Лекция 23-ая «Молекулярная эволюция белков »

Спецкурс "Информационная Биология"

Лектор - к.б.н. Д.А.Афонников.

Слайд 1.
Настоящая лекция посвящена введению в предмет молекулярной эволюции.

Слайд 2.
Под термином эволюция в науке понимается изменение чего-либо во времени. Предметом теории молекулярной эволюции является изучение закономерностей эволюционных изменений генетических макромолекул в живых организмах. В основе лежит эволюционная теория - комплекс знаний об общих закономерностях и движущих силах исторического развития живой природы. Основой эволюционной теории служит утверждение о том, что все ныне существующие организмы произошли от ранее существовавших путем длительного их изменения под воздействием внешних и внутренних факторов.

Слайд 3.
Основными задачами теории молекулярной эволюции являются изучение закономерностей эволюции генетических макромолекул, а также реконструкция эволюционной истории генов и организмов.

Слайд 4.
При решении этих задач используются результаты исследований в других областях науки: палеонтологии, генетики, молекулярной биологии, биофизики, математики и информатики.

Слайд 5.
История становления и развития эволюционной теории тесно связана с развитием биологии. До возникновения эволюционного учения господствовала единственная точка зрения - о неизменности живого мира, созданного сверхъестественной силой, креационизм. В 1809 г. Ж.Б. Ламарк сформулировал первое целостное эволюционное учение о трансформации организмов. Согласно Ламарку, интенсивно функционирующие органы организмов усиливаются и развиваются, не находящие употребления ослабевают и уменьшаются, а самое главное - эти функционально-морфологические изменения передаются по наследству. Само же употребление или неупотребление органов зависит от условий окружающей среды и от присущего любому организму стремления к совершенствованию. Перемена во внешних условиях ведет к изменению потребностей животного, последнее влечет за собой изменение привычек, далее - усиленное употребление определенных органов и т.д.
Наибольшее влияние на развитие идей эволюции оказало учение Ч. Дарвина, которое было сформулировано им в работе "Происхождение видов" в 1859 году.

Слайд 6.
В основе учения Дарвина лежали следующие положения. В пределах каждого вида существует изменчивость по морфологическим, физиологическим и др. признакам. Эти вариации случайные.
Особь дает потомство, которое наследует признаки родителей. Эффективность воспроизводства особей называют приспособленностью.
Организмы дают большее потомство, чем это необходимо для простого воспроизводства численности вида.
Из-за ограниченности внешних ресурсов возникает естественный отбор, который приводит к выживанию и размножению наиболее приспособленных особей.
В целом, теорию эволюции Дарвина можно рассматривать как "хаос с обратной связью".

Слайд 7.
Развитие теории молекулярной эволюции тесно связано с развитием биологии.
В конце XIX века Мендель сформулировал понятие гена как единицы наследственности. В начале XX века концепция гена была развита в работах де Фриза, Вейсмана, Моргана и др. В 30х годах XX в. в работах математиков Фишера, Райта и Холдейна были сформулированы основы популяционной генетики - науки о генетической структуре популяций.

Слайд 8.
В середине XX века были установлены структуры генетических макромолекул - ДНК и белков. Были разработаны методы секвенирования последовательностей ДНК. Это позволило проводить сравнение последовательностей ДНК разных видов и организмов. Это позволило исследовать эволюционную изменчивость видов на уровне последовательностей ДНК и белков. Анализ этой изменчивости привел с формулированию нейтральной теории молекулярной эволюции М. Кимурой в 60-х годах XX в.
Успехи экспериментальной молекулярной биологии к концу XX в. позволили решить задачу расшифровки последовательностей ДНК полных геномов ряда живых организмов.

Слайд 9.
На современном этапе развития молекулярной биологии можно выделить несколько ключевых направлений развития биологии. Это развитие теории генных сетей, секвенирование полных геномов большого количества (в том числе и высших) организмов, проведение их полномасштабного сравнительного анализа, развитие структурной биологии.

Слайд 10.
Обратимся теперь к объектам исследования молекулярной биологии. Это генетические макромолекулы - ДНК, РНК и белки. Каждый из классов этих молекул выполняют специфические функции. ДНК является носителем генетической информации, РНК служат матрицей для синтеза белка и участвуют в других биохимических процессах клетки). Белки выполняют структурные, регуляторные, каталитические функции. Специфика этих молекул (их структурная организация и функции) определяет и особенности их эволюции. Далее мы рассмотрим особенности структурно-функциональной организации трех типов этих макромолекул.

Слайд 11.
Основной функцией ДНК является кодирование и наследственная передача генетической информации. Отметим, что последовательность ДНК постоянно подвергается стохастическим изменениям (мутациям).

Слайд 12.
Структура ДНК имеет несколько уровней организации. Первичная структура - последовательность из 4 нуклеотидов. Пространственная организация ДНК включает две цепи, связанных водородными связями комплиментарных пар оснований. Комплиментарные пары бывают сильные (C-G) и слабые (A-T). Пространственная имеет вид двойной спирали. Первичная структура ДНК может быть записана в виде строк 4-буквенного алфавита.

Слайд 13.
Согласно основной догме молекулярной биологии, генетическая информация реализуется по принципу ДНК->РНК->белок. Процесс формирования молекулы РНК называется транскрипцией, затем на рибосоме происходит синтез белка по матричной РНК

Слайд 14.

Генетическая информация записана в последовательности ДНК сложным образом. ДНК может содержать часть, которая кодирует другие генетические макромолекулы - РНК и белки. Доля кодирующей части ДНК варьируется от организма к организму и у человека составляет менее 2%. Некодирующая часть может содержать регуляторные районы, повторяющиеся участки. Функция некоторых районов ДНК пока неизвестна.

Слайд 15.
Ключевым в биологии является понятие гена. Его определение изменялось с накоплением молекулярно-биологических знаний. Согласно современным представлениям ген - это сегмент ДНК, который обуславливает фенотип или функцию. В отсутствии данных о функции ген может быть охарактеризован последовательностью, транскриптом или по гомологии. Количество генов в разных организмах различно. Так, например, у человека предполагается около 30000 генов.

Слайд 16.
Гены кодируют белки, РНК и имеют сложную структуру. На рисунке представлено схематически строение гена эукариот. В структуре гена можно выделить кодирующие участки (экзоны), некодирующие участки (интроны), регуляторные районы (промотор). В кодирующей части есть стартовый кодон, определяющий начало аминокислотной последовательности и стоп-кодон, который ее завершает.

Слайд 17.
Молекулы РНК состоят из 4 мономеров (A,U,G,C), с заменой основания тимин (T) на урацил (U). Длина молекул РНК составляет обычно от 100 до 1000 пар оснований. Это гораздо меньше, чем длина цепочки ДНК, которая может достигать более миллиона пар оснований. Молекулы РНК функционируют в виде одной нити, однако могут содержать взаимно комплиментарные участки, которые могут образовывать пары за счет водородных связей. Спаренные участки (стебли) и неспаренные участки (петли) образуют вторичную структуру РНК. В растворе РНК принимает сложную конформацию, в которой даже монономеры, удаленные по цепи и не образующие пар могут контактировать (дальние контакты). На рисунке приведена пространственная (третичная) структура молекулы транспортной РНК (слева) и ее вторичная структура (справа).

Слайд 18.
Третьим важным классом генетических макромолекул являются белки. Это нерегулярные гетерополимеры, состоящие из 20 мономеров - аминокислот. Структура аминокислоты включает основную цепь и боковую группу. Основная цепь одинакова у всех аминокислот. Боковые группы аминокислот различаются по своей химической структуре и свойствам. Именно они определяют физико-химические особенности и взаимодействия аминокислотных остатков. Аминокислоты в белках связаны между собой пептидной связью. Существует несколько уровней организации белка. Первый из них - его аминокислотная последовательность (первичная структура). Вторичная структура это локально упорядоченные участки основной цепи белка. Наиболее часто встречаются два типа вторичных структур - альфа спирали и бета-нити. Вторичная структура образует жесткие сегменты полипептидной цепи, которые укладываются в третичную структуру. Стабильность упаковки пространственной структуры белка обеспечивается взаимодействиями удаленных по первичной последовательности остатков. Необходимо отметить, что белковая глобула гетерогенна и содержит внутреннюю гидрофобную часть (ядро) и полярную оболочку.

Слайд 19.
На данном слайде схематически представлены различные уровни структурной организации белка - от первичной структуры до четвертичной.

Слайд 20.
Последовательность белка кодируется в последовательности матричной РНК по правилам генетического кода. Каждая тройка нуклеотидов (кодон), кодирует аминокислоту. Генетический код является вырожденным, т.е. несколько кодонов могут кодировать одну аминокислоту. Разные аминокислоты могут кодироваться разным числом кодонов. Например, триптофану соответствует один кодон - UGG, а лейцину - 6. Отметим, что тип кодируемой аминокислоты определяется чаще всего тремя первыми нуклеотидами и слабо зависит от третьей позиции. Генетический код так же содержит три стоп кодона, которые прекращают трансляцию белка.

Слайд 21.
Отдельные изменения генотипа называются мутациями. Мутация - основа наследственной изменчивости в живой природе. Мутации могут приводить к изменениям больших сегментов ДНК. К таким мутациям относятся транслокации - перенос гена в другое место генома; дупликации - удвоение участка гена; инверсия - поворот участка гена на 180 градусов; делеции - удаление участка гена.
Существует так же класс точечных мутаций, которые сводятся к заменам одного нуклеотида в последовательности ДНК.

Слайд 22.
Точечные замены нуклеотидов могут быть нескольких типов. Во-первых, мутация может приводить к замене аминокислоты в последовательности белка (миссенс-мутация). Мутация может приводить к возникновению стоп-кодона (нонсенс-мутация). В третьих, вставки или делеции 1-2 нуклеотидов могут приводить к сдвигу рамки считывания белка. В четвертых, вставка или делеция трех нуклеотидов может приводить к вставке/делеции одной аминокислоты. На рисунке приведены несколько примеров мутаций разных типов, которые вызывают наследственное заболевание "кистозный фиброз".

Слайд 23.
Молекулярная эволюция исследует закономерности временных изменений наследственной информации на различных уровнях ее организации, хромосомном, на уровне последовательностей ДНК, РНК и белков. В дальнейшем, однако, мы будем рассматривать изменение последовательностей макромолекул (ДНК, РНК и белков). Источником таких вариаций являются мутации. Поэтому можно сказать, что молекулярная эволюция это теория о возникновении и дальнейшем поведении мутаций. В этой связи одной из важных задач является задача о распределении частот мутантных генов в популяции. Рассмотрим классическую модель популяции особей с двойным (диплоидным) набором генов в генотипе и бесконечно большой численностью. Предполагается, что у гена есть две формы (аллеля) - дикий тип A и мутантный тип B. В этом случае для диплоидного генома возможны три генотипа - два гомозиготных AA,BB, с одинаковыми аллелями, и один гетерозиготный AB. Пусть частота гена A в популяции - p, гена B - q, p+q=1. В этом случае частоты генотипов AA,AB и BB равны, соответственно, pp,2pq, qq и выполняется закон Харди-Вайнберга, $p^2+2pq+q^2=1$. Таким образом, зная частоты генов можно вычислить частоты генотипов и наоборот. Например, известно, что заболевание "кистозный фиброз" вызывается мутантным аллелем нормального гена (A®B) и проявляется в случае генотипа BB. Частота этого заболевания известна, и равна $1/1700= q^2$. В таком случае, q=0.024; p=0.976; $p^2=0.953$. Это пример простейшей модели распределения мутантных генов в популяции.

Слайд 24.
Другой важной с точки зрения теории молекулярной эволюции задачей является исследование временной динамики частот генов в популяции. Как и в предыдущем случае рассмотрим ген, имеющий дикий тип A и мутантный аллель B, генотипы AA,AB,BB.

Численность популяции является бесконечно большой. В качестве единицы временной шкалы будем считать номер поколения. Усложним модель, по сравнению с предыдущей. Пусть генотипы обладают различной приспособленностью. За единицу приспособленности возьмем генотип AA. Относительное изменение приспособленности генотипов AB и BB обозначим sAB и sBB. Динамика изменения частот генов в популяции зависит, очевидно, от соотношения приспособленностей. Назовем фиксацией мутантного аллеля событие, когда все особи популяции имеют данный ген (этот ген становится геном дикого типа). На рисунке показано три типа поведения динамики в зависимости от величин sAB и sBB - селективного преимущества. В первом случае, приспособленность гетерозиготной особи не отличается от приспособленности гомозиготной по аллелю A, а приспособленность BB выше. Видно, что за счет низкой начальной частоты гена B его частота растет вначале медленно, затем за очень малый относительный промежуток времени этот ген фиксируется. Во втором случае, даже в составе гетерозиготы ген B дает преимущество в приспособленности по сравнению с геном AA, поэтому ген B фиксируется практически сразу и за короткое время. В третьем случае, происходит резкое увеличение частоты гена B, однако за счет преимущества гетерозигот, частота гена A падает очень медленно.
В целом, для вариантов соотношения приспособленностей (1) и (2) характерной особенностью динамики является окончательная фиксация гена, имеющего большую приспособленность, и то, что промежуток времени, когда частоты генов A и B сравнимы - очень мал.

Слайд 25.
Ситуация меняется существенно, если считать, что численность популяции конечна. Теоретические исследования такой модели провел М.Кимура и показал, что для конечных популяций важным фактором является случайные флуктуации (в этом случае ими пренебрегать нельзя). Основной вывод теории Кимуры состоит в том, что вероятность фиксации мутации зависит не только от ее приспособленности, но и от численности популяции (см. формулу). В этом случае даже для некоторых мутаций, дающих отрицательный вклад в приспособленность генотипа, существует ненулевая вероятность фиксации. На рисунке показана зависимость вероятности фиксации мутации (ось Y) от ее селективного преимущества s (ось X) и численности популяции N (100, 1000 и 10000). Видно, что чем меньше размер популяции, тем существеннее ее поведение отклоняется от классической модели.
Таким образом, в конечных популяциях могут фиксироваться даже селективно вредные мутации.

Слайд 26.
Эффект конечности популяции помогает объяснить высокую степень полиморфизма, т.е. тот факт, что в реальных популяциях частота мутантных аллелей является достаточно высокой на протяжении многих поколений. Процесс динамики мутантных генов в стохастической модели Кимуры изображен на рисунке. При этом предполагается, что основная доля мутаций, фиксировавшихся в популяции, это селективно нейтральные мутации. Для *нейтральных* мутаций существуют два основных параметра, определяющих их динамику в популяции - время фиксации мутации t_{fix} и частота фиксации мутации K, которая равна частоте их возникновения u. Для адаптивных замен $K=4Nsu$.

Слайд 27.
Основные положения теории молекулярной эволюции, разработанной Кимурой, заключаются в следующем.

- Адаптивные мутации очень редки;
- Деструктивные мутации быстро элиминируются

- Подавляющее большинство фиксировавшихся мутаций селективно нейтральны
- Скорость фиксации замен в генах является постоянной (гипотеза молекулярных часов, Zukercandle & Poling
- Функционально важные районы функционируют медленнее, менее важные - быстрее
  На рисунке показана схема изменения частот мутаций разного типа до и после естественной селекции.

Слайд 28.
Рассмотрим известные примеры, которые согласуются с нейтральной теорией эволюции. Широко известен факт постоянства скоростей аминокислотных и нуклеотидных замен (этот эффект по аналогии с атомными часами в физике был назван Цукеркендлом и Полингом "молекулярными часами"). На рисунке приведен график зависимости геологического времени расхождения видов (ось Y) и числа накопленных аминокислотных замен (ось X). Видно, что эта зависимость линейна, что отражает постоянство скорости фиксации мутаций. Причем для разных белков эта скорость различна. Для белков, функциональная нагрузка которых считается невысокой (фибринопептиды), эта скорость наибольшая из приведенных. Этот факт хорошо согласуется с положениями Кимуры о том, что , скорости фиксации замен постоянны и они выше для позиций (белков), которые несут меньшую функциональную нагрузку.

Слайд 29.
Еще одним интересным фактом может служить сравнение скоростей замен в кодонах по первой и второй позиции (т.н. несинонимические замены), и по третьей позиции, замены в которых слабо влияют на замену соответствующих аминокислот в последовательности белка (синонимические замены). Из приведенной таблицы видно, что скорости несинонимических замен в несколько раз меньще, чем синонимических. Для белков отвечающих за упаковку ДНК в нуклеосомах (гистоны) эта скорость практически равна 0. Т.е. их эволюция подвержена сильным ограничениям. Однако частоты несинонимических замен в них близки к частотам в таких сильно вариабельных белках, как иммуноглобулины.

Слайд 30.
Перейдем теперь к математическим моделям, описывающим эволюцию последовательностей ДНК, РНК и белков. Цель заключается в наиболее полном описании процесса одиночных нуклеотидных/аминокислотных замен.
В основе большинства моделей лежит теория Марковских цепей с конечным числом состояний. Полнота описания достигается как правило за счет усложнения моделей и переходу к большему числу параметров модели.

Слайд 31.
В качестве простейшей модели рассмотрим последовательность нуклеотидов бесконечной длины. Эту модель можно описать одним параметром - вероятностью замены a одного нуклеотида на другой за единичный промежуток времени. Этот параметр одинаков для всех возможных пар замен (представленных в виде схемы слева). Такая модель была предложена Джуксом и Кантором и носит их имя. Отметим, что замены в позиции происходят случайным образом, независимо от других позиций последовательности. Поэтому вероятности наблюдать нуклеотиды можно оценить по частотам их встречаемости в последовательности.
Рассмотрим, как зависит вероятность наблюдать нуклеотид A в момент времени t+1 - PA (t+1), при условии, что в момент времени t частота нуклеотида A известна и равна $P_A(t)$.

$$P_A(t+1) = (1-3\alpha)P_A(t) + \alpha[1 - P_A(t)] = -4\alpha P_A(t) + \alpha$$

Здесь первое слагаемое отражает уменьшение частоты нуклеотида A за счет мутаций в

любой из трех других нуклеотидов (A->X), второе слагаемое отражает увеличение частоты нуклеотида A за счет мутаций трех других нуклеотидов (X->A). Если перейти к непрерывному времени, то это выражение сведется к дифференциальному уравнению

$$\frac{dP_A(t)}{dt} = -4\alpha P_A(t) + \alpha$$

Слайд 32.
Решение этого уравнения записывается как

$$P_A(t) = 0.25 + (P_A(0) - 0.25)e^{-4\alpha t}$$

Где $P_A(0)$ - вероятность наблюдать нуклеотид A в момент времени t=0. Если

$$P_A(t) = 0.25 + 0.75e^{-4\alpha t}$$

(верхний график на рисунке слева),
Если PA(0) = 0, то

$$P_A(t) = 0.25 - 0.25e^{-4\alpha t}$$

(нижний график на рисунке слева). Отметим ряд важных особенностей модели эволюции. Во-первых, при времени большом эволюции, частоты всех нуклеотидов стремятся к своим равновесным значениям, одинаковым и равным 1/4.
Во вторых, зависимость $P_{AA}(t)=P_{ii}(t)$ (вероятность нуклеотида остаться неизменным через промежуток времени t) не зависит от типа нуклеотида. В третьих, вероятность нуклеотида мутировать на любой другой $P_{ij}(t)$ так же не зависит от конкретного типа нуклеотида.

Известно, что в ДНК частоты замен зависят от типа нуклеотидов. В частности частоты замен типа A↔T и A↔G,C различаются. Замены типа A↔T и G↔C называются трансверсиями, типа A↔G,C - транзициями. Различие частот трансверсий и транзиций было выявлено при сравнении ДНК разных организмов. Для описания таких различий Кимура модифицировал модель нуклеотидных замен, введя в нее дополнительный параметр. В модели Кимуры вероятности транзиций ане равны вероятностям трансверсий Бета. Из этого вытекает изменение динамики замен. Во-первых, поведение модели зависит от двух параметров. Во вторых, если вероятности P_{ii} не зависят от типа нуклеотида i, то P_{ij} уже зависят от того, к какому типу относится замена (транзиция или трансверсия). Однако, неизменным остается экспоненциальный характер зависимости частот нуклеотидов от времени, который приводит к тому, что через бесконечно большой промежуток времени частоты нуклеотидов примут стационарные значения.

Слайд 34.
В общем случае вероятность одномоментных замен нуклеотидов может быть описана т.н. матрицей замен. Размер этой матрицы 4x4, она описывает линейное преобразование вектора P частот нуклеотидов за короткий промежуток времени (см. выражения для частот слева). В Векторной форме это преобразование записывается как , где p - вектор частот в момент времени t, p' - вектор частот в момент времени t'=t+dt. Свойства матрицы замен таковы: сумма ее элементов по строкам составляет 1; эволюция за n дискретных промежутков времени равна умножению на матрицу M^n. Если M=const, то существуют равновесные частоты, которые находятся из уравнения
p=Mp.

Слайд 35.
Обратимся теперь к модели аминокислотных замен в белках. В этом случае алфавит описывается набором из 20 символов - канонических аминокислот. Таким образом размерность вектора частот равна 20. Простейшая модель аминокислотных замен на основе

матрицы замен 20x20 была предложена Маргарет Дайхофф, в 70х гг. Эта модель и сейчас широко используется при анализе аминокислотных последовательностей. В работе Дайхофф матрица замен M была определена эмпирически на основе сравнения гомологичных белков нескольких семейств. Особенности матрицы Дайхофф таковы.

- Равновесные частоты равны частотам встречаемости аминокислот в последовательностях белков.
- Наиболее часты замены аминокислот на аминокислоты, сходные по физико-химическим свойствам.
- Исходная матрица нормирована на время, эквивалентное 1 замене на 100 позиций (1PAM), т.е. время измеряется в единицах PAM.
- Для оценки вероятности замен через время t=n надо матрицу 1PAM возвести в степень n.

Слайд 36.
Рассмотрим теперь проблему определения эволюционного расстояния между последовательностями. Она заключается в определении взаимосвязи между различиями на уровне последовательностей ДНК/РНК/белков и временем (геологическим) эволюции видов. Очевидно, знание такой зависимости позволяет производить датировку молекулярных событий (расхождение видов) на основе сравнения их последовательностей макромолекул. В простейшем случае, когда замены редки (или время эволюции мало) можно предположить, что число замен в паре последовательностей прямо пропорционально времени их эволюции. На рисунке показана предковая последовательность (изменившиеся позиции показаны синим) и две дочерних, которые после дивергенции видов от предковой эволюционировали независимо (изменения последовательности показаны красным). Заметим, что расстояние между двумя последовательностями отражает удвоенное время, прошедшее с момента дивергенции (при условии, что скорости накопления замен в двух последовательностях были одинаковы). Для определения времени эволюции в такой модели достаточно знать долю p несовпадающих символов в двух последовательностях. Расстояние будет вычисляться как $d=-\log_e(1-p)$. Зная скорость замен l можно вычислить время эволюции, или зная d и t можно вычислить скорость замен. Недостатком такого подхода является его неприменимость на больших временах эволюции, когда сказываются повторные и обратные замены нуклеотидов.

Слайд 37.
Ситуацию можно улучшить, применяя выражения, связывающие число замен и время эволюции, полученные на основании более сложных моделей эволюции, описанных нами ранее. Так, в приближении модели Джукса-Кантора

$$d = -\frac{3}{4}\log_e(1-\frac{4}{3}p).$$

При использовании модели Кимуры (двухпараметрической)

$$d = -\frac{1}{2}\log_e[(1-2P-Q)\sqrt{1-2Q}]$$

(P-частота транзиций, Q-частота трансверсий). В самом общем случае матрицы замен

$$d = 2(\sum_1^4 p_i\lambda_i)t$$

Слайд 38.
В случае эволюции аминокислотных последовательностей рассматриваются матрицы семейства PAM. Матрицы серии PAM-N отражают вероятность замены аминокислот ai->aj за

время эволюции, эквивалентное N PAM единиц. Зависимость между расстоянием PAM и числом замен так же не является линейной (представлена на рисунке). Она близка к линейной на малых временах эволюции (в пределах 30% замен или ~50 PAM).

Слайд 39.
В молекулярной эволюции эволюционные отношения между последовательностями из разных видов описываются филогенетическими деревьями. Как бравило это бинарные деревья. Точки ветвления соответствуют моментам разделения биологических видов, после которых потомки продолжают эволюцию независимо. Филогенетические деревья могут иметь корень - предковую по отношению ко всем современным последовательность. Как правило, нам известны последовательности современных видов, т.е. только конечных вершин (таксонов). Последовательности внутренних узлов неизвестны, хотя и существуют некоторые методы их восстановления.

Слайд 40.
Не все методы построения филогенетических деревьев могут выдавать точное положение корня. Некоторые алгоритмы безразличны к его положению (т.е. не дают оценку положения корня). В принципе положение корня может быть выбрано на любой из ветвей дерева (см. рисунок), выбор зависит от представлений биолога или требует дополнительных данных.

Слайд 41.
Обратимся теперь к некоторым наиболее известным методам построения филогенетических деревьев. Наиболее давний из них - UPGMA (Unweighted Pair Group Method with Arithmetic Mean). При построении такого дерева вначале выбирается ближайшая пара таксонов. Далее из оставшихся таксонов выбирается тот, среднее расстояние от которого до пары объединенных таксонов (кластера) наименьшее. Процедура продолжается далее, причем расстояние между кластером X и кластером Y равно среднему от парных расстояний между последовательностями этих кластеров. В итоге получаем дерево с корнем. Метод UPGMA имеет следующие особенности.

- Предполагает равномерность замен (молек.часы) для всех таксонов
- Расстояние=2*длину ветви
- Дает всегда дерево с корнем
  Отклонение от постоянства скоростей замен может привести к неправильно восстановленной топологии дерева.

Слайд 42.
Метод ближайшего соседа не требует постоянства скоростей замен по таксонам, он позволяет правильно восстанавливать топологию дерева, даже если такое постоянство нарушено. Сущность его заключается в определенном правиле пересчета эволюционных расстояний.

Слайд 43.
Метод ближайшего соседа восстанавливает топологию дерева без учета положения корня. Примеры таких деревьев для тех же последовательностей, что и в случае UPGMA - приведены на рисунке.

Слайд 44.
Топологию дерева можно перестраивать (при постоянном числе конечных вершин), перестановкой соседних вершин, сокращением и перестройкой поддеревьев, изменением длин вершин. Иногда говорят о "пространстве топологий". Это пространство велико. Известно, что полное число различных топологий бинарных деревьев для данного числа

конечных вершин растет экспоненциально.

Слайд 45.
Задачу построения филогенетического дерева можно переформулировать. Пусть имеется для данного набора таксонов всевозможный набор филогенетических деревьев. Задача заключается в поиске такой топологии, которая "наилучшим образом" описывала бы наблюдаемые различия в последовательностях генов. В этом случае задача построения дерева сводится к выбору оптимума в пространстве топологий. Понятие "наилучшим образом" может трактоваться по-разному. В частности, широко используется метод максимальной парсимонии. Здесь критерием оптимума является минимизация числа замен по ветвям дерева. Принцип парсимонии продемонстрирован на рисунке.
1)В первом дереве изменения происходят только один раз (+)
2)Во втором дереве 1 появляется (+) и теряется (*)
3)В третьем дереве 1 появляется независимо два раза (+)
Дерево (1) содержит минимальное число эволюционных событий - его и выбираем.
Отметим, что при такой постановке возникает самостоятельная нетривиальная задача поиска оптимума в пространстве деревьев.

Слайд 46.
Дальнейшим развитием такого подхода является метод максимального правдоподобия. В этом случае целевой функцией является функция правдоподобия L(D|T) - вероятность наблюдения данных (D), при условии, что эволюция происходила по данной топологии (T) - функция правдоподобия.

- Вычисляется вероятность наблюдения данных (D), при условии, что эволюция происходила по данной топологии (T) - функция правдоподобия L(D|T)оВычисляется рекурсивно от листьев к вершине
- Выбирается дерево, которое дает max(L)
- Решение зависит от выбора модели замен

Слайд 47.

- С помощью методов МП можно оценивать и другие параметры: так как и матрица замен, и скорости замен могут быть такими параметрами. L=L(T,M,t…..).
- Можно усложнять модель, добавляя новые параметры.
- Метод имеет статистическое обоснование
- Но требует большого количества вычислений
- Не эффективен при большом числе параметров и малом количестве данных

Слайд 48.
Примером успешного применения методов филогенетического анализа служит открытие царства архебактерий. Woese et al (1970-е) сравнивали последовательности малых субъединиц 16S rRNA и обнаружили, что архебактерии ближе к эукариотам, чем к бактериям. Архебактерии были выделены в отдельное царство. (см. филогенетическое дерево).
Слайд 49.
Отметим, что в природе встречаются отклонения от базовых моделей замен нуклеотидов/аминокислот. При этом сильное влияние оказывает структура макромолекул. Возникает потребность в совершенствовании существующих моделей с обнаружением новых

фактов/явлений.

Слайд 50.
В частности особенностью РНК являются зависимые (коррелированные) замены пар нуклеотидов, формирующих взаимодействия - шпильки. Таким образом вторичная структура РНК (шпильки) накладывает ограничения на паттерны замен.

Слайд 51.
Гораздо более сложная ситуация возникает при попытке более точно описать эволюцию белков. Белки - гетерогенные объекты и содержат различные типы вторичных структур, различные участки доступности растворителю, функциональные сайты. Поэтому давление отбора для разных участков белка различно.

Слайд 52.
В качестве отклонений от модели эволюции, предложенной Дайхофф можно указать следующие:

- скорости фиксации замен для различных ветвей эволюционного дерева могут быть различными, что не соответствует предположению о "молекулярных часах" (Ayala et al., 1997; Ayala, 1997);
- скорости фиксации замен аминокислот могут различаться для различных позиций белка вдоль его последовательности (Uzzel and Corbin, 1971; Morozov et al., 2000);
- вероятности замен аминокислот (определенные как элементы матрицы аминокислотных замен) для различных участков глобулы или даже для отдельных позиций белковой последовательности могут быть различными (Overington, 1992).
- параметры эволюционного процесса (в частности элементы матрицы M) могут зависеть от времени (Benner et al., 1994).
- мутации в различных позициях белка могут фиксироваться зависимым образом.

Слайд 53.
Пример неравномерности скоростей замен по первичной последовательности белка приведен на рисунке. Показан профиль относительного темпа замен в позициях k-цепи иммуноглобулина - белка системы иммунного ответа.

Слайд 54.
Другой особенностью замен в белке является различие в структуре и функции позиций белка. Это приводит к тому, что режимы замен (матрицы вероятностей замен M) оказываются для разных позиций разными. Koshi and Goldstein оценили матрицы замен для разных структурных участков белка. Сравнение матриц замен для петель на поверхности белка (слева) и погруженных спиралей (справа) приведено на рисунке. Видно, что в петлях частыми являются вставки и делеции (замены на символ "-" - крайний левый столбец матрицы) а так же повышена частота замен между полярными аминокислотами (верхний правый угол матрицы). Во внутренних участках белка превалируют замены гидрофобных остатков на гидрофобные (нижний левый угол соответствующей матрицы).

Слайд 55.
Для учета эффекта гетерогенности белка в ряде работ предложено было рассматривать матрицы замен, специфические либо для белковых семейств, либо для участков одинаковых по структурно-функциональным свойствам.

Слайд 56.

Очевидным недостатков такого подхода является слишком большое число оцениваемых параметров (матриц замен, каждая из которых содержит 210 параметров) и чем точнее мы желаем описать эволюцию белка - тем большее число параметров нам необходимо учитывать.

Слайд 57.
Koshi and Goldstein предложили следующее решение такой проблемы. Они предложили параметризовать матрицы замен меньшим числом параметров. Это было сделано в предположении о больцмановском распределении частот встречаемости аминокислот, которое зависит от некоторой функции приспособленности остатка. Функцию же приспособленности можно описать небольшим числом параметров. В результате для определения матрицы замен необходимо было оценить всего от 3 до 5 параметров.

Слайд 58.
Другой особенностью замен аминокислот в белках могут служить координированные замены . Это такие замены, которые происходят зависимым образом в силу того, что приспособленность белка зависит от определенного сочетания аминокислот в двух или более позиций белка. На рисунке приведен иллюстративный пример координированной фиксации замен в паре остатков, формирующих солевой мостик. Белки с одноименно заряженными парами остатков менее стабильны и под давлением отбора будут элиминированы в ходе эволюции. Белки с разноименными парами будут иметь селективное преимущество. Таким образом, набор родственных последовательностей современных белков будет содержать характерные паттерны замен аминокислот, свидетельствующих об их статистической зависимости.

Слайд 59.
Одной из моделей, которая позволяет описывать зависимые замены является ковариионная модель. Согласно этой модели число аминокислотных позиций, в которых замены могут фиксироваться в каждый момент времени, является постоянным, однако набор этих "потенциально" мутирующих позиций меняется от таксона к таксону. Для обозначения множеств таких наборов позиций, способных фиксировать мутации, Фитчем был введен термин covarion (сокращение от "concomitantly variable codons" - "одновременно меняющиеся кодоны"). Рассматривается три класса позиций белка (Miyamoto and Fitch, 1995): 1) пул коварионов, с набором позиций, которые могут мутировать в конкретный момент времени на эволюционной шкале (обозначение - c); 2) класс потенциально вариабельных сайтов, в которых в текущий момент замены фиксироваться не могут (ti - temporary invariable); 3) категория полностью инвариантных позиций (pi - permanently invariable). Скорость, с которой происходит изменение класса позиций c->ti, обозначается v (Fitch, 1971).

Слайд 60.
Среди методов выявления координированных замен остатков в белках отметим следующие:

- Методы теории информации
- Анализ статистики парных частот встречаемости символов
- Методы максимального правдоподобия
- Анализ корреляций физико-химических свойств

Слайд 61.
Например, в нашей лаборатории разработан пакет программ анализа координированных замен CRASP. Его схема представлена на слайде. На вход подается множественное выравнивание белков, затем эти данные преобразуются в матрицу чисел, по которой

оцениваются корреляции. Матрица коэффициентов корреляции выдается как в текстовом, так и графическом виде. Отдельный блок программ позволяет проводить анализ интегральных характеристик. Этот пакет программ имеет гипертекстовый интерфейс и расположен на сайте лаборатории теоретической генетики ИЦиГ.

Слайд 62.
Нами так были проанализированы координированные замены в ДНК-связывающих доменах семейства гомеодомен. Выборка составила 372 последовательности, анализировались физико-химические свойства, отражающие заряд остатка (изоэлектрическая точка аминокитслот). По матрице парных коэффициентов корреляции этого свойства в позициях белка было построено дерево близости позиций, на котором при уровне значимости 99.9% выделено два кластера.

Слайд 63.
Характерной особенностью первого кластера является то, что все значимые коэффициенты корреляции оказались отрицательными, т.е. замены в нем являются компенсаторными. Мы предположили, что для этого кластера консервативным свойством будет суммарная величина изоэлектрической точки. Табличный критерий показал, что вклад координированных замен в эту характеристику является значимым консервативным. Тесты Монте-Карло показали, что ни в одной из 100000 случайных выборок величина дисперсии этой характеристики не была меньше реального значения. Интересно, что все остатки этого кластера расположены в районе контакта двух спиралей, а часть образует солевые мостики. Так как величины изоэлектрической точки характеризуют заряд остатка, мы предположили, что компенсаторные замены могут обуславливать постоянство суммарного заряда остатков в районе контакта спиралей 1 и 2, необходимого для стабилизации их упаковки.

Слайд 64.
Одним из интересных взглядов на эволюцию последовательностей является теория, предложенная в работах Эйгена, Шустера, Кауффмана и др. В них эволюция белка рассматривается как движение по ландшафту приспособленности в многомерном пространстве последовательностей. (Впервые идея ландшафта приспособленности была предложена Райтом). Полиморфизм популяции можно описать в такой модели как облако точек, каждая из которых представляет собой мутантный вариант белка, а все это облаго движется в направлении максимумов на таком ландшафте.

Слайд 65.
Разрешите в заключение упомянуть ряд ссылок на Интернет ресурсы, связанные с молекулярной эволюцией (и не только с молекулярной), которые могут быть интересными.