК / под ред. М.С. Уотермена, перевод с англ. – М.: Мир, 1999. – 349 с.

27. Меллер Г.Д. Избранные работы по генетике / Г.Д. Меллер. – М.-Л.: Огиз-Сельхозгиз, 1937. – 350 с.

28. Павлинов, И.Я. Кладистический анализ (методологические проблемы). – М.: Изд-во МГУ, 1990. – 160 с.: ил.

29. Павлинов, И.Я. Введение в современную филогенетику (кладогенетический аспект) / И.Я. Павлинов. – М.: изд-во КМК, 2005.

30. Паун Г. ДНК-компьютер. Новая парадигма вычислений / Г. Паун, Г. Розенберг, А. Саломаа; пер. с англ. Д.С. Ананичева, И.С. Киселевой, О.Б. Финогеновой, ред. М.В. Волков. – М.: Мир, 2004. – 527 с.

31. Порозов, Ю.Б. Биоинформатика: учебно-метод. пособие / Ю.Б. Порозов, – СПб: НИУ ИТМО, 2012. – 52 с.

32. Приставка А.А. Большой практикум по биоинженерии и биоинформатике: учеб.-метод. пособие: в 3 ч. / А.А. Приставка. В.П. Саловарова – Иркутск: Изд-во ИГУ, 2013. – Ч. 1: Белки. – 2013. – 121 с.

33. Сидякин, В.Г. Основы научных исследований. Биология: учеб. пособие для биол. фак. ун-тов / В.Г. Сидякин, Д.И. Сотников, А.М. Сташков. – Киев: Высш. шк., 1987. – 196 с.

34. Структура и функционирование белков: применение методов биоинформатики / пер. с англ.: В.Н. Новоселецкий, Е.Д. Балицкая, Т.В. Науменкова; ред. В.Н. Новоселецкий. – М.: УРСС: Ленанд, 2014. – 414 с.

35. 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.

36. Хакен Г. Синергетика / Г. Хакен. – М.: Мир, 1985. – 423 с.

37. Чернавский Д.С. Синергетика и информация. Динамическая теория информации / Д.С. Чернавский. – М.: Либроком, 2009. – 304 с.

38. Шеннон Л. Работы по теории информации и кибернетике / К. Шеннон, Е. Бандвагон. – М.: Иностр. лит., 1963. – С. 667.

39. Шредингер Э. Что такое жизнь? Физический аспект живой клетки / Э. Шредингер. – М. – Ижевск: РХД, 2002. – 92 с.

40. Эйген М. Гиперцикл. Принципы самоорганизации макромолекул / М. Эйген, П. Шустер. – М.: Мир, 1982. – 270 с.

41. Fulekar M.H. Bioinformatics: Applications in Life and Environmental Sciences / M.H. Fulekar. – Berlin: Springer, 2009. – 247 p.

42. 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.

43. Kangueane P. Bioinformation Discovery. Data to Knowledge in Biology / P. Kangueane. – Berlin: Springer, 2009. – 166 p.

44. Liebler D.C. Introduction to Proteomics. Tools for the New Biology / D.C. Liebler. – Totowa: Humana Press, 2002. – 198 p.

45. Lesk A.M. Introduction to Bioinformatics / Lesk A.M. – Oxford: Oxford University Press, 2002. – 255 p.

46. Marcus F.B. Bioinformatics and Systems Biology. Collaborative Research and Resources / F.B. Marcus. – Berlin: Springer, 2008. – 287 p.

47. Molecular Biomethods. Handbook / Ed. by J.M. Walker, R. Rapley. – Totowa: Humana Press, 2008. – 1124 p.

48. Stephenson F.H. Calculations for Molecular Biology and Biotechnology / F.H. Stephenson. – Amsterdam: Elsevier, 2003. –302 p.

49. Ramsden J. Bioinformatics. An Introduction / J. Ramsden. – Berlin: Springer, 2009. – 271 p.

50. Selzer P.M. Applied Bioinformatics. An Introduction / P.M. Selzer, R.J. Marhöfer, A. Rohwer. – Berlin: Springer, 2008. – 287 p.

51. Statistical Bioinformatics. A Guide for Life and Biomedical Science Researchers / Ed. by J.K. Lee. – New York: Wiley, 2010. – 350 p.

Учебное издание

ЯНОВСКАЯ Виктория Владимировна

БИОИНФОРМАТИКА

Курс лекций

Технический редактор *Г.В. Разбоева*

Компьютерный дизайн *В.Л. Пугач*

Подписано в печать 04.10.2022. Формат 60x84 $^{1}/_{16}$. Бумага офсетная.
Усл. печ. л. 4,82. Уч.-изд. л. 4,37. Тираж 9 экз. Заказ 172.

Издатель и полиграфическое исполнение – учреждение образования
«Витебский государственный университет имени П.М. Машерова».

Свидетельство о государственной регистрации в качестве издателя,
изготовителя, распространителя печатных изданий
№ 1/255 от 31.03.2014.

Отпечатано на ризографе учреждения образования
«Витебский государственный университет имени П.М. Машерова».
210038, г. Витебск, Московский проспект, 33.