

Министерство образования Республики Беларусь
Белорусский государственный университет
Факультет биологический
Кафедра общей экологии и методики преподавания биологии

СОГЛАСОВАНО

Заведующий кафедрой

Гричик В.В

«03» января 2025 г.

СОГЛАСОВАНО

Декан факультета

Хрусталёв В.В

«16» января 2025 г.

Молекулярная и биохимическая эволюция живых систем

электронный учебно-методический комплекс
для специальности 1-31 80 23 «Биоинформатика»

Регистрационный № 2.4.2-24 / 601

Автор:

Воронова-Барте Н.В., доцент

Рассмотрено и утверждено на заседании Научно-методического совета БГУ
20.03.2025 г., протокол № 8.

Минск 2025

УДК 575.852(075.8)

В 754

Утверждено на заседании Научно-методического совета БГУ.

Протокол № 8 от 20.03.2025 г.

Решение о депонировании вынес
Совет биологического факультета.

Протокол № 6 от 16.01.2025 г.

А в т о р:

Воронова-Барте Нина Владимировна, доцент, кафедра общей экологии и методики преподавания биологии факультета биологии БГУ

Рецензенты:

Михаленко Е.П., ведущий научный сотрудник лаборатории экологической генетики и биотехнологии государственного научного учреждения «Институт генетики и цитологии НАН Беларуси», кандидат биологических наук;

Филиппова С.Н., доцент кафедры клеточной биологии и биоинженерии растений БГУ, кандидат биологических наук.

Воронова-Барте, Н. В. Молекулярная и биохимическая эволюция живых систем : электронный учебно-методический комплекс для специальности 1-31 80 23 Биоинформатика / Н. В. Воронова-Барте ; БГУ, Биологический фак., Каф. общей экологии и методики преподавания биологии. – Минск : БГУ, 2025. – 38 с. – Библиогр.: с. 38.

Электронный учебно-методический комплекс предназначен для магистрантов специальности 1-31 80 23 «Биоинформатика», включает в себя разделы: теоретический, контроля знаний и вспомогательный, содержащий учебно-программные материалы и список рекомендованной литературы. Курс «Молекулярная и биохимическая эволюция живых систем» предполагает изучение общих вопросов, касающихся особенностей изменчивости биологических последовательностей и методов работы с ними.

СОДЕРЖАНИЕ

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА	4
1. ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ.....	7
1.1. Модели нуклеотидных и аминокислотных замещений.....	7
1.2. Особенности построения филогений	16
1.3. Байесовский метод при построении филогений	20
1.4. Сравнение методов и тестирование достоверности	22
1.5. Теория коалесценции и видовые деревья.....	24
1.6. Молекулярные часы и вычисление времени дивергенции таксонов.....	26
1.7. Нейтральная и адаптивная эволюция белков.....	29
1.8. Симуляция молекулярной эволюции	30
2. ПРАКТИЧЕСКИЙ РАЗДЕЛ	33
2.1. Примерный перечень заданий для управляемой	33
самостоятельной работы магистрантов	33
2.2. Примерная тематика практических занятий.....	33
2.3. Примерная тематика реферативных работ.....	34
3. РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ.....	35
3.1. Вопросы и задания для самоконтроля.....	35
3.2. Вопросы для подготовки к экзамену.....	35
4. ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ.....	38
4.1. Рекомендуемая литература	38
4.2. Электронные ресурсы	38

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Электронный учебно-методический комплекс (ЭУМК) по учебной дисциплине «Молекулярная и биохимическая эволюция живых систем» создан в соответствии с требованиями Положения об учебно-методическом комплексе на уровне высшего образования и предназначен для магистрантов специальности 1-31 80 23 Биоинформатика. Содержание разделов ЭУМК соответствует образовательным стандартам углубленного высшего образования данной специальности. Главная цель ЭУМК – оказание методической помощи магистрантам в систематизации учебного материала в процессе подготовки к итоговой аттестации по курсу «Молекулярная и биохимическая эволюция живых систем».

Структура ЭУМК включает:

Учебно-методическое обеспечение дисциплины

1.1. Теоретический раздел (учебное издание для теоретического изучения дисциплины в объеме, установленном типовым учебным планом по специальности): включает тезисы отдельных лекций, содержащих структурированный теоретический материал по 8 темам дисциплины «Молекулярная и биохимическая эволюция живых систем».

2. Контроль знаний (материалы текущей и итоговой аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к зачету, задания, вопросы для самоконтроля, тематика рефератов и др.).

3. Вспомогательный раздел.

3.1. Учебно-программные материалы (учебная программа для учреждения высшего образования по учебной дисциплине).

3.2. Информационно-аналитические материалы (список рекомендуемой литературы, перечень электронных образовательных ресурсов и их адреса и др.).

Работа с ЭУМК должна включать на первом этапе ознакомление с тематическим планом дисциплины, представленным в учебной программе. С помощью учебной программы по дисциплине можно получить информацию о тематике лекций, перечнях рассматриваемых вопросов и рекомендуемой для их изучения литературы. Для подготовки к промежуточным зачетам необходимо, в первую очередь, использовать материалы, представленные в разделе учебно-методическое обеспечение дисциплины, а также материалы для текущего контроля самостоятельной работы. В ходе подготовки к итоговой аттестации рекомендуется ознакомиться с требованиями к компетенциям по дисциплине, изложенными в учебной программе, структурой рейтинговой системы, а также перечнем вопросов к зачету. Для написания рефератов могут быть использованы информационно-аналитические материалы, указанные в соответствующем разделе ЭУМК.

Цель учебной дисциплины – сформировать у студента глубокое представление об основных теоретических и методологических подходах, применяемых в биоинформатике при изучении молекулярных механизмов эволюции живых систем.

Задачи учебной дисциплины:

- знакомство студентов магистратуры с современными теориями молекулярной эволюции;
- изучение основных методов и алгоритмов оценки эволюционных событий, происходящих в нуклеотидных и аминокислотных последовательностях;
- изучение стратегии адаптации организмов на уровне биомолекул;
- изучение основных методов реконструкции эволюционных событий;
- знакомство магистрантов с основными биоинформационными методами изучения молекулярной и биохимической эволюции, включая прогностические подходы;
- формирование представления о практических областях использования знаний о молекулярной и биохимической эволюции, а также о наиболее современных исследованиях в этой области.

Место учебной дисциплины в системе подготовки специалиста с углубленным высшим образованием

Учебная дисциплина относится к модулю «Клеточные и молекулярные основы функционирования и эволюции живых систем» компонента учреждения образования.

Учебная программа составлена с учетом межпредметных связей и программ по дисциплинам: «Функциональная биология клетки», «Биоинформационные подходы в исследовании белков и метаболитов», «Анализ геномов» и др.

Требования к компетенциям

Освоение учебной дисциплины «Молекулярная и биохимическая эволюция живых систем» должно обеспечить формирование следующей специализированной компетенции:

СК-3. Применять биоинформационные методы эволюционного анализа геномных и протеомных данных, филогенетического анализа нуклеотидных последовательностей и пространственных биомолекулярных структур, а также графического представления биоинформационных данных различного типа.

В результате освоения учебной дисциплины обучающийся должен:

знать:

- современные теории эволюционной динамики, реализуемой на уровне макромолекул;
- особенности и ограничения методов, применяемых для изучения молекулярной эволюции;
- основные статистические подходы, используемые для оценки получаемых результатов в эволюционных исследованиях;
- особенности разворачивания нейтральной и адаптивной эволюции в крупных таксонах живых организмов;
- принципы реконструкции эволюционных событий на основе данных секвенирования ДНК.

уметь:

– актуализировать аналитические задачи, подбирать и использовать релевантные методы биоинформатики для работы с данными в области молекулярной эволюции;

– проводить биологическую интерпретацию результатов, получаемых методами эволюционной биоинформатики.

владеть:

– основными методами эволюционной биоинформатики;

– понятийным и аналитическим аппаратом биоинформатики, необходимыми для работы с данными в эволюционных исследованиях.

Структура учебной дисциплины

Дисциплина изучается во 2 семестре. Всего на изучение учебной дисциплины «Молекулярная и биохимическая эволюция живых систем» отведено:

– для очной формы получения углубленного высшего образования: 102 часа, в том числе 40 аудиторных часа, из них: лекции – 16 часов, практические занятия – 16 часов, управляемая самостоятельная работа – 8 часов, в том числе внеаудиторный контроль УСР – 4 часа, управляемая самостоятельная работа с применением дистанционных образовательных технологий – 4 часа (ДОТ).

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма промежуточной аттестации – экзамен.

1. ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

Курс «Молекулярная и биохимическая эволюция живых систем» предполагает изучение общих вопросов, касающихся эволюции биологических последовательностей и методов работы с ними.

Занятия по курсу «Молекулярная и биохимическая эволюция живых систем» предполагают освоение студентами магистратуры разделов курса, по предложенным тематикам лекций.

1.1. Модели нуклеотидных и аминокислотных замещений

Современная молекулярная биология рассматривает эволюцию макромолекул (ДНК, РНК и белков) как процесс, во многом определяемый случайными мутациями и селективными факторами, действующими на различные участки генома. Многочисленные исследования показывают, что скорость накопления мутаций может существенно варьироваться в зависимости от функциональной значимости гена и его структуры. Ключевые современные представления включают концепцию молекулярного дрейфа и баланса между нейтральными, вредными и полезными мутациями, а также учёт гетерогенности эволюционных процессов в разных геномных регионах.

Быстро эволюционирующие гены часто связаны с адаптацией к меняющимся условиям среды, иммунным ответом, взаимодействием с патогенами или специализированными функциями (например, репродуктивные белки). Высокая скорость эволюции таких генов обусловлена либо сниженным селективным давлением на структуру (если изменения не сильно влияют на жизнеспособность организма), либо, наоборот, повышенным позитивным отбором (когда быстрые изменения дают преимущество в определённых условиях). Медленно эволюционирующие гены чаще всего выполняют жизненно важные, высококонсервативные функции (например, гены репликации ДНК, транскрипции, трансляции, а также «древние» структурные белки). Уровень замены нуклеотидов и аминокислот в таких генах мал, поскольку большинство мутаций может оказываться летальными или существенно снижать приспособленность организма. Таким образом, скорость эволюции гена (или белка) определяется совокупностью функциональной важности, уровнем селективных ограничений и необходимостью адаптации.

В рамках одного и того же генома разные участки (промоторные регионы, регуляторные элементы, интроны, экзоны) эволюционируют с разной скоростью. Кодрующие (экзонные) последовательности, особенно отвечающие за критические структуры и активные центры белков, демонстрируют относительно низкую скорость эволюции из-за сильного очищающего отбора (purifying selection). Напротив, регуляторные области и интроны могут быть подвержены более слабому отбору или даже нейтральному дрейфу, если мутации там незначительно сказываются на приспособленности.

Для описания вероятностей замены одного нуклеотида другим широко используются Марковские модели, предполагающие, что изменение состояния в следующий момент времени зависит только от текущего состояния (однородная цепь Маркова). Простейшая модель JC69 (Jukes-Cantor, 1969) исходит из одинаковой вероятности перехода между любыми нуклеотидами. Более усложнённые модели (K80 – Kimura двухпараметрическая, HKY85 – модель Хасэгавы–Кишино–Яно, F84, TN93 и др.) учитывают различия между транзициями и трансверсиями, а также неодинаковые равновесные частоты нуклеотидов.

JC69 (Jukes-Cantor 1969): предполагает одинаковую вероятность перехода между любыми нуклеотидами. Это самая простая модель с равными равновесными частотами нуклеотидов и без разделения транзиций/трансверсий.

K80 (Kimura двухпараметрическая): вводит различие между транзициями (Ti) и трансверсиями (Tv), поскольку транзиции в природе встречаются чаще. Однако частоты нуклеотидов остаются равными ($A = G = C = T$).

HKY85 (Hasegawa-Kishino-Yano 1985): учитывает различия в скоростях транзиций и трансверсий, а также позволяет неодинаковые равновесные частоты нуклеотидов ($\pi_A, \pi_G, \pi_C, \pi_T$). F84, TN93: дальнейшее развитие идей HKY85 с более гибким учётом соотношения Ti/Tv, равновесных частот и особенностей переходов между конкретными парами нуклеотидов.

TN93 (Tamura-Nei) позволяет отдельно моделировать транзицию пурин→пурин и пиримидин→пиримидин. Чем сложнее модель, тем больше в ней параметров, которые нужно оценить из данных, но тем точнее она описывает реальные эволюционные процессы.

При вычислении генетической дистанции модели корректируют наблюдаемые процентные расхождения с учётом скрытых множественных замен. При больших эволюционных дистанциях реальное число замен может быть существенно выше, чем то, что мы видим напрямую (процент расхождения по сайтам). Когда в одном и том же сайте происходили несколько последовательных замен, часть информации «перезаписывается», и мы можем недооценивать истинную дивергенцию. Это явление называется множественными заменами (multiple hits).

JC69: даёт простую экспоненциальную поправку (если мы наблюдаем $x\%$ различий, модель предполагает, что реальное число замен больше x из-за перекрывающихся замен).

K80: отдельно поправляет число транзиций и трансверсий, учитывая, что они имеют разные скорости и разную вероятность «перекрыться».

HKY85, F84, TN93: ещё более тонко корректируют процент расхождений, вводя поправки и на различающиеся равновесные частоты нуклеотидов, и на разные подтипы транзиций.

Если не вводить корректировку на множественные замены, то при больших эволюционных расстояниях расстояние будет сильно занижено, что искажает филогенетические выводы.

В разных случаях оправданно применение различных моделей нуклеотидных замещений:

JC69 – когда последовательности очень близкие и имеется ограниченное количество данных. Применима для самых простых случаев.

K80 – когда важно разделить транзиции и трансверсии, но равновесные частоты приблизительно равны.

HKY85 – более общий случай, когда неизвестны равновесные частоты нуклеотидов и есть существенная разница T_i/T_v .

F84, TN93 – когда нужны ещё более гибкие параметры эволюции, например, разные скорости внутри типа транзиции ($A \leftrightarrow G$) и ($C \leftrightarrow T$).

GTR (General Time Reversible) – самый общий из «обратимых» (time-reversible) классов. Применяют при максимальной гетерогенности данных.

Отношение транзиций (T_i) к трансверсиям (T_v) является важным параметром в эволюционной генетике, так как транзиции, как правило, встречаются чаще (особенно в кодирующих участках) и могут сохраняться в силу меньшего влияния на структуру ДНК. Транзиции (пурин $\leftrightarrow$ пурин или пиримидин $\leftrightarrow$ пиримидин) часто менее заметно влияют на структуру ДНК, поэтому происходят и/или сохраняются чаще, чем трансверсии (пурин $\leftrightarrow$ пиримидин).

Параметр T_i/T_v используется для подбора оптимальной модели нуклеотидных замен, т.к. помогает точнее учесть характер мутаций. В прикладном смысле оценка T_i/T_v важна в популяционных исследованиях и в клинической генетике, поскольку даёт нам информацию о механизмах мутагенеза и потенциальной функциональной значимости изменений.

Модели часто вводят поправочные коэффициенты для транзиций и трансверсий, что отражает неоднородность эволюционного процесса в разных участках гена. В разных участках гена частота транзиций может отличаться в силу различной функциональной нагрузки и степени селективного давления. Для кодирующих районов третья позиция кодона (часто синонимичная) более склонна к транзициям, в то время как трансверсии там могут сильнее «наказываться» отбором. Известно, что среднее соотношение T_i/T_v для большинства организмов колеблется обычно в диапазоне 2–5, но может отличаться у микроорганизмов, у митохондриальных геномов, в высококонсервативных сегментах и т.д. Из T_i/T_v -соотношения можно сделать выводы о степени селективных ограничений и механизмах мутаций в конкретной филогенетической группе.

Метод максимального правдоподобия (Maximum Likelihood, ML) в филогенетике заключается в поиске набора параметров эволюционной модели, который с наибольшей вероятностью объясняет наблюдаемые нуклеотидные последовательности. Метод Maximum Likelihood (ML) ищет такие значения параметров эволюционной модели (скорости транзиции/трансверсии, равновесные частоты нуклеотидов, уровень гетерогенности сайтов и т.д.), при которых вероятность (правдоподобие) получить конкретные наблюдаемые последовательности максимальна.

Основная формула метода ML:

$$L(\theta) = P(\text{Данные} / \theta),$$

где θ – вектор параметров (скорости мутаций, распределение скоростей по сайтам, топология дерева), P – вероятность.

ML-оценка параметра θ , это такая точка, где $L(\theta)$ достигает максимума.

Поиск параметров в общем смысле происходит следующим образом:

1. Задаём начальные значения параметров (часто эвристические).
2. Вычисляем правдоподобие для данных (выравнивание последовательностей) с учётом текущих параметров.
3. Меняем параметры (например, методом градиентного подъёма, ЕМ-алгоритмом, или другими численными методами).
4. Повторяем, пока не достигнем (локального) максимума функции правдоподобия.

При сравнении нескольких моделей (например, JC69, HKY85, GTR и др.) используют *информационные критерии* (AIC, BIC и пр.), позволяющие выбрать наилучшую модель с учётом точности и количества свободных параметров.

AIC (Akaike Information Criterion):

$$AIC = 2k - 2\ln(L),$$

где k – число свободных параметров, L – максимизированное правдоподобие.

При этом, чем меньше значение AIC, тем лучше модель описывает реальную ситуацию в выборке последовательностей.

BIC (Bayesian Information Criterion):

$$BIC = \ln(N) \cdot k - 2\ln(L),$$

где N – размер выборки (число позиций или последовательностей).

BIC сильнее штрафует за большое число параметров, чем AIC, т.е. является более строгим.

Выбор критерия зависит от размера выборки, предполагаемой сложности модели и предпочтений исследователя (AIC более «либерален» по отношению к количеству параметров, BIC обычно более «консервативен»).

Оптимизация моделей по критерию ML важна для адекватной реконструкции филогенетического дерева и точного вычисления генетических дистанций. «Оптимизировать модель по ML» – значит найти такие параметры (и, возможно, топологию дерева), при которых функция правдоподобия достигает максимума, а затем выбрать модель (JC69, HKY85, GTR и т.д.), дающую наилучший баланс между подгонкой данных и числом параметров (по AIC/BIC).

Марковские цепи лежат в основе большинства моделей нуклеотидных замещений, поскольку позволяют формально описать вероятности переходов от одного состояния (нуклеотида) к другому во времени. Матрица переходных

вероятностей в однородной цепи Маркова остаётся постоянной, и каждое новое состояние зависит лишь от текущего.

В однородной цепи Маркова считается, что вероятность перехода из состояния i в состояние j (где состояния – это нуклеотиды) определяется матрицей скоростей (или матрицей перехода за бесконечно малое время) Q . За конечный интервал времени t реальная матрица переходных вероятностей $P(t)$ получается путём экспоненцирования:

$$P(t) = \exp(Qt).$$

Строго говоря, Q – это матрица, в которой на диагональных элементах стоят отрицательные суммы скоростей исхода из состояния i , а вне диагонали – положительные скорости перехода q_{ij} .

Преимущества использования Марковских цепей:

1. Позволяют формализовать вероятностный характер замен: каждое новое состояние зависит только от текущего (нет необходимости учитывать более сложные «исторические» эффекты);

2. Математически элегантны: существуют хорошо изученные методы экспоненцирования матриц, вычисления стационарных распределений, правдоподобий при заданных деревьях;

3. Применимы во всех распространённых моделях нуклеотидных замен (JC69, K80, HKY85, TN93, GTR и др.), где важна непрерывная модель временных переходов с учётом скорости эволюции.

UNREST (unequal transition rates) и GTR (General Time Reversible) – это более общие, «расслабленные» модели эволюции нуклеотидов, которые учитывают различные скорости замен между всеми парами нуклеотидов и, в случае GTR, удовлетворяют условию обратимости во времени (time-reversibility).

«Расслабленная» (relaxed) – значит, что модель не навязывает жёстких ограничений (например, равенства скоростей замены $A \rightarrow G$ и $G \rightarrow A$, или одинаковых равновесных частот нуклеотидов). Модель позволяет независимые скорости для каждой пары ($i \rightarrow j$).

Обратимость во времени (time-reversibility) предполагает, что эволюционный процесс выглядит так же, если «запустить его назад», при условии достижения стационарного распределения (т.е. $\pi_i q_{ij} = \pi_j q_{ji}$).

Параметры таких моделей подбираются статистически, чтобы максимизировать правдоподобие наблюдаемых данных. Подбор параметров осуществляется методами ML или байесовскими методами: алгоритм ищет набор скоростей q_{ij} и равновесных частот π_i , которые максимизируют правдоподобие данных. При этом обратимость накладывает связь между q_{ij} и q_{ji} .

Расчёт генетических дистанций по таким моделям сложнее, но даёт более точное описание процесса дивергенции. Учёт индивидуальных скоростей замен между любыми парами нуклеотидов более реалистичен, ведь биохимические и структурные ограничения разные для каждой пары ($A \rightarrow G$ может отличаться от $G \rightarrow A$ и т.п.). Это сводит к минимуму систематические ошибки, возникающие при попытке «усреднить» скорость замен.

При построении эволюционных моделей на основе попарных сравнений последовательностей возникает риск недооценить сложность многократных замен, а также игнорировать гомоплазию (сходство, возникшее независимо у разных линий). Кроме того, корректировка большого количества парных расстояний может приводить к статистическим погрешностям и усложнять оценку достоверности результатов. При большом числе попарных сравнений вероятность накопления ложных совпадений растёт (ошибки множественного тестирования). Игнорирование гомоплазии (независимого появления схожих замен в разных ветвях) может исказить оценку истинного эволюционного расстояния. При сильной дивергенции между ветвями попарная оценка может давать «недооценку» числа замен (множественные замены в одном сайте). Эти ограничения порождают необходимость в более комплексных методах анализа многопоследовательственного выравнивания и филогенетических реконструкций. Современные методы (ML, Байесовские, парсимония и т. д.) обрабатывают сразу весь набор последовательностей, учитывая конфликты и сигналы во всех позициях и у всех таксонов. Это позволяет лучше учесть гомоплазию и корректировать пропуски данных, повышая надёжность результата по сравнению с простым суммированием попарных расстояний.

Для анализа белковых последовательностей разрабатываются специальные матрицы замен (например, PAM, BLOSUM), основанные на обширных эмпирических данных о частотах замещений аминокислот.

Эмпирические матрицы (PAM, BLOSUM):

PAM (Point Accepted Mutation) основана на выравнивании близкородственных белков и подсчёте частот замен аминокислот. Существуют разные версии (PAM1, PAM250 и др.), различающиеся «эволюционной глубиной».

BLOSUM (BLOcks SUBstitution Matrix): строится на основе консервативных блоков белков и делится по уровням идентичности (BLOSUM62, BLOSUM80 и т.д.).

Движение по матрице замен (например, BLOSUM62) означает, что если мы видим замену аминокислоты X на Y, то оценивается вероятность/«скор» этой замены. В матрице каждая клетка (X,Y) отражает, насколько часто или редко $X \rightarrow Y$ встречается в эволюционно родственных белках. Т.о. эти матрицы учитывают неодинаковые вероятности замен между разными аминокислотами в зависимости от их химических и физических свойств.

Частота и «степень допускаемости» замен зависят от:

1. полярности (заряженные, гидрофобные, полярные незаряженные);
2. размеров боковых цепей (большие объёмные группы vs. мелкие, как Gly);
3. ароматичности;
4. возможности образовывать водородные связи, дисульфидные мостики;
5. участия в каталитическом центре и т.д.

Механистические модели (например, структурно-информированные) пытаются учесть детальные биохимические особенности, включая функциональные ограничения и взаимодействия аминокислот в белках.

Структурно-информированные модели учитывают реальное положение аминокислоты в пространстве, её контакты с соседями, важность для структуры белка. Модели вроде LG, WAG, JTT – это расширенные эмпирические матрицы, построенные на больших наборах реальных белковых выравниваний, но с учётом элементов механистического понимания (консервация семей белков, функциональные мотивы). Преимущество механистических моделей – более точный учёт биохимических реалий, однако они сложнее в параметризации и требуют данных о структуре белка или функциях определённых сайтов.

Гетерогенность между сайтами (site heterogeneity) отражает факт, что различные позиции в последовательности эволюционируют с разной скоростью из-за различной функциональной важности. Например, сайты, отвечающие за активный центр фермента, могут быть высококонсервативными, а петли на поверхности белка более вариабельны. В моделях часто используют гамма-распределение для описания распределения скоростей эволюции по сайтам, что повышает точность оценки генетических расстояний и филогенетических связей.

Гамма-распределение позволяет описать разброс скоростей эволюции по сайтам. Оно задаётся двумя параметрами: α (shape) и β (rate) или иногда α и средняя скорость.

$$f(x;\alpha,\beta) = \beta^\alpha / \Gamma(\alpha) * x^{\alpha-1} e^{-\beta x},$$

где x – скорость эволюции сайта, $\Gamma(\alpha)$ – Гамма-функция.

Без учёта гетерогенности замен любая модель подразумевает одну «среднюю» скорость замен для всех позиций. Но функционально важные сайты эволюционируют медленно, а менее критичные – быстрее. Гамма-распределение даёт возможность взвешенно корректировать вклад разных сайтов и в итоге точнее восстанавливать истории замен и оценивать ветви филогенетического дерева.

Для упрощённого описания эволюции белков нередко применяют Пуассоновскую модель, которая предполагает равную вероятность замены каждой аминокислоты и независимость событий замены. Пуассоновская модель упрощённо предполагает, что каждая аминокислота заменяется с постоянной средней скоростью, независимо от текущего состояния. Фактически это тоже «марковский» процесс, но с жёстким допущением: все аминокислоты имеют одинаковую вероятность перехода друг в друга, и нет структуры «4×4» или «20×20» с разными скоростями. Это упрощение подходит, если эволюция равномерна и нет сильных различий между типами замен. Однако эта модель не учитывает гетерогенность сайтов, поэтому более точными считаются подходы с гамма-поправкой (Gamma distance), где учитывается вариабельность скоростей по разным позициям. Если мы хотим учесть вариабельность скоростей по сайтам, накладываем на Пуассоновскую модель гамма-поправку (т.е. вводим

распределение скоростей для разных сайтов). Информация о вариабельности (параметр α) получается либо апостериорно при подгонке модели, либо путём отдельных оценок (например, ML-оценки из данных). Это даёт более реалистичную оценку эволюционного расстояния между белками.

Кодонные модели рассматривают эволюцию на уровне триплетов нуклеотидов, позволяя различать синонимичные и несинонимичные замены. Такие модели (например, модель Goldman–Yang, Muse–Gaut) учитывают вероятность возникновения мутаций в каждом из трёх положений кодона, а также влияние синонимичных замен, обычно менее подверженных селективному давлению. Кодонная модель рассматривает триплет нуклеотидов как единое состояние, позволяя напрямую различать:

1. синонимичные замены, не меняющие аминокислоту (чаще менее подвержены отбору);

2. несинонимичные замены, меняющие аминокислоту (часто подлежащие более сильному отбору). 1

Yang & Nielsen (YN), GY94 (Goldman–Yang 1994), MG94 (Muse–Gaut 1994) – классические кодонные модели, различаются в том, как учитывают равновесные частоты кодонов и ставки (рэйты) мутаций.

Choice of model зависит от размеров и качества выборки, предполагаемого уровня отбора, наличия функционально важных сайтов.

Плюсы кодонных моделей состоят в том, что они дают возможность напрямую оценивать dN и dS . Минусы этих моделей состоят в том, что они нуждаются в большем числе параметров, им нужны большие наборы данных и вычислительные ресурсы. Кодонные модели широко используются для изучения процессов адаптивной эволюции и оценки селективных эффектов на уровне белковых кодирующих генов.

В контексте кодирующих последовательностей часто оценивают скорости синонимичных (dS) и несинонимичных (dN) замен. Существуют различные подходы к вычислению этих показателей: от более простых парсимонийных методов до сложных ML-методов, учитывающих неоднородности эволюции и неодинаковые равновесные частоты нуклеотидов.

Основные методы вычисления, это:

1. метод Nei–Gojobori – относительно простой метод подсчёта доли синонимичных и несинонимичных сайтов и замещений;

2. метод Li – более точный учёт переходов и трансверсий;

3. метод Goldman–Yang (ML-подход) учитывает неодинаковые равновесные частоты, гамма распределение по сайтам и т.д.

ML-оценки часто считаются более достоверными, но они требуют большого количества компьютерных ресурсов.

Показатель dN/dS , равный отношению скорости несинонимичных замен к синонимичным, служит индикатором направления отбора. Значение $dN/dS < 1$ указывает на действие отрицательного (очищающего) отбора, сдерживающего несинонимичные мутации, тогда как $dN/dS > 1$ обычно свидетельствует о позитивном (движущем) отборе. Значение, близкое к 1, может указывать на

нейтральную эволюцию, при которой изменения не оказывают существенного влияния на приспособленность организма.

В сравнительной геномике анализ dN/dS позволяет выявлять гены, подвергшиеся ускоренной адаптивной эволюции в отдельных линиях (например, при адаптации к новым условиям или патогенам). Метод широко используется для поиска сигнатур отбора (характерных паттернов в геноме, указывающих на недавнее или текущее действие естественного отбора, например, резкое изменение частоты определённого аллеля), исследования эволюционных “горячих точек” (hotspots – регионы ДНК/белка, где замены происходят чаще, чем в среднем, и где может идти активная адаптация, например, иммунные гены, гены репродукции, рецепторы и т.д.) и функционально значимых мутаций.

В сочетании с филогенетическими методами вычисления dN/dS даёт интегральное представление об адаптивном потенциале определённых групп генов.

В заключение поясним ключевой момент про матрицу переходов в моделях нуклеотидных/кодонных замен. Матрица переходных вероятностей описывает вероятность перехода от одного состояния (нуклеотида или кодона) к другому за фиксированный эволюционный интервал. Для моделей нуклеотидных замен эта матрица обычно параметризуется скоростями α для транзиций и β для трансверсий, либо более сложными параметрами при учёте неодинаковых равновесных частот.

Например.

1. Имеется матрица скоростей Q размерности 4×4 (для нуклеотидов) или 61×61 (для кодонов без стоп-кодонов).

2. Элемент q_{ij} отражает скорость замены состояния i на j (часто зависящую от типа мутации и равновесных частот). На диагонали Q стоит $-\sum_{j \neq i} q_{ij}$.

3. Чтобы найти вероятность перехода из i в j за время t , необходимо вычислить $P(t) = \exp(Qt)$. Это матричная экспонента:

$$\exp(Qt) = \sum (Q^k t^k) / k!.$$

4. На практике используют численные методы (диагонализацию, серии Паде и т.д.), встроенные в программы филогенетического анализа.

5. Полученная $P(t)$ даёт вероятность каждого возможного перехода ($i \rightarrow j$) за определённый эволюционный интервал. Затем эта информация суммируется по дереву (например, методом Felsenstein) для вычисления общего правдоподобия выравнивания.

Преимущество данного подхода состоит в том, что мы получаем формально обоснованные вероятности всех возможных историй замен в каждой позиции, что позволяет проводить точный статистический и филогенетический анализ.

Таким образом, современные методы изучения эволюции на уровне ДНК и белков основаны на тонко настроенных моделях замен (Markov chain models, кодонных моделях), учёте множественных замен и гетерогенности скоростей замещений (гамма-распределение), корректной оптимизации параметров (ML, байесовские подходы) с использованием информационных критериев (AIC, BIC),

расчёте генетических дистанций и оценке dN/dS для выявления сигнатур отбора и понимания молекулярных механизмов эволюции.

Каждый из инструментов (JC69, HKY85, GTR, пуассоновская и кодонные модели) имеет свою область применения и уровень сложности. Грамотно выбранная модель и корректные статистические методы обеспечивают наиболее точную реконструкцию эволюционной истории и позволяют делать достоверные выводы о природе отборных процессов, скорости изменений и филогенетических взаимоотношениях таксонов.

Генетические дистанции отражают меру эволюционной дивергенции между последовательностями. Малая дистанция может свидетельствовать об общей недавней эволюционной истории или высокой степени консервативности, тогда как большая дистанция указывает на более глубокое расхождение линий или интенсивный отбор. В биологическом контексте эти показатели позволяют судить о родственных связях между организмами, реконструировать филогенетическое дерево и выявлять участки, подвергшиеся ускоренному или замедленному эволюционному процессу.

1.2. Особенности построения филогений

Различают две основные разновидности филогенетических деревьев:

1. Укоренённые (*rooted*) деревья.

Имеют условный или реальный «корень», от которого последовательно «ветвятся» все линии. Корень чаще всего определяется, исходя из наличия внешней группы (намеренно включенные в выборку последовательности, заведомо не входящие в анализируемую группу). Укорененные деревья отражают предполагаемое направление эволюции во времени: от корневого предка к современным таксонам.

2. Неукоренённые (*unrooted*) деревья.

Показывают только топологию (кто ближе к кому), но не указывают направление эволюционного движения. Используются там, где сложно или невозможно подобрать аутгруппу; часто встречаются в алгоритмах первичного построения. Все звездчатые или неразрешенные деревья относятся к неукорененным.

1. Видовые (*species tree*) деревья.

Видовое дерево моделирует историю дивергенции видов и предполагает единую эволюционную линию для каждого вида.

2. Генное дерево (*gene tree*).

Генное дерево иллюстрирует эволюцию конкретного гена/локуса, которая может расходиться с историей вида из-за неполной сортировки родительских линий, горизонтального переноса генов, гибридизации и прочих факторов.

Неполная сортировка родительских генетических линий (Incomplete Lineage Sorting, ILS), это специфическое явление, возникающее при построении филогенетических деревьев. Когда близкородственные виды расходятся в процессе видообразования, гены, имеющие несколько копий в геноме, могут распределяться между дочерними линиями случайным образом. Если время

между последовательными событиями видообразования мало, некоторые варианты генов «не успевают» фиксироваться и происходит, своего рода, перемешивание аллельных вариантов генов. В результате генное дерево, построенное по такому участку генома, может не совпадать с истинной историей видов. Это сильно затрудняет интерпретацию получаемых филогений и требует использование комплексных подходов (многолокусный анализ, коалесцентные методы и т.д.).

Основные методы построения филогенетических деревьев делятся на:

1. Дистанционные (*Distance-Based*) методы

В этих методах сначала происходит расчет матрицы попарных расстояний между последовательностями (процент расхождений, модельно-корректированные расстояния и т. д.). Далее дерево строится так, чтобы лучше всего отражать эти расстояния (минимизация суммарной ошибки), а суммарная дистанция по всем ветвям была минимальной.

К таким методам относятся, например,

Neighbor Joining (NJ). Этот метод итеративно объединяет пары таксонов, имеющих минимальное «расстояние», формируя ветвление на дереве. NJ работает быстро даже для больших наборов данных, но может быть чувствителен к выбору метода расчета дистанции и эффектам типа «притяжения длинных ветвей».

Fitch–Margoliash (метод минимальных квадратов). Стремится минимизировать разницу между эмпирическими расстояниями и квадратами длин ветвей получаемого дерева. Более вычислительно сложен, чем NJ.

Minimum Evolution. Ищет дерево с минимальной суммарной длиной всех ветвей при условии заданной матрицы расстояний. Является наиболее быстрым.

К преимуществам дистанционных методов относятся:

1. Простота реализации, высокая скорость.
2. Подходящи для больших наборов данных, когда другие методы слишком ресурсоемки.
3. Известны также ограничения методов, основанных на расчете дистанций:
4. Сжатие всей информации о последовательностях в единственную матрицу расстояний может привести к потерям подробных сигналов (например, при гомоплазии, неоднородности скоростей замещений по сайтам и т.д.).
5. Подвержены систематическим ошибкам, если модель расстояний выбрана неправильно (особенно при больших эволюционных дистанциях).

2. Дискретные методы

К дискретным методам относятся, например.

Парсимония (Parsimony). Основная идея этого метода, это минимизация общего числа эволюционных изменений (мутаций), необходимых для объяснения наблюдаемых различий между последовательностями. Взвешенная парсимония придает разный «вес» разным типам изменений (например, транзиции могут быть «дешевле», чем трансверсии).

Преимущества этого метода состоят в том, что он подходит и для морфологических, и для молекулярных данных. Часто интуитивно понятен для интерпретации (базируется на принципе «Бритвы Оккама»).

К недостаткам метода относится его чувствительность к эффекту «притяжения длинных ветвей». Кроме того, метод не учитывает вероятностные аспекты: одинаковый «вес» длинного набора конвергентных событий и короткого пути.

Максимальное правдоподобие (Maximum Likelihood, ML). В этом методе происходит последовательный анализ сайтов с моделированием вероятности наблюдаемых символов (нуклеотидов или аминокислот) на дереве с заданной эволюционной моделью (матрица скоростей, распределение скоростей по сайтам и т.п.). ML рассматривается как дискретный метод, так как учитывает конкретные состояния сайтов и не сводится к анализу дистанций.

3. Байесовские методы

Байесовские методы используют априорную информацию (например, предположения о распределении длин ветвей) и получают апостериорное распределение топологий дерева и параметров модели (скоростей, частот и т. д.) на основе применения алгоритмов МСМС (цепей Маркова Монте-Карло).

Особенности байесовских методов состоят в том, что они дают не одно «лучшее дерево», а целое распределение возможных деревьев, позволяя оценить неопределённость топологий. Методы обходят проблемы локальных максимумов за счёт стохастического блуждания в пространстве деревьев.

Недостатки этих методов также существенны. А именно, высокая вычислительная стоимость, особенно при большом числе таксонов и сложных моделях. Метод требует настройки гиперпараметров (априорных распределений, «горячих/холодных» цепей и т. д.), что может влиять на результат.

Исчерпывающие и эвристические методы поиска лучшей топологии

При наличии нескольких альтернативных деревьев возникает вопрос выбора наилучшего дерева, т.е. дерева максимально близко к реальности отображающего эволюционные события в группе с учетом выбранной эволюционной модели. При этом выделяют две рупные группы методов – исчерпывающие и эвристические.

Исчерпывающий (exhaustive) перебор, это подход, при котором строятся и перебираются все возможные топологии. Для каждого дерева вычисляют «стоимость» (длину, правдоподобие) и выбирают лучшую. Сложность этого подхода заключается в так называемом комбинаторном взрыве: число возможных деревьев растёт факториально с числом таксонов и быстро становится непрактичным, значительно увеличивая требуемое машинное время и мощность.

Эвристические алгоритмы (hill climbing, branch swapping, nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR)).

При этом работа начинается с некоторой не оптимальной топологии, затем дерево локально модифицируется («обмен ветвей» или перестановка поддеревьев), после чего оценивается, улучшилась ли метрика (правдоподобие, длина и т.п.).

Локальные (ложные) максимумы, это точки в пространстве топологий, где дальше улучшение «по месту» невозможно, поскольку каждое следующее дерево имеет худшую оценку, чем исходное. Однако локальный максимум не обязательно является глобальным максимумом, т.е. при движении по пространству иным путем глобальная оценка (метрика) могла бы быть лучше. Это

связано с тем, что движение в пространстве деревьев во многом определяется случайными причинами.

Чтобы «выйти» из локального максимума, часто используют различные случайные перестановки (random restarts), более крупные перестановки ветвей или комбинируют результаты нескольких независимых поисков. При этом, например, одна или несколько ветвей «обрезаются» и заново «прививаются» к другому месту на дереве, чтобы найти потенциально более оптимальную топологию.

Стохастические алгоритмы (МСМС)

Метод Монте-Карло строит цепь Маркова, которая со временем сходится к стационарному распределению, совпадающему с апостериорным.

Процесс МСМС включает:

1. Определение целевого распределения (апостериорного).
2. Генерацию последовательности случайных выборок, соответствующих этому распределению.
3. Отбрасывание начальных значений (burn-in) для уменьшения влияния выбора начальной точки на схождение цепи.
4. Оценку статистик на основе оставшихся выборок.

Принцип работы МСМС при построении филогенетических деревьев заключается в следующем:

1. Фиксируется вероятностная модель эволюции (распределение для длин ветвей, скорости замен, равновесные частоты и т. д.).
2. Запускается цепь Маркова, которая случайным образом меняет топологию дерева и значения параметров (длину ветвей, вариации скоростей эволюции, гамма-распределение и т. п.).
3. Новое состояние принимается с вероятностью, зависящей от отношения апостериорных вероятностей (аналог алгоритма Метрополиса–Гастингса).
4. Постепенно цепь «собирает» выборку из распределения возможных деревьев, давая и среднюю топологию, и оценку неопределённости (posterior probability).

В алгоритме МСМС *гиперпараметрами* (hyperparameters) называются, например, параметры априорных распределений (например, среднее и дисперсия длины ветвей), предлагаемые шаги при перестройке дерева (proposal distribution), количество итераций, «прогорание» (burn-in) цепи, частота выборки (sampling frequency) и т.д. От корректной настройки гиперпараметров зависит сходимость, скорость и стабильность работы марковской цепи.

При построении филогений особенно важно обеспечить высокое качество входных данных, поскольку особенности входных данных существенно влияют на получаемый результат.

Важнейшим параметром качества данных является качественное *выравнивание*. Ошибочное выравнивание может исказить истинную сигнатуру замен, поэтому желательно использовать проверенные программы (MAFFT, MUSCLE, Clustal Omega, etc.) и вручную корректировать проблемные области.

Часто требуется удаление участков с неоднозначными выравниваниями (gappy regions) или псевдогенами, исключение сильно укороченных или явно «шумных» последовательностей.

Филогенетические деревья на основе аминокислотных последовательностей

Часто в работе более предпочтительным является использование аминокислотных, а не нуклеотидных деревьев. АК деревья свободны от влияния синонимичных замен, устойчивы к «шуму» в третьей позиции кодона. АК деревья более предпочтительны при работе с высокодивергированными и изменчивыми последовательностями белок-кодирующих генов.

Для построения АК филогений используются дистанционные методы, которые аналогичны нуклеотидным, но расстояние рассчитывается на основе матриц аминокислотных замен (PAM, BLOSUM и др. матриц).

ML/Байесовские: применяются механистические или эмпирические матрицы замен (JTT, WAG, LG и т.д.) для расчёта вероятностей перехода одной аминокислоты в другую.

Партишенинг (Partitioning) композитных последовательностей

При работе с несколькими участками генома все последовательности можно объединить в одну выборку, не строя несколько отдельных филогений. Для этого несколько генов (или участков) объединяются в одну композитную последовательность, и общая выборка составляется из композитных последовательностей, в которых части соединены по принципу «голова к хвосту».

Поскольку характер эволюции разных последовательностей различается, эволюционные параметры выборки (скорости, частоты и т.п.) оцениваются отдельно по «разделам», т.е. по каждому участку выборки в отдельности (partition). Такую возможность предоставляют программы PartitionFinder, ModelFinder (в составе IQ-TREE), RAxML с опцией partitioning. Они автоматически подбирают оптимальное разбиение и модель эволюции для каждой партии (гена) или даже для каждого кодонного положения.

1.3. Байесовский метод при построении филогений

Теорема Байеса и её применение в биологии

Теорема Байеса является фундаментальным принципом вероятностного вывода, позволяющим пересчитывать вероятность гипотез с учетом новых данных. Она формулируется следующим образом:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

где:

$P(H|D)$ – апостериорная вероятность гипотезы H после получения данных D ;

$P(D|H)$ – правдоподобие данных при условии истинности гипотезы;

$P(H)$ – априорная вероятность гипотезы;

$P(D)$ – нормализующий множитель (маргинальное правдоподобие данных).

В биологии теорема Байеса широко применяется в анализе последовательностей ДНК, филогенетике, предсказании структуры белков, анализе экспрессии генов и других областях.

Байесовские методы используются для построения эволюционных деревьев с учетом вероятностных моделей мутаций. Метод МСМС позволяет семплировать вероятные деревья на основе последовательностей ДНК или белков, оценивая их правдоподобие и выявляя наиболее вероятные филогенетические связи.

Также Байесовские методы применяются для предсказания мотивов, гомологичных последовательностей и сайтов связывания. Например, при предсказании сайтов связывания ДНК-белков можно использовать апостериорное распределение вероятности наличия определенного сайта в последовательности.

В анализе экспрессии генов Байесовские методы позволяют оценить уровень активности генов с учетом шума и технических ошибок. Например, можно моделировать вероятностное распределение экспрессии на основе RNA-Seq данных и учитывать априорную информацию о функциональных группах генов.

Моделирование сложных взаимодействий между биологическими процессами возможно с использованием Байесовских сетей. Они позволяют оценивать вероятностные связи между переменными, такими как генные взаимодействия, регуляция экспрессии и метаболические пути.

Байесовские методы используются для классификации биомедицинских данных, диагностики заболеваний и анализа многомерных данных. Например, Байесовские классификаторы могут предсказывать вероятность наличия заболевания на основе совокупности биомаркеров.

Байесовская вероятность и её принципиальные отличия

Байесовская вероятность интерпретируется как мера уверенности в гипотезе, тогда как классическая (частотная) вероятность определяется на основе повторяемых экспериментов.

Использование априорного знания – Байесовский подход – позволяет включать в модель предварительные знания о параметрах модели. Это означает, что вместо того, чтобы рассматривать вероятность как объективную частоту события, мы можем использовать накопленную информацию и экспертные оценки для формирования начальных предположений. Например, в филогенетике априорные знания могут включать информацию о вероятности конкретных мутаций или структуре родственных связей между видами.

Понятие «обновление вероятностей» используется для описания процесса, в котором вероятности корректируются по мере получения новых данных. Это ключевой аспект Байесовского подхода, который позволяет динамически адаптировать модели. Например, при анализе экспрессии генов новая экспериментальная информация может уточнять вероятностное распределение активности генов в изменяющихся условиях.

Байесовские методы также позволяют проводить «оценку неопределенности» в параметрах модели. Вместо определения единственного оптимального значения параметров, Байесовский анализ оценивает полное апостериорное распределение,

что позволяет лучше понимать возможные вариации модели и учитывать их в дальнейших расчетах.

Марковские цепи Монте-Карло (МСМС)

МСМС — это метод, позволяющий приближенно вычислять апостериорное распределение параметров. Он используется в сложных моделях, где прямой расчет вероятностей невозможен, например, при построении филогенетических деревьев.

Метод Метрополиса-Гастингса

Этот алгоритм применяется в МСМС для выбора новых состояний модели с учетом их вероятности. В биоинформатике он помогает оптимизировать деревья, корректировать параметры мутационных моделей и искать наилучшие топологии.

Семплирование по Гиббсу

Это пошаговое обновление параметров на основе условных распределений. Применяется в анализе экспрессии генов и моделировании Байесовских сетей для оценки вероятностей взаимодействий.

Анализ апостериорных вероятностей

Этот метод позволяет находить наилучшие модели на основе данных. В филогенетике апостериорные вероятности помогают выбрать наиболее вероятное эволюционное дерево из множества возможных вариантов.

Оценка маргинального правдоподобия

Используется для сравнения моделей и выбора наилучшей гипотезы. Например, при анализе альтернативных эволюционных гипотез можно рассчитать их маргинальное правдоподобие и выбрать наиболее вероятную.

Благодаря этим методам Байесовский подход обеспечивает надежные и интерпретируемые результаты в биологических исследованиях, позволяя учитывать как новые данные, так и накопленные знания.

1.4. Сравнение методов и тестирование достоверности филогенетических деревьев

Филогенетические методы (максимальная парсимония, максимальное правдоподобие, байесовский подход) часто дают различные топологии для одного набора данных. Статистический анализ помогает понять, насколько эти различия значимы и какой метод лучше описывает данные.

Статистический анализ филогенетических деревьев включает в себя несколько ключевых аспектов:

1. Оценка надежности узлов — используются методы бутстрэп-анализа и байесовского вывода для определения уровня поддержки каждого узла в дереве.

2. Сравнение различных методов построения деревьев — включает методы на основе максимального правдоподобия (ML), байесовских методов и метода наименьших квадратов (Least Squares).

Оценка длины ветвей используется для измерения эволюционных расстояний между видами или последовательностями. Длина дерева (в парсимонии), это число шагов (эволюционных изменений), необходимых для объяснения данных.

Минимизация длины получаемых деревьев, это цель парсимонии.

Для оценки качества деревьев применяются различные критерии:

Правдоподобие (Likelihood) – вероятность получения наблюдаемых данных при данной модели эволюции.

Согласованность (Consistency) – способность метода давать правильное дерево при увеличении объема данных.

Эффективность (Efficiency) – вычислительная сложность метода, определяющая скорость и точность построения деревьев.

Надежность (Robustness) – устойчивость метода к неточностям и неполноте данных, особенно в условиях шумовых данных.

Проверка достоверности топологии включает несколько ключевых методов:

Перестановочные тесты (Permutation Tests): Оценка того, насколько вероятность наблюдаемых данных превышает вероятность случайных данных.

Сравнение топологий методом Shimodaira-Hasegawa (SH-test): Определение, значительно ли различаются вероятности различных деревьев.

Методы реконструкции консенсусных деревьев, такие как Majority Rule Consensus и Strict Consensus Trees.

Бутстрэп-анализ оценивает устойчивость узлов дерева.

Алгоритм метода:

1. Генерируются случайные выборки из исходного набора данных, т.е. создаются псевдореплики путем случайной перевыборки сайтов с заменой (например, 1000 раз). Для каждой псевдореплики строится дерево.

2. Подсчитывается частота появления каждого узла (в процентах).

3. Поддержка > 70% считается умеренной, > 95% – сильной.

Следует иметь в виду, что Бутстрэп переоценивает поддержку при систематической ошибке (например, неверной модели эволюции).

Следует обратить внимание, что в последние десятилетия в филогенетических исследованиях чаще всего используются подходы, основанные на моделях (в частности, ML и BI). Методы парсимонии все реже используются для анализа больших выборок, т.к. он склонен к неправильной интерпретации гомопластичных состояний признаков как гомологичных, что приводит к неправильной группировке таксонов с увеличивающейся поддержкой по мере включения дополнительных данных, свойство, называемое несогласованностью.

Эта проблема наиболее распространена в так называемой «зоне Фельзенштейна», области теоретического пространства дерева, в которой расходящиеся несестринские таксоны с параллельными преобразованиями состояний признаков соединяются притяжением длинных ветвей.

В целом, филогенетические методы на основе моделей можно разделить на подходы, основанные на конкатенации (или суперматрице) и коалесценции.

Анализ на основе конкатенации:

Обычно используемое программное обеспечение для построения деревьев методом ML включает garli, raxml, iq-tree и examl. raxml и iqtree относительно удобны для пользователя, на их веб-страницах доступны хорошо аннотированные руководства. iqtree, в частности, чрезвычайно хорошо документирован и включает в себя множество учебных пособий, доступных на веб-сайте разработчика. exml и

garli, в свою очередь, требуют более глубокого знакомства с основами написания скриптов и работы в средах командной строки. exml хорошо подходит для анализа очень больших наборов данных, в то время как garli примечателен тем, что включает в себя широкий спектр моделей эволюции кодонов.

Наиболее известным программным обеспечением для ВІ являются mrbayes, exabayes, phylobayes, revbayes, beast и beast2.

Mrbayes, это хорошо зарекомендовавшая себя программа с очень полезной документацией и множеством доступных онлайн-руководств. phylobayes реализует модель САТ, модель сайт-гетерогенной смеси, разработанную для учета сайт-специфичных особенностей эволюции белка. revbayes, в свою очередь, обеспечивает больший контроль пользователя над отдельными параметрами.

beast и beast2 в основном используются для совместной оценки дерева видов и времени расхождения и обсуждаются более подробно далее. Из всех этих программ наиболее подходящими для больших филогеномных наборов данных являются revbayes и exabayes, которые подходят для массового распараллеливания и многопоточности на компьютерных кластерах.

1.5. Теория коалесценции и видовые деревья

Теория коалесценции, это математическая модель в популяционной генетике, которая описывает, как гены в популяции или между видами связаны через общих предков. В контексте построения филогенетических деревьев теория коалесценции используется для моделирования генеалогии генов, учитывая случайные процессы, такие как генетический дрейф, при оценке эволюционных отношений. Иными словами, теория коалесценции отслеживает линии генов назад во времени, пока не будет найден их общий предок (коалесцентное событие). Это позволяет построить генеалогическое древо, отражающее историю расхождения генов.

Теория коалесценции учитывает не только видообразование, но и внутривидовые процессы, такие как неполное расхождение линий (incomplete lineage sorting, ILS). Это особенно важно для близкородственных видов, где гены могут иметь разные эволюционные истории. Методы коалесценции строят отдельные деревья для каждого гена и интегрируют их в общее видовое древо. Они лучше, чем методы конкатенации, подходят для близкородственных видов, где ILS может быть значительным.

Методы, основанные на конкатенации, выводят филогении из одной объединенной матрицы генов и предполагают, что все гены имеют одинаковую или очень похожую эволюционную историю. Даже когда набор данных разделен таким образом, что к отдельным локусам применяются разные модели эволюции (произведен партишенинг), древо связано по всем разделам, и в результате анализа получается единое древо видов. Однако это предположение может быть неверным, поскольку биологические процессы, такие как гибридизация и неполная сортировка по линиям (ILS), могут приводить к расхождениям между генными деревьями и видовым деревом. Чтобы преодолеть эту проблему, в качестве альтернативы были предложены подходы на основе коалесценции, в

которых генным деревьям разрешено варьироваться между генами, а предположение о том, что все гены имеют одинаковую историю, смягчается. Из этих подходов модель коалесценции для нескольких видов является наиболее популярной для анализа филогеномных данных.

Наиболее часто используемые методы на основе коалесценции в филогеномике называются подходами «сокращенной» или суммарной коалесценции, например, *stem*, *mdc*, *steac*, *star*, *mp-est*, *svdquartets* и *astral*, в которых генные деревья и дерево видов не оцениваются совместно. Эти методы не анализируют данные последовательностей напрямую, а вместо этого используют отдельные деревья генов в качестве входных данных для расчета одного дерева видов, при этом учитывая вариации между генеалогиями.

Методы коалесценции, которые совместно оценивают деревья генов и дерево видов, такие как *beast*, более желательны, но в настоящее время не являются вычислительно разрешимыми для больших наборов данных. *astral* и *svdquartets*, в то же время, являются быстрыми и удобными для пользователя, и было показано, что они как статистически последовательны в многовидовой модели коалесценции, так и хорошо работают в условиях высокого ILS.

Методы конкатенации и коалесценции различаются подходом к обработке данных и предположениями об эволюционной истории. В частности, конкатенация объединяет последовательности всех локусов в одну матрицу (например, 10 генов по 1000 bp = 10000 bp), в итоге предлагая единую топологию для всех локусов. Методы конкатенации используются в ML или парсимонии.

Следует отметить, что конкатенация игнорирует неполную сортировку линий (ILS) и независимую эволюцию генов, в связи с чем может переоценивать поддержку узлов из-за систематической ошибки.

Теория коалесценции помогает проследить, как гены в популяциях и видах связаны через общих предков, двигаясь назад во времени. Она моделирует, как гены в выборке «объединяются» в общих предков, учитывая случайные процессы, такие как генетический дрейф. Методы коалесценции в своей сути «признают», что гены могут расходиться, эволюционировать независимо, и пытаются учесть эти расхождения.

Методы коалесценции рекомендуются для анализа близкородственных видов, где вероятны ILS, например, для оценки демографической истории или видового древа. Они также полезны для задач, связанных с генетическим разнообразием и миграцией. Однако их лучше не использовать при значительных рекомбинациях, так как стандартная модель коалесценции предполагает отсутствие рекомбинации внутри локусов. Например, для древних расхождений конкатенация может быть более подходящей.

Для одного вида модель коалесценции используется в популяционной генетике для изучения генетического разнообразия, оценки эффективного размера популяции и демографической истории. Это помогает понять, как изменялся размер популяции во времени, например, через байесовские графики горизонта.

«Обратная машина времени», это метафора для процесса коалесценции, который идёт назад во времени, чтобы найти общих предков генов. Модель Фишера-Райта описывает случайный дрейф в популяции фиксированного

размера, где каждая особь имеет равные шансы оставить потомков. Коалесценция, это обратный процесс этой модели, фокусирующийся на генеалогии выборки

Нейтральная коалесценция, это стандартная модель, где гены нейтральны, то есть не подвержены отбору, и эволюция определяется только дрейфом и мутациями. Это основа для многих анализов, например, оценки времени коалесценции. В нейтральной модели время до общего предка для двух генов в популяции размера N_e составляет в среднем $2N_e$ поколений.

Симуляция коалесценции проводится с помощью программ, таких как ms, SIMCOAL и coala, для тестирования гипотез, оценки методов и понимания влияния демографических процессов на выборку. Например, она используется для генерации данных под разными сценариями, чтобы проверить, как методы справляются с ILS.

Демографические процессы, такие как изменения размера популяции, миграция и половозрастная структура, влияют на форму филогенетических деревьев. Например, бутылочные горлышки могут привести к потере разнообразия, а рост популяции – к звездообразным деревьям. Миграция может усложнить анализ, вызывая ретикулянтную (сетчатую) эволюцию.

Процессы Пуассона моделируют время коалесцентных событий. Гомогенные процессы имеют постоянную скорость, что подходит для популяций с фиксированным размером, а негомогенные – переменную, что отражает изменения размера популяции во времени.

Гомогенные процессы имеют так называемую постоянную интенсивность и используются, например, для моделирования коалесценции в стабильных популяциях. Негомогенные процессы имеют переменную интенсивность, что важно для учёта демографических изменений, таких как рост или спад численности.

Непараметрические демографические модели, это гибкие методы оценки истории размера популяции без предположений о её форме, например, байесовские графики горизонта, полезные для сложных демографических сценариев. При этом для асексуальных видов генеалогия проста, без рекомбинации, а для сексуальных видов учитывает рекомбинацию, что требует использования сложных моделей, таких как MSC.

Модель MSC используется для оценки видового древа, интегрируя генетические деревья и учитывая ILS, с помощью методов, таких как *BEAST или ASTRAL.

Байесовское разграничение видов, это метод, использующий байесовский подход для оценки вероятности существования разных видов в выборке на основе генетических данных, часто с помощью BPP. При этом для определения монофилии или парафилии биоинформатическими методами строится филогенетическое древо, и проверяется, образует ли группа единый кластер (монофилия) или исключает часть потомков (парафилия), с использованием тестов, таких как Shimodaira-Hasegawa.

1.6. Молекулярные часы и вычисление времени дивергенции таксонов

Молекулярные часы, это метод оценки времени расхождения видов на основе накопления мутаций в ДНК, что возможно в предполагаемых условиях относительно постоянной скорости эволюции. *Глобальные часы* предполагают одну скорость для всех линий, *локальные* – разные скорости для разных частей дерева, а *расслабленные часы* учитывают вариации скоростей в разных ветвях и с течением времени.

Калибровка молекулярных часов осуществляется с помощью ископаемых остатков, предоставляющих известные даты расхождения. Молекулярные часы применяются для датировки происхождения вирусов, используя известные даты изоляции штаммов, и крупных таксонов – ссылаясь на древние ископаемые.

Говоря более детально, молекулярные часы, это метод, который использует скорость накопления мутаций в ДНК или белках для оценки времени расхождения видов. Исследования показывают, что мутации накапливаются примерно с постоянной скоростью, что позволяет использовать генетические различия для определения момента, когда сравниваемые два вида имели последнего общего предка. Например, если два вида отличаются на 5 % в ДНК и известно, что скорость мутаций составляет 1 % за миллион лет, то расхождение могло произойти около 5 миллионов лет назад.

Эта теория, предложенная Зукеркандлем и Полингом в 1960-х, основана на наблюдении, что различия в аминокислотных последовательностях, например, гемоглобина, пропорциональны времени, прошедшему с момента расхождения видов, насколько мы можем установить это время при помощи палеонтологических данных.

Для проверки того, сохраняется ли постоянство молекулярных часов (является ли скорость эволюции константной), используются статистические методы, такие как тест относительных скоростей и тесты отношения правдоподобия.

Тест относительных скоростей сравнивает генетические расстояния между двумя видами и внешней группой: если скорости одинаковы, различия должны быть схожими. Например, если человек и шимпанзе одинаково отличаются от обезьяны Нового Света, это указывает на одинаковую скорость эволюции. Тесты отношения правдоподобия сравнивают модели с и без предположения о постоянной скорости эволюции, используя максимальное правдоподобие для оценки, какая модель лучше описывает данные. Эти методы помогают выявить, есть ли значительные вариации скоростей между линиями, что важно для выбора подходящей модели часов.

Молекулярные часы имеют несколько ограничений, которые могут повлиять на их точность.

1. Вариации в скорости эволюции:

Скорость эволюции может различаться из-за скорости смены поколений, размеров популяции, уровня метаболизма (считается, что высокий уровень метаболизма приводит к окислительному стрессу в клетках и повышает скорость мутации) и эффективности работы систем репарации ДНК.

2. Неполнота ископаемых данных:

Ископаемые остатки могут не отражать точное время расхождения линий, а их возраст может быть лишь приблизительной оценкой, что увеличивает неопределенность.

3. Качество последовательностей:

Ограниченное количество или низкое качество данных может привести к статистическим ошибкам в получаемых оценках.

4. Ограничения применяемых моделей:

Выбор неподходящей модели эволюции может привести к искажениям, особенно если не учитывать рекомбинацию или отбор.

Индекс дисперсии скорости эволюции

Индекс дисперсии, или отношение дисперсии к среднему числу замен, используется для измерения вариативности числа замен между линиями. При постоянной скорости, как в распределении Пуассона, этот индекс равен 1.

Если индекс больше 1, это указывает на «передисперсию» (чрезмерную дисперсию), что означает большее разнообразие замен, чем ожидается при постоянной скорости. Этот индекс применяется для тестирования гипотезы молекулярных часов, помогая определить, есть ли отклонения от ожидаемой постоянной скорости, что важно для выбора модели (например, строгих или расслабленных часов).

Глобальные молекулярные часы предполагают, что скорость эволюции одинакова для всех ветвей филогенетического дерева. Простая модель.

Локальные молекулярные часы допускают разные скорости для разных частей дерева, часто используя расслабленные модели, где скорость может варьироваться между линиями. Например, случайные локальные часы предлагают серию локальных часов, каждый из которых охватывает подрегион филогении, с возможными изменениями скорости на ветвях. Эти модели помогают учитывать гетерогенность скоростей, что важно для точных оценок времени расхождения эволюционных линий.

Эвристические подходы для сглаживания различий в скоростях эволюции включают разделение данных на подгруппы с похожими скоростями эволюции или использование алгоритмов кластеризации для группировки генов по сходству их деревьев. Например, можно применять отдельные модели часов для кластеров генов, которые демонстрируют схожие паттерны вариации скоростей между ветвями, что улучшает итоговую точность оценок.

Расслабленные молекулярные часы, это модели, которые позволяют учитывать вариации скоростей эволюции между линиями, в отличие от строгих часов, предполагающих постоянную скорость. Они включают автокоррелированные модели, где скорость коррелирует между соседними ветвями, и некоррелированные, где скорости независимы и следуют определенному распределению. Эти модели широко используются в байесовских анализах, особенно для данных с высокой гетерогенностью скоростей.

Калибровка с помощью ископаемых остатков включает использование возраста ископаемых для установки времени расхождения на определенных узлах филогении. Например, если ископаемое вида датируется 15 миллионами лет назад

и известно, что оно является общим предком двух современных видов, это время используется для калибровки. Однако ископаемые могут представлять лишь минимальную оценку, и их идентификация может быть неопределенной, что увеличивает общую неопределенность калибровок.

Байесовские методы используют МСМС для оценки времен дивергенции. Например, программа МСМС Tree позволяет анализировать большие геномные наборы, используя расслабленные часы и гибкие калибровки. Это позволяет учитывать неопределенности, такие как вариации скоростей и ошибки калибровки, предоставляя доверительные интервалы для оценок.

Приблизительный расчет правдоподобия используется для ускорения вычислений в больших наборах данных, где точные вычисления могут быть слишком ресурсоемкими. Например, TreeTime использует итеративную оптимизацию для приближенного максимального правдоподобия, что помогает в анализе быстро эволюционирующих вирусов. Это особенно полезно для молекулярных часов, где требуется обработка больших филогенетических деревьев.

Метод молекулярных часов особенно полезны для анализа эволюции вирусов (из-за их быстрой эволюции). Они калибруются с использованием известных дат изоляции вируса или исторических записей об инфекционных вспышках, вызванным этим вирусом. Для крупных таксонов, таких как филы или классы, молекулярные часы калибруются с помощью древних ископаемых или известных времен расхождения. Например, расхождение между млекопитающими и птицами может быть калибровано с помощью ископаемых их последнего общего предка, а затем часы используются для оценки времен дивергенции других групп. Это позволяет оценить время появления крупных групп, таких как цветковые растения или основные группы животных, с учетом неопределенностей из-за ограничений ископаемых данных и вариаций скоростей.

1.7. Нейтральная и адаптивная эволюция белков

В соответствии с нейтральной теорией молекулярной эволюции, большинство генетических изменений на молекулярном уровне являются нейтральными, то есть не влияют на приспособленность организма и не подвергаются естественному отбору. Эти нейтральные мутации накапливаются в популяциях преимущественно благодаря случайному генетическому дрейфу.

Среди фактов, подтверждающих теорию нейтральной эволюции можно назвать следующие:

1. Высокий уровень генетического полиморфизма, обнаруженный в популяциях, трудно объяснить только действием естественного отбора, что поддерживает идею о значительной роли нейтральных мутаций.

2. Наблюдения показывают, что молекулярные часы, т.е. скорость накопления мутаций, остаются относительно постоянными для различных видов, что согласуется с предположением о случайном накоплении нейтральных мутаций.

Для проверки соответствия генетических данных нейтральной теории разработаны специальные статистические тесты.

D-тест Таджимы.

Тест Таджимы сравнивает два разных оценочных параметра генетического разнообразия: среднее число парных различий и количество полиморфных сайтов. Если наблюдаемое значение D значительно отклоняется от нуля, это может свидетельствовать о действии отбора или демографических изменений. Тест используется для выявления отклонений от нейтральной эволюции в популяционных данных.

Тест Макдональда-Крейтмана.

Предложенный в 1991 году Джоном Макдональдом и Мартином Крейтманом, этот тест сравнивает соотношение синонимичных и несинонимичных замен внутри вида (полиморфизм) и между видами (дивергенция). Отклонения от ожидаемого соотношения могут указывать на положительный или отрицательный отбор. Тест широко применяется для оценки роли естественного отбора в молекулярной эволюции.

Для идентификации филогенетических линий, подвергшихся адаптивной эволюции, используются эвристические методы, такие как анализ соотношения несинонимичных и синонимичных замен (dN/dS). Значение $dN/dS > 1$ может свидетельствовать о положительном отборе. Также могут быть использованы статистические модели, оценивающие вероятность различных сценариев эволюции и позволяющие выявлять участки генома под действием отбора.

1.8. Симуляция молекулярной эволюции

Компьютерная симуляция процессов, происходящих в ДНК, позволяет воссоздать эволюционные изменения с течением времени в виртуальной модели. Такие симуляции используют алгоритмы, моделирующие мутации, рекомбинацию, генетический дрейф, естественный отбор и другие процессы, которые влияют на генетический состав популяций. Это позволяет исследователям протестировать гипотезы, сравнить модели эволюции с реальными данными и предсказать возможное развитие генетического разнообразия. Симуляция позволяет сравнивать теоретические ожидания с наблюдаемыми данными; тестировать методы построения филогенетических деревьев или оценки параметров популяционной генетики; прогнозирование эволюционных сценариев. Прогнозирование сценариев особенно актуально для патогенов, где предсказание мутаций может помочь в разработке вакцин и методов лечения.

Для симуляции эволюционных процессов применяются статистические и стохастические модели. Такие модели могут основываться на принципах нейтральной теории, моделировать процессы отбора, изменяющиеся размеры популяций, миграции и другие демографические события. Программные инструменты используют данные модели для генерации синтетических последовательностей ДНК или целых геномов.

Основные подходы к симуляции эволюции

Симуляция последовательностей на фиксированном дереве проводится на заранее заданном филогенетическом дереве. При таком подходе дерево (топология и длины ветвей) фиксированы, и выбранная модель эволюции используется для генерации последовательностей на концах ветвей, получение “поздних” разветвлений и т.д.

Этапы симуляции эволюции на фиксированном дереве:

Выбирается или реконструируется филогенетическое дерево на основе эмпирических данных. Дерево фиксируется, и его параметры (длины ветвей, порядок ветвления) служат основой для симуляции.

Определяются параметры модели подстановок, которые описывают скорость мутаций, вероятности различных замен и прочие характеристики эволюционного процесса.

Исходная (родительская) последовательность эволюционирует по дереву, где в каждой точке ветвления происходят случайные мутации согласно заданной модели. Программы типа Seq-Gen используют этот подход для генерации синтетических данных

Симуляция случайных деревьев и множественных псевдо-реплик, это подход, при котором случайным образом генерируются филогенетические деревья согласно заданным статистическим свойствам (например, равновероятное распределение топологий или использование модели рождения-смерти). Такие деревья могут отражать разнообразие гипотетических эволюционных сценариев.

Путём генерации большого числа случайных деревьев можно создать набор псевдо-реплик данных, который позволяет оценить стабильность статистических оценок (аналогично бутстрэп-анализу) и проверить гипотезы относительно эволюционных процессов. Такие реплики помогают выявить, насколько результаты анализа чувствительны к изменению топологии дерева.

Симуляция популяционных процессов

Моделирование динамики популяций включает генерацию сценариев, в которых учитываются флуктуации частот аллелей, эффект генетического дрейфа, мутации, миграция и естественный отбор. Т.е. при моделировании эволюции внутри популяций учитываются не только изменения последовательностей, но и демографические процессы.

Для симуляции популяционных процессов могут быть использованы стохастические модели. Например, модель Винера или модели на основе пуассоновских процессов позволяют моделировать случайные изменения в составе популяции. Такие модели помогают предсказать, как флуктуации численности популяции или эффекты бутылочного горлышка влияют на генетическое разнообразие.

Программы типа ms или msms используют метод Монте-Карло для генерации эволюционных сценариев, что позволяет учитывать случайность в мутациях, рекомбинации и генетическом дрейфе. Это особенно важно для изучения эволюционной истории и распределения генетического полиморфизма в популяциях. Такие симуляции позволяют оценивать вероятность возникновения определённых генетических структур, моделировать распространение новых

мутаций и анализировать влияние демографических событий на генетическую структуру популяции.

Программное обеспечение для симуляции эволюционных процессов

Существует множество программных инструментов, каждый из которых имеет свои особенности.

Seq-Gen: инструмент для симуляции эволюции последовательностей по фиксированному дереву с использованием различных моделей подстановок. Поддерживает различные модели подстановок и позволяет генерировать синтетические данные для тестирования алгоритмов построения деревьев. Удобен для создания данных по фиксированному дереву

ms и *msms*: симулятор популяционной генетики, который моделирует эволюционные процессы с учетом генетического дрейфа, рекомбинации, демографических событий и отбора. Эти инструменты особенно полезны для симуляций, основанных на моделях нейтральной эволюции и для глубокого анализа популяционных процессов

SimPhy: используется для симуляции не только последовательностей, но и целых генеалогических деревьев, что позволяет моделировать конфликт между генными и видовыми деревьями. Подходит для оценки многоуровневых эволюционных сценариев

BEAST и *MrBayes*: программы для байесовского анализа филогенетических деревьев, которые могут включать эволюционные симуляции для оценки временных параметров и модели изменений.

2. ПРАКТИЧЕСКИЙ РАЗДЕЛ

2.1. Примерный перечень заданий для управляемой самостоятельной работы магистрантов

В качестве заданий для управляемой самостоятельной работы студенты магистратуры выполняют компьютерное тестирование по результатам изучения соответствующих разделов дисциплины.

Раздел 1. Эволюционные процессы в ДНК и белках. Характеристика эволюционных процессов в ДНК и аминокислотных последовательностях. Биологические интерпретации соотношения dN/dS . Применение dN/dS в сравнительной геномике. Характеристика эволюционных процессов в ДНК и аминокислотных последовательностях. Статистический анализ получаемых деревьев.

Тема 1.4 Сравнение методов и тестирование достоверности филогенетических деревьев. Построение сложных филогений в биологических исследованиях.

Раздел 2. Эволюционные процессы в популяциях и крупных таксонах. Примеры построения сложных филогений в биологических исследованиях. Молекулярные часы. Неопределенность в вычислениях времени дивергенции. Нейтральная теория эволюции. Биологическое значение симуляции молекулярной эволюции.

Тема 2.4 Симуляция молекулярной эволюции. Получение множественных псевдо-реплик. Предсказание флуктуаций частот аллелей.

2.2. Примерная тематика практических занятий

Занятие № 1. Программное обеспечение для вычисления генетических дистанций. Подбор оптимальной модели нуклеотидных и аминокислотных замещений. Вычисление dS и dN .

Занятие № 2. Построение филогенетических деревьев: подготовка исходных данных, составление композитных последовательностей, применение дистанционных и дискретных методов построения филогений, интерпретация выходных данных.

Занятие № 3. Байесовский метод при построении филогений. Валидация MCMC алгоритма. Оптимизация выбора алгоритма для проведения филогенетического анализа.

Занятие № 4. Методы оценки правдоподобия, согласованности, эффективности и надежности получаемых реконструкций филогений.

Занятие № 5. Программное обеспечение для построения видовых деревьев. Анализ конфликтных ситуаций при построении видовых и генных деревьев. Байесовское разграничение видов.

Занятие № 6. Модели молекулярных часов и их калибрование. Расчет относительной скорости эволюции в разных таксонах. Байесовское вычисление времени дивергенции.

Занятие № 7. Анализ эволюции белковых молекул при помощи D-теста Таджимы, теста Макональда-Крейтмана и Branch-site теста. Сравнение методов для детекции позитивного отбора.

Занятие № 8. Программы, используемые для симуляции эволюционных процессов. Симуляция эволюции вирусов и других патогенов.

2.3. Примерная тематика реферативных работ

1. Спорные филогенетические линии в наше время
2. Актуальные проблемы и способы решения построения сложных филогений
3. Достижения в области симуляции эволюции патогенных организмов
4. Современные примеры адаптивной эволюции в разных таксонах
5. Актуальные программы для реконструкции филогений: сильные и слабые стороны
6. Конфликтные ситуации при построении видовых и генных деревьев и возможные пути их решения
7. Теории эволюции: история возникновения и основополагающие постулаты
8. История развития дерева жизни

3. РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

3.1. Вопросы и задания для самоконтроля

1. Характеристика эволюционных процессов в ДНК, РНК и белках.
2. Выравнивание последовательностей. Вероятностные алгоритмы выравнивания последовательностей.
3. Основные матрицы нуклеотидных и аминокислотных замен, принципы их применения.
4. Скорость эволюции и принципы ее вычисления.
5. Принципы и основные методы построения филогений. Байесовские филогении.
6. Марковские цепи и их использование в эволюционных исследованиях.
7. Датирование происхождения таксонов – основные методы и их ограничения.
8. Адаптивная эволюция и позитивная селекция – методы их вычисления.
9. Симуляция эволюционных процессов на уровне нуклеотидных и белковых последовательностей – принципы и ограничения.
10. Наиболее известные примеры использования симуляции эволюционных процессов в биологии.

3.2. Вопросы для подготовки к экзамену

1. Характеристика эволюционных процессов в ДНК и аминокислотных последовательностях.
2. Характеристика Марковских моделей нуклеотидных замещений. Выбор лучшей модели нуклеотидных замещений.
3. Марковские цепи: характеристика и область применения.
4. Способы вычисления отношения транзиций и трансверсий (T_i , T_v , T_i/T_v).
5. Эмпирические и механистические модели аминокислотных замещений.
6. Модель Пуассона и Гамма-дистанция как способ вычисления дистанции между белками. Основные модели кодонных замещений.
7. Вычисление dS и dN . Биологические интерпретации соотношения dN/dS . Применение dN/dS в сравнительной геномике.
8. Классификация методов построения филогенетических деревьев.
9. Стохастические методы построения деревьев.
10. Отличительные особенности построения аминокислотных и нуклеотидных филогенетических деревьев.
11. Композитные последовательности. Секционирование.
12. Методы построения филогенетических деревьев: наименьших квадратов, минимум эволюции, присоединение ближайшего соседа, максимального правдоподобия.
13. Способы вероятностного вычисления предковой последовательности. Протяжение длинных ветвей.
14. Биологическая интерпретация результатов филогенетического анализа.

15. Характеристика и область применения Марковской цепи Монте-Карло (МСМС). Валидация МСМС алгоритма.
16. Характеристика Алгоритма Метрополиса-Гастингса. Отношение вероятностей и семплирование по Гиббсу.
17. Байесовский метод при построении филогений: Байесовский вывод, фактор и маргинальное правдоподобие.
18. Подходы к оптимизации выбора алгоритма для реконструкции филогений.
19. История крупномасштабных геномных изменений: вопросы биологии и вычислительные задачи.
20. Статистические критерии при анализе филогенетических деревьев.
21. Бутстрэп-анализ и тест длины внутренних ветвей при проверке достоверности топологии филогенетического дерева.
22. К-Н тест при проверке достоверности топологии филогенетического дерева. Индексы, используемые в анализе парсимоний.
23. Определение апостериорной вероятности и высокой апостериорной вероятности.
24. Характеристика NNI, SPR и TBR статистики для неукорененных и укорененных деревьев.
25. Описание парадокса звездчатых деревьев.
26. Сложные филогении в биологических исследованиях - ограничения применимости филогений.
27. Сложные филогении в биологических исследованиях - решаемые задачи.
28. Модель коалесценции в применении к единичному виду и системе из нескольких видов.
29. Модель Фишера-Райта и нейтральная коалесценция. Симуляция коалесценции.
30. Изменения размера популяций. Непараметрические демографические модели для популяций.
31. Демографические процессы в популяциях. Генеалогические методы для однополых и двуполых видов.
32. Построения видовых деревьев. Конфликт видовых и генных деревьев.
33. Байесовское разграничение видов. Монофилетические и парафилетические группы видов.
34. Ограничения применимости молекулярных часов при реконструкции филогений.
35. Модели глобальных и локальных молекулярных часов. Расслабленные молекулярные часы.
36. Байесовское вычисление времени дивергенции. Неопределенность в вычислениях времени дивергенции.
37. Способы датирования происхождения крупных таксонов.
38. Нейтральная теория эволюции. Тестирование нейтральности эволюции.

39. Эвристический метод при определении линий, осуществляющих адаптивную эволюцию.

40. Характеристика тестов, применяемых в позитивной селекции: тест правдоподобия позитивной селекции с использованием случайных сайтов, branch-site тест.

41. Сравнение и ограничения методов для детекции позитивного отбора.

42. Адаптивно эволюционирующие гены. Адаптивная эволюция в крупных таксонах растений и животных.

43. Характеристика способов симуляции процессов, происходящих в ДНК с течением времени. Программы, используемые для симуляции.

44. Характеристика способов симуляции процессов, происходящих в популяциях. Программы, используемые для симуляции.

45. Симуляция молекулярной эволюции: биологическое значение, применяемые программы. Примеры симуляции эволюции различных патогенов.

4. ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

4.1. Рекомендуемая литература

Основная

1. Гены по Льюину / Дж. Кребс, Э. Голдштейн, С. Килпатрик ; пер. с англ. [И. А. Кофиади и др.] под редакцией Д. В. Ребрикова и Н. Ю. Усман. - 5-е изд., испр. - Москва: Лаборатория знаний, 2022. - 919 с.
2. Компо,Ф. Алгоритмы биоинформатики / Филипп Компо и Павел Певзнер ; [пер. с англ. И. Л. Люско]. - Москва : ДМК Пресс, 2023. - 681 с.

Дополнительная

1. Геномика с молекулярно-генетическими основами / В. В. Попов. - Изд. 2-е. - Москва: URSS, 2012. - 298 с.
- 1.Higgs P. G., Attwood T. K. Bioinformatics and Molecular Evolution / P. G. Higgs, T. K. Attwood, 1st edition, Malden, MA: Wiley-Blackwell. – 2005. – 384 p.
2. Pevzner P., Shamir R. Bioinformatics for Biologists, 1st edition, Cambridge University Press. – 2011. – 394 p.
3. Yang Z. Molecular Evolution: A Statistical Approach / Z. Yang, Oxford: Oxford University Press. – 2014. – 512 p.
4. Shapiro J. A. Evolution: A View from the 21st Century / J. A. Shapiro, 1st edition, Upper Saddle River, NJ. – 2013. – 239 p.
5. Forsdyke Evolutionary Bioinformatics / Forsdyke, 3rd ed., New York, NY: Springer. – 2016. – 508 p.
6. Gagniuc P. A. Algorithms in Bioinformatics: Theory and Implementation / P. A. Gagniuc, 1st edition, Hoboken, NJ: Wiley. – 2021. – 528 p
7. Cornish-Bowden A. Biochemical Evolution: The Pursuit of Perfection / A. Cornish-Bowden, 2nd edition, New York: Garland Science. – 2016. – 300 p.
8. Haubold B., Börsch-Haubold A. Bioinformatics for Evolutionary Biologists: A Problems Approach / B. Haubold, A. Börsch-Haubold, Softcover reprint of the original 1st ed., Cham: Springer. – 2019. – 331 p.

4.2. Электронные ресурсы

1. Data Science Workbook [Electronic resource]. – Mode of access: <https://datascience.101workbook.org>. – Date of access: 06.01.2025.