Филогенетика и кладистика

Автор: Никитин Павел

Версия 2.0

В настоящем пособии автор попытался максимально, с его точки зрения, подробно рассмотреть довольно молодой раздел биологических наук – филогенетику. В пособии рассмотрен современный подход к объяснению эволюционных событий и классификации живых организмов с той стороны, как это видят биоинформатики. Издавна эволюционный путь изображался в виде дерева жизни, сегодня же от него осталась лишь схема – граф. Именно поэтому большой акцент в пособии ставится на математику. Также рассмотрены базовые понятия, необходимые для интерпретации филогенетических деревьев, которые формируют кладистику, как таковую. Концепция состоит в том, что наиболее объективные деревья можно получить, основываясь не на морфологических признаках видов, а на их геномах и протеомах, поэтому в пособии рассматривается процесс получения последовательности мономеров нуклеиновых кислот, называемый секвенированием. Далее задача становится в чистом виде комбинаторной, поэтому речь заходит о выравнивании последовательностей и непосредственно построении деревьев. В целом, в пособии рассматривается биоинформатика последовательностей.

Пособие основано на цикле статей интернет-сообщества «Наука и рок-н-ролл» http://vk.com/klausius со многими дополнениями. Изначальный замысел цикла выглядел несколько иначе, по сравнению с тем, что получилось. В итоге была немного нарушена структура, введено много не поясненных терминов, а также обещаний рассказать что-то позднее. Накопившейся информации было настолько много, что было принято решение оформить все это в виде отдельного пособия. Вероятно, даже в нем не удастся исправить все неточности, поэтому в начале заявлена версия пособия. Читатель всегда имеет право предложить свой вариант исправления в том или ином месте. Стоит также оговориться, что здесь на данный момент не рассмотрены вопросы сборки последовательностей после секвенирования и множественного выравнивания последовательностей. Это будет исправлено в будущем. Во второй версии было добавлено некоторое количество информации про анализ белков и принципы выравнивания последовательностей.

Оглавление

Введение

Все вы наверняка видели, что эволюционный путь чаще всего изображают в виде дерева. Оно могло быть схематичным или красочным с иллюстрациями, прямо растущим вверх или круглым. Но в любом случае это дерево, по которому можно проследить эволюционный путь и сестринские отношения между группами. Для чего это важно, и почему издревле систематики пытались уложить классификацию под эволюционный путь? Любая область биологии, будь то зоология или биохимия, неразрывно связана с представлениями об эволюции. Нам важно знать именно родственные отношения между живыми существами, чтобы понимать различные биологические процессы, но природа, порой, дает нам много различных подвохов, которые усложняют эту задачу. Например, часто встречающаяся конвергентная эволюция дает нам очень похожие внешне организмы, но они могут быть очень далекими родственниками.

Собственно здесь мы познакомимся с основными понятиями филогенетики, поймем, как строятся филогенетические деревья, какие молекулярные методы в них заложены, а потом научимся составлять такие деревья самостоятельно.

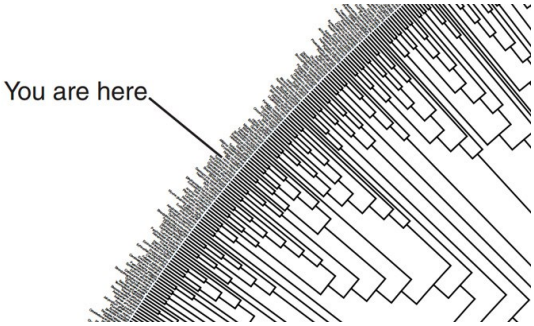
Начнем с нескольких основных понятий:
Филогенетика - наука, изучающая процессы образования биоразнообразия.
Кладистика - направление филогенетической систематики. Характерные особенности кладистической практики состоят в использовании так называемого кладистического анализа (строгой схемы аргументации при реконструкции родственных отношений между таксонами), строгом понимании монофилии и требовании взаимно-однозначного соответствия между реконструированной филогенией и иерархической классификацией. Кладистический анализ — основа большинства принятых в настоящее время биологических классификаций, построенных с учётом родственных отношений между живыми организмами.

Кладистика относится к числу трёх ведущих таксономических школ, доминирующих в современной биологической систематике. Ей противостоят фенетика, основанная на количественной оценке так называемого общего сходства, и эволюционная таксономия, которая, подобно кладистике, при построении системы опирается на эволюционную близость (то есть общность происхождения), однако не требует строгого соответствия системы и филогении (в частности, это выражается в признании права на существование в системе парафилетических групп).

Что важно помнить - сравнение внешних признаков не всегда точно передает родственные отношения организмов, но есть одна штука, которая однозначно ответит на поставленный вопрос о взаимосвязи - наследственная информация в форме ДНК и РНК. Как вообще узнать нуклеотидную последовательность организма? Для этого есть отличный способ секвенирования. О нем подробно мы поговорим далее, но в целом можно сказать, что молекулу делят на фрагменты и прогоняют их через аппарат, у которого есть молекулы-датчики, проверяющие комплементарность нуклеотидов. Все это обрабатывается программой и выдается исследователю в виде последовательности букв.

Секвенатор Illumina MiSeq

Для собственного интереса вы можете воспользоваться сайтом с филогенетическим деревом огромного числа живых организмов: tree.opentreeoflife.org



Деревья и все-все-все

Здесь мы познакомимся с основным объектом нашего дальнейшего изучения - филогенетическим древом, что это и с чем его едят. И чтобы подробно в этом разобраться, нам нужно погрузиться в теорию графов

Дерево — это связный ациклический граф. Связность означает наличие путей между любой парой вершин, ацикличность — отсутствие циклов и то, что между парами вершин имеется только по одному пути.
Лес — упорядоченное множество упорядоченных деревьев.
Ориентированное (направленное) дерево — ацикличный орграф (ориентированный граф, не содержащий циклов), в котором только одна вершина имеет нулевую степень захода (в неё не ведут дуги), а все остальные вершины имеют степень захода 1 (в них ведёт ровно по одной дуге). Вершина с нулевой степенью захода называется корнем дерева, вершины с нулевой степенью исхода (из которых не исходит ни одна дуга) называются концевыми вершинами или листьями.

Связанные определения

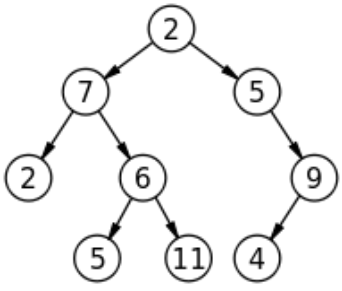
Степень вершины — количество инцидентных ей ребер.
Концевой узел (лист, терминальная вершина) — узел со степенью 1 (то есть узел, в который ведёт только одно ребро; в случае ориентированного дерева — узел, в который ведёт только одна дуга и не исходит ни одной дуги).
Узел ветвления — неконцевой узел.
Дерево с отмеченной вершиной называется корневым деревом.
Уровень узла — длина пути от корня до узла. Можно определить рекурсивно:

1. уровень корня дерева T равен 0;
2. уровень любого другого узла на единицу больше, чем уровень корня ближайшего поддерева дерева T, содержащего данный узел.

Остовное дерево (остов) — это подграф данного графа, содержащий все его вершины и являющийся деревом. Рёбра графа, не входящие в остов, называются хордами графа относительно остова.
Несводимым называется дерево, в котором нет вершин степени 2.
Лес — множество (обычно упорядоченное), не содержащее ни одного непересекающегося дерева или содержащее несколько непересекающихся деревьев.

Двоичное дерево

Простое бинарное дерево размера 9 и высоты 3, с корнем значения 2. Это дерево не сбалансировано и не отсортировано.



Термин двоичное дерево (применяется так же термин бинарное дерево) имеет несколько значений:

- Неориентированное дерево, в котором степени вершин не превосходят 3.

- Ориентированное дерево, в котором исходящие степени вершин (число исходящих рёбер) не превосходят 2.
- Абстрактная структура данных, используемая в программировании.

N-арные деревья

N-арные деревья определяются по аналогии с двоичным деревом. Для них также есть ориентированные и неориентированные случаи, а также соответствующие абстрактные структуры данных.

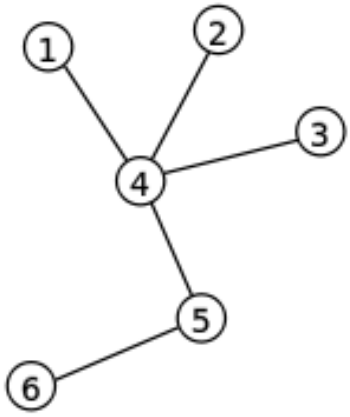
- N-арное дерево (неориентированное) — это дерево (обычное, неориентированное), в котором степени вершин не превосходят N+1.
- N-арное дерево (ориентированное) — это ориентированное дерево, в котором исходящие степени вершин (число исходящих рёбер) не превосходят N.

Свойства

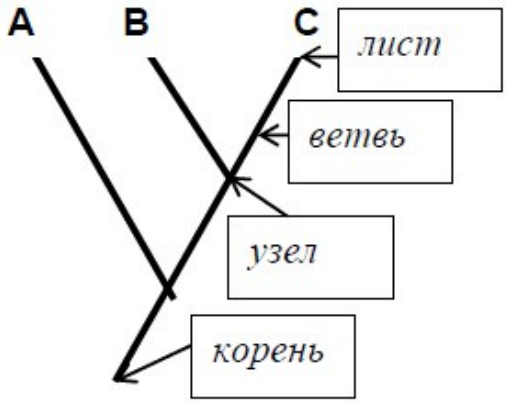
- Дерево не имеет кратных рёбер и петель.
- Любое дерево с n вершинами содержит n-1 ребро. Более того, конечный связный граф является деревом тогда и только тогда, когда B-P=1 где B— число вершин, P— число рёбер графа.
- Граф является деревом тогда и только тогда, когда любые две различные его вершины можно соединить единственной простой цепью.
- Любое дерево однозначно определяется расстояниями (длиной наименьшей цепи) между его концевыми (степени 1) вершинами.
- Любое дерево является двудольным графом.
- Любое дерево, множество вершин которого не более чем счётное, является планарным графом.
- Для любых трёх вершин дерева, пути между парами этих вершин имеют ровно одну общую вершину.

Кодирование деревьев

- Дерево можно кодировать наборами из нулей и единиц.
- Код Прюфера однозначно сопоставляет произвольному конечному дереву последовательность.
- Дерево можно задать набором скобок где пара скобок соответствует одной вершине, которые соединены ребром если соответствующие скобки непосредственно одна в другой. Например дерево на рисунке можно записать как ((()()(()))), взяв корень вершину 1, или (()()()(())), взяв за корень вершину 4.



Древо в нашем случае отражает родственные отношения организмов, по которому можно проследить кто кому родственно ближе, кто дальше. Применим полученные знания теории графов к нашей теме. Дерево формально состоит из следующих частей:



Листья – рассматриваемые объекты (A, B и C);
Узлы – точки схождения ветвей (узел, указанный стрелкой, объединяет объекты B и C, а нижележащий узел уже объединяет группу B+C и объект A);

Ветви – линии, соединяющие листья с узлами и узлы друг с другом;
Корень – узел, объединяющий все рассматриваемые объекты в одну группу.

Тут важное замечание: все объекты, которые мы рассматриваем, даже если они вымерли (а они состоят из всех особей той или иной систематической группы, чаще всего вида) лежат исключительно на листьях, то есть на тех местах, где схематичные ветви прерываются. То есть на самих ветвях никто не лежит, ветви нужны для измерения времени эволюции (не всегда), генетической разницы между видами. Узлы являются лишь группирующими объектами, которые с некоторым приближением можно назвать гипотетическим предком, который дивергировал на 2 группы. То есть была популяция организмов, которая из-за определенных условиях разделилась на две и теперь эти популяции доэволюционировали независимо до новых видов. Фактически популяция организмов, давшая начало двум группам реально существовала, но в палеонтологической летописи мы их скорее всего не найдем. Вероятность того, что найденный организм принадлежит именно этой популяции настолько мала, что ей пренебрегают и найденный организм выносят в маленькую веточку от узла. Ну это в том случае, если организм очень похож на узловой, составленный нами аналитически.

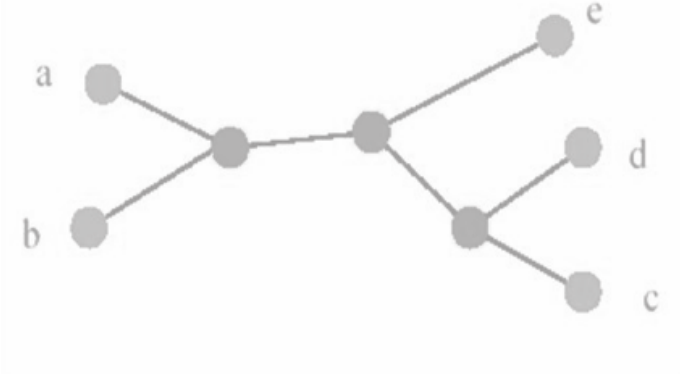
Внимание: чем больше узлов соединяет 2 листа, тем менее родственны соответствующие биологические объекты относительно других, представленных на древе. С глобальной точки зрения деревья бывают укорененные и неукорененные. Неукорененное древо показывает лишь относительную эволюционную близость рассматриваемых объектов и не содержит информацию, по которой можно предсказать базальные группы.

Какие бывают деревья

Укорененное дерево (rooted tree) отражает направление эволюции (есть общий предок)

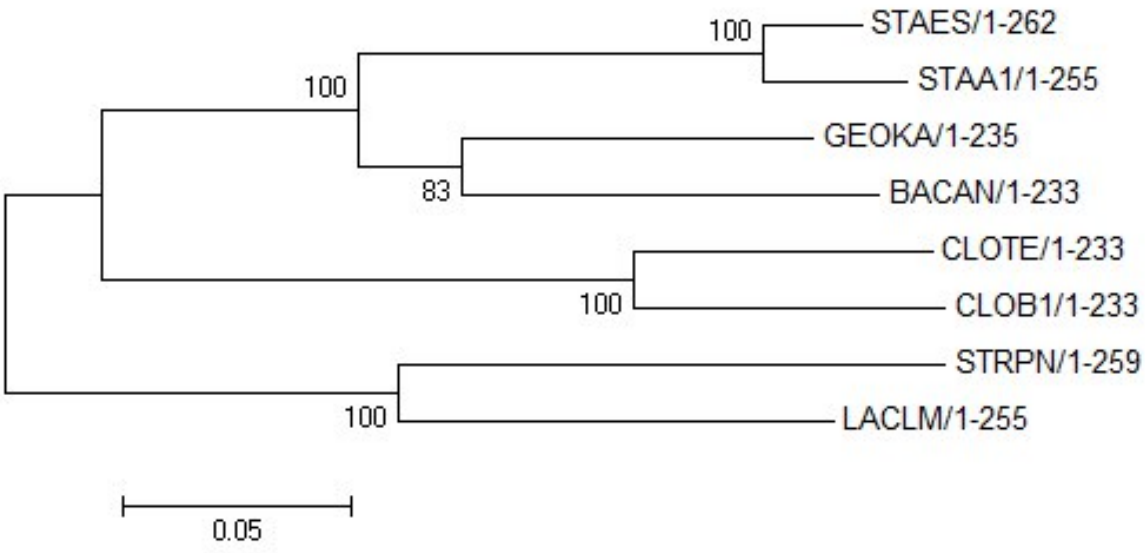
Если число листьев равно n, существует (2n-3)!! разных бинарных укорененных деревьев.
По определению, (2n-3)!! = 1*3*…*(2n-3)

Неукорененное дерево (unrooted tree) показывает связи только между узлами (общий предок не предлагается)



Существует (2n-5)!! Разных неукороннных деревьев с n листьями

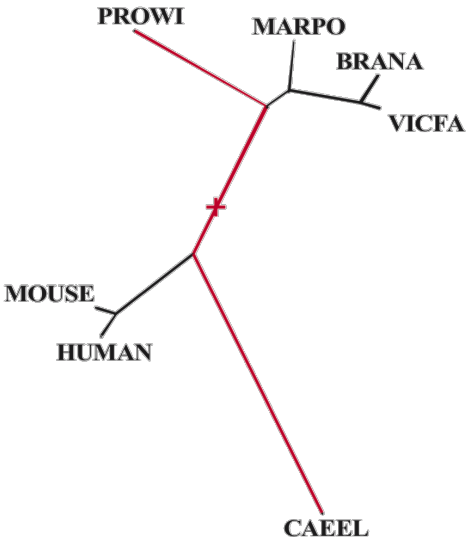
На деревья часто наносят дополнительную информацию, например, длины ветвей могут показывать степень их родства (чем короче, тем больше) или время дивергенции (чем короче, тем позднее произошла). Также дополнительная численная информация может быть указана в узлах древа. Часто это так называемая бутстреп-поддержка, измеряемая в процентах, которая прямо пропорциональна надежности формирования данной группы. Ну например я напишу в узел 80, и это будет означать, что я могу ошибиться с вероятностью в 20%.



Неукорененные деревья можно укоренять! Вариантов в таком случае несколько:

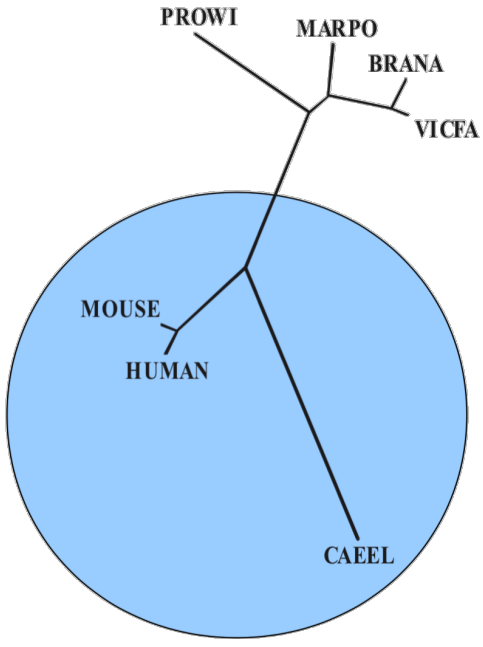
В среднюю точку (он же «по самой длинной ветви») – midpoint rooting

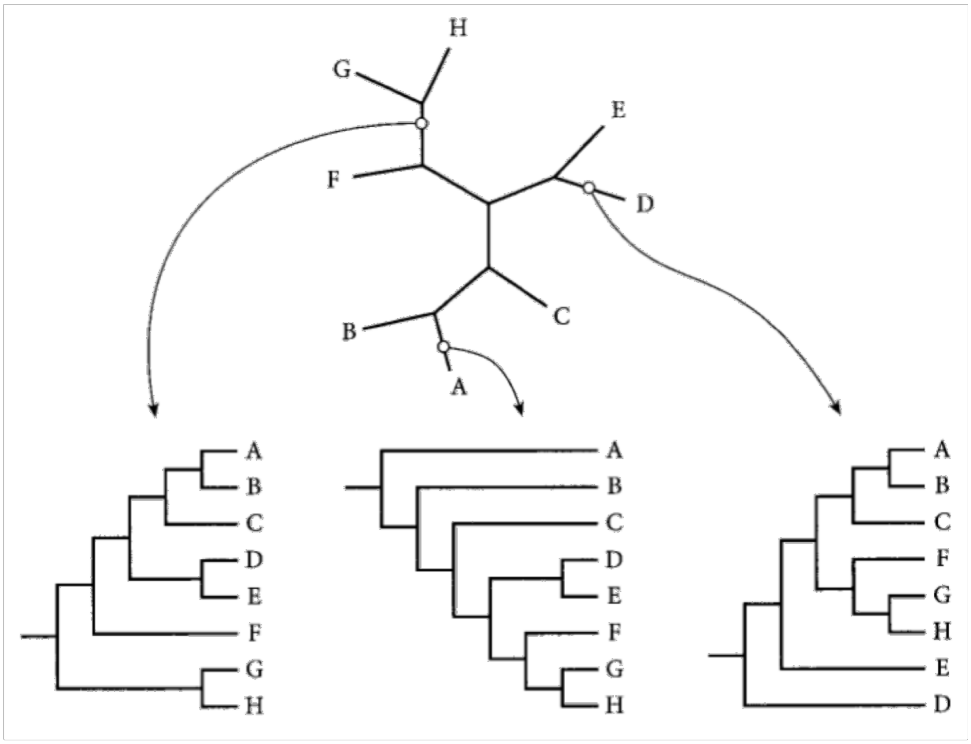
Находим на дереве самый длинный путь от листа к листу и за корень принимаем середину этого пути



Используя внешнюю группу (outgroup)

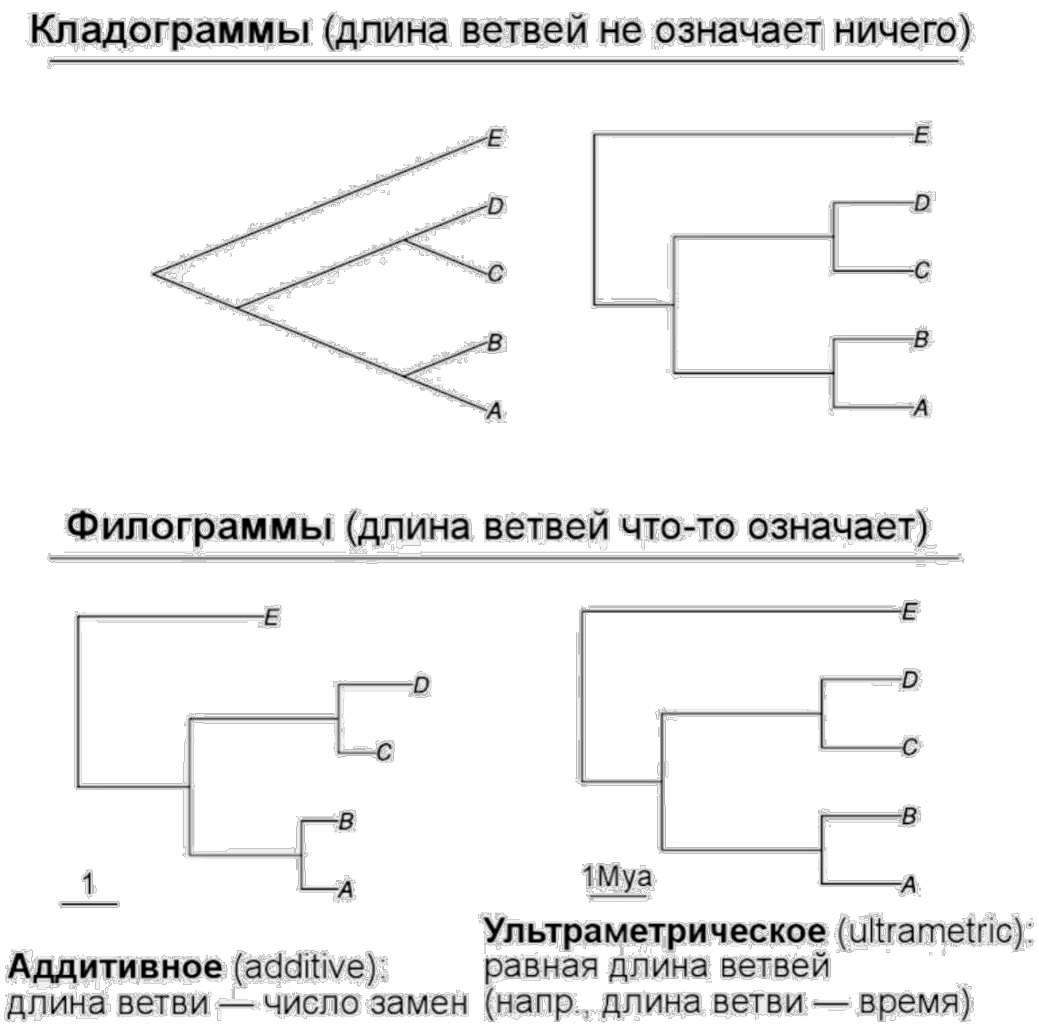
В данном случае укоренено дерево четырех растений, для чего пришлось строить дерево с участием внешней группы – трех животных в синем круге



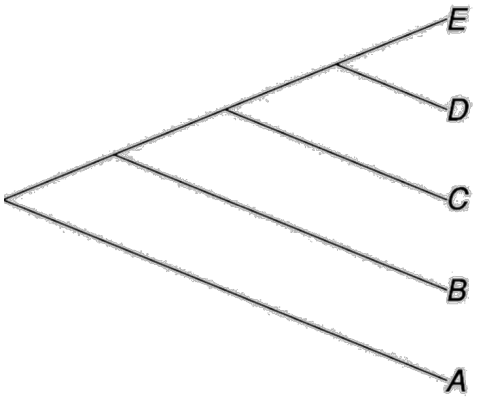


Пример укоренения

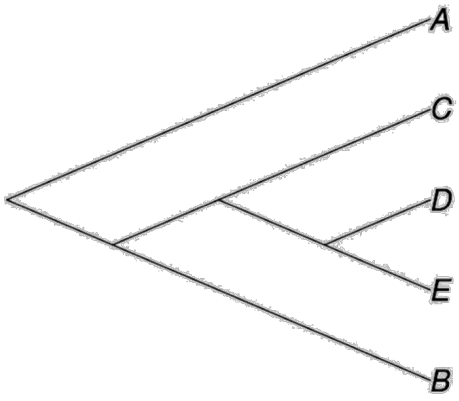
В свете различий филогенетики и кладистики различают *филограммы* и *кладограммы* соответственно



Одна и та же топология дерева может быть изображена по-разному. Проблема в том, что кодироваться они будут по-разному, хотя и отображают один и тот же эволюционный сценарий.

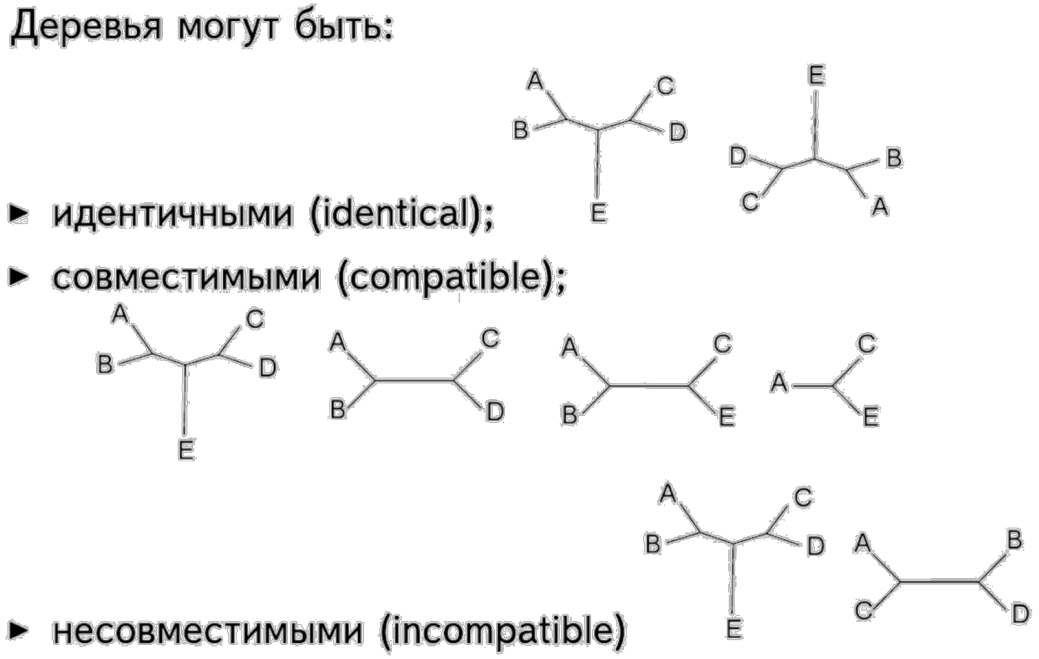


(A,(B,(C,(D,E))));



((B,((E,D),C)),A);

Как мы поняли, деревья, построенные по одним и тем же данным, могут быть идентичными и неидентичными. Кроме того, если мы представим себе чуть более сложную ситуацию. Если мы сравнивали немного разные данные, и у нас получились немного разные концевые данные деревьев, то можем ли мы сравнивать их топологию? Да. Но тогда нужно ввести понятие совместимости и несовместимости.
Совместимые деревья – такие деревья разного размера, топология ветвей которых совпадает. Есть процесс, название которого не имеет аналогов в русском языке и называется pruning. Это имеет прямое отношение к обрезке плодовых деревьев, а формально означает создание более маленького дерева из более большого с сохранением топологии (ну допустим, нам нужен лишь фрагмент от всего дерева).



Также деревья можно оценивать по разным критериям:

Оценка расстояния между деревьями

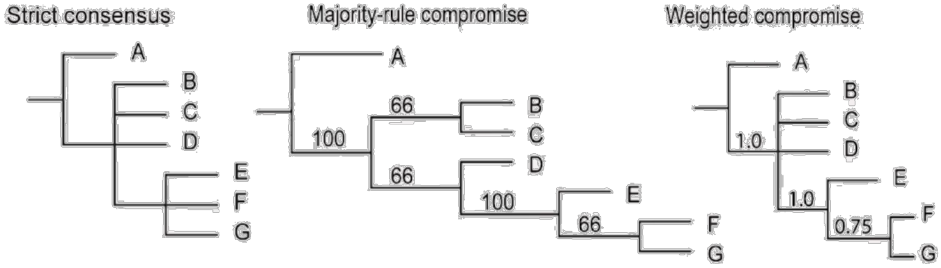
- Число общих клад
- Минимальное число SPR (subtree pruning reinraphment)

Прямая оценка качества дерева

- Индекс состоятельности (consistency index, CI)
- Индекс гомоплазии (homoplasy index, HI)
- Индекс удержания (retention index, RI)
- Decay index или Bremer support

Равно как мы вырезали участок дерева из большого, можно проводить обратное действие, то есть суммировать деревья.

- строгий консенсус (strict consensus);
- правило большинства (majority rule);
- комбинация (взвешенные варианты);



Sharkey MJ *et al.* (2013). Weighted compromise trees: a method to summarize competing phylogenetic hypotheses. Cladistics, 29(3), 309-314.

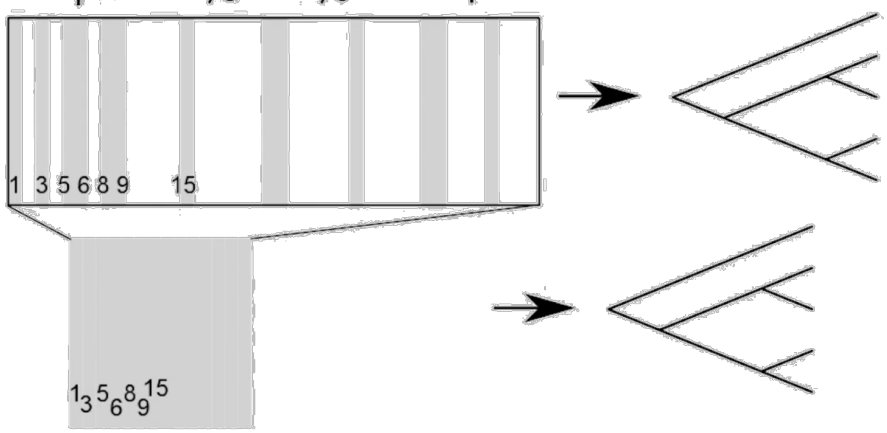
- филогенетические сети.

Вернемся к методам численной оценки качества дерева. По сути наше дерево – лишь точечная оценка. Вероятность, что мы идеально восстановили топологию дерева довольно мала. Поэтому, неплохо, если мы восстановили правильно хотя бы часть узлов. Основная идея оценки – псевдорепликация. В чем суть: выборки у нас нет, есть только точечная оценка в виде одного дерева, но иметь выборку очень хочется. Первым исторически был предложен метод «складной нож». Суть его в том, что мы из нашего выравнивания убираем от 1% до половины позиций. У нас остается какое-то выравнивание по-меньше и строим дерево. И так много раз. Далее смотрим, в каком проценте деревьев сохраняется конкретная клада. Это и есть поддержка конкретной клады.

Jackknife

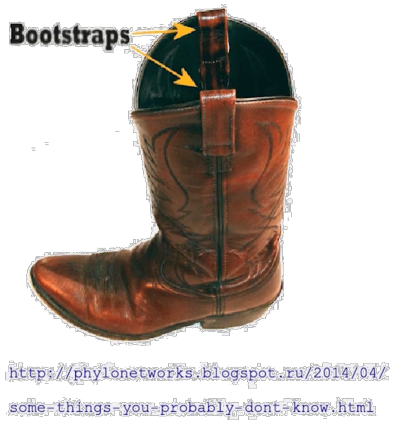
http://www.wisegeek.com/what-is-a-jackknife.htm

Убираем 1%–50% позиций



Наиболее же часто используемый метод – bootstrap (бутстрэп). Суть – та же псевдорепликация, выборка того же объема, но с повторяющимися элементами. Происхождение слова довольно интересное. Бутстрэпы – вот такие язычки на картинке с сапогом, а в современности такие есть на пятке кроссовок. В английском языке есть выражения типа to pull somebody by one's bootstraps, почти как Мюнгхаузен вытягивал себя за волосы. Похожее мы делаем с псевдорепликацией – когда мы создаем выборку на пустом месте, как бы вытаскивая себя за волосы

Bootstrap



http://phylonetworks.blogspot.ru/2014/04/some-things-you-probably-dont-know.html

Варианты анализа на основе бутстрепа

- непараметрический (non-parametric);
- Shimodaira–Hasegawa (SH) test;
- Approximate-Unbiased (AU) test;

Другие варианты статистического анализа топологии:

- fast / rapid bootstrap;
- approximately likelihood ratio test.

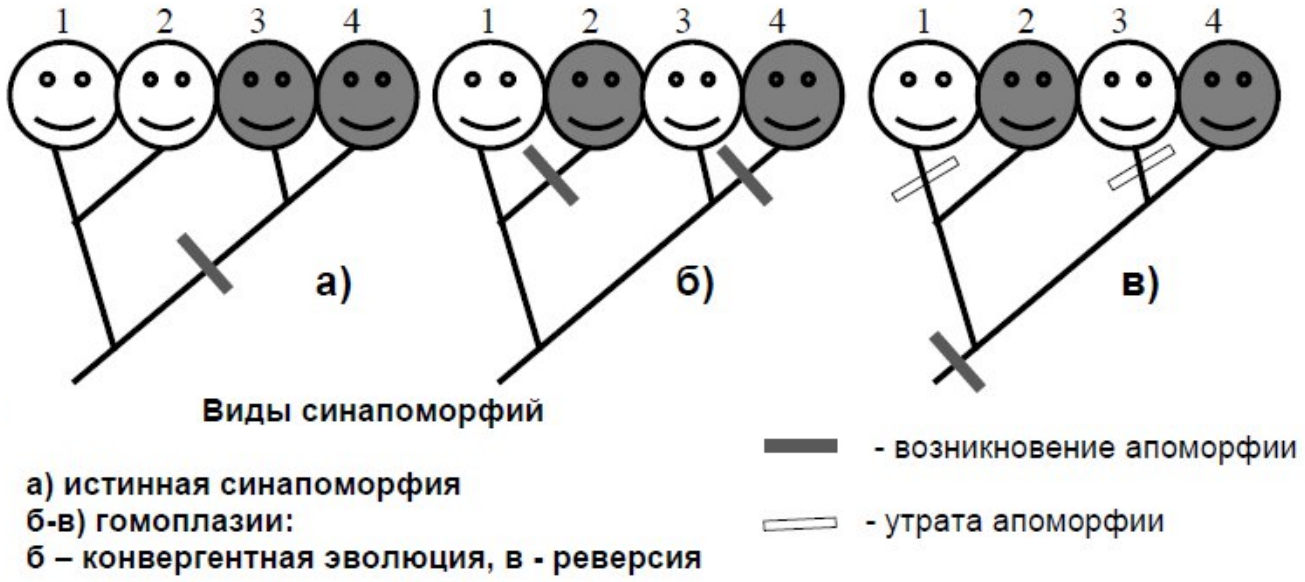
+ апостериорная вероятность

Признаки и типы групп организмов

Вспомним самые азы из школьной общей биологии.
Признак – совокупность состояний определенного атрибута исследуемых объектов. Например, цвет глаз - признак, а голубой – его *состояние*. Однако, очень часто под термином «признак» одновременно понимают и его состояние

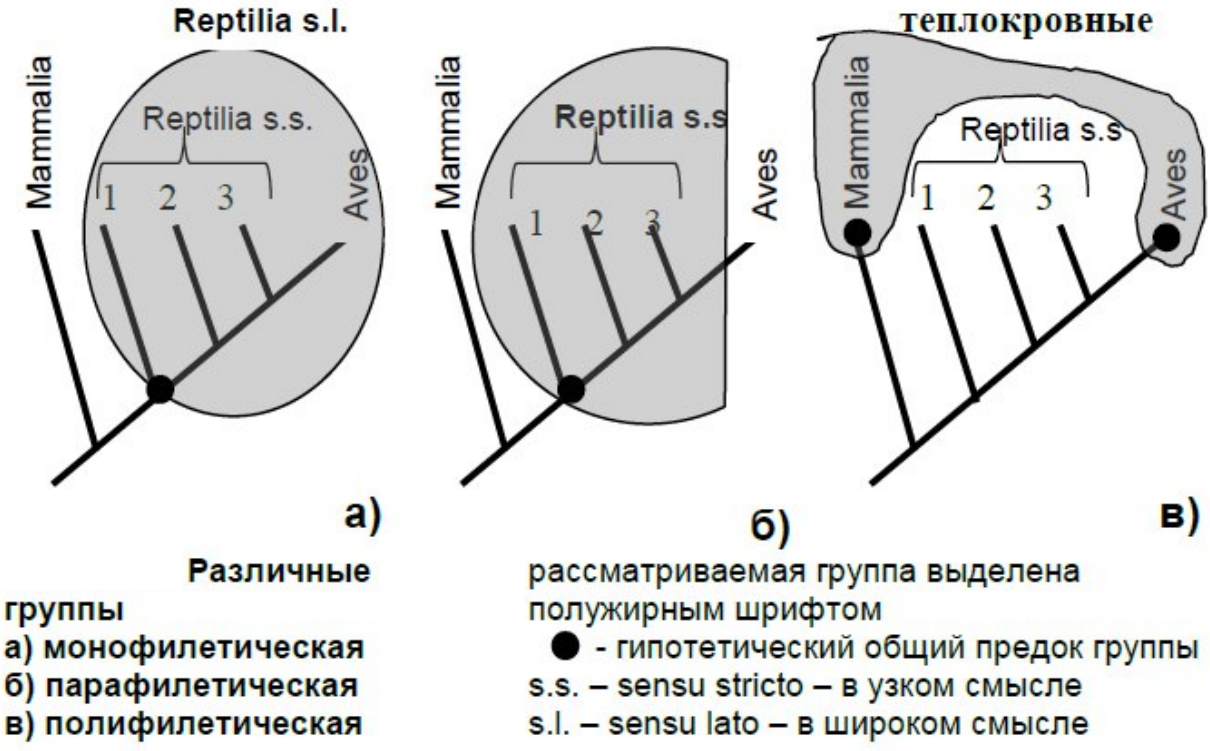
В филогенетике признак может иметь следующие состояния:
Плезиоморфное (предковое, базальное) – исходное состояние признака, которое характерно для гипотетического общего предка
Апоморфные (приобретенные) – производные состояние признака. Как я и говорил, самым похожим на общего предка является тот, кто ответвился прямо от него и на дереве не имеет ветвлений. Еще такая группа называется внешней и служит для калибровки дерева.

Внимание, очевидность: сходство в эволюции может достигаться двумя путями: общностью происхождения и конвергентным развитием, приведшими к эволюции различных структур подобным образом. В первом случае сходные структуры называют гомологичными, во втором – аналогичными. Восстановление филогенеза представляет собой процесс выделения в сходстве полезной информации о гомологиях и шума в виде аналогий.
Еще немножко терминологии: если организм конвергентно достиг одного внешнего признака с другим организмом, то такое явление на древе по этому признаку называется ***синапоморфией***. Синапоморфия может быть:
Истинной, то бишь образующие ее апоморфии гомологичны (синапоморфия возникла в эволюции единожды) и все потомки гипотетического предка, у которого она возникла, сохранили ее
Ложной (гомоплазия) – все иные случаи, образовавшиеся вследствие конвергентной эволюции, то есть апоморфии являются аналогичными (возникли в эволюции несколько раз) *или* реверсии (обратного перехода) апоморфного признака к плезиоморфному.



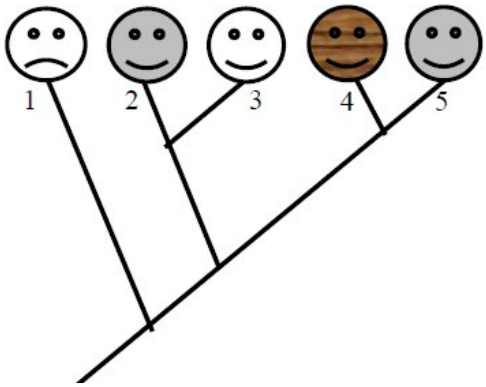
Объекты, рассматриваемые в филогенетике, на основании их признаков можно сгруппировать в группы. Используются в обиходе следующие:

- ***монофилетическая*** (клада) – группа включает всех потомков гипотетического общего предка. Пример на рисунке- Рептилии s.l
- ***парафилетическая*** – группа включает не всех потомков такого гипотетического общего предка. Пример на рисунке - Рептилии s.s
- ***полифилетическая*** – группа включает потомков различных гипотетических предков. Пример на рисунке - гомойотермные

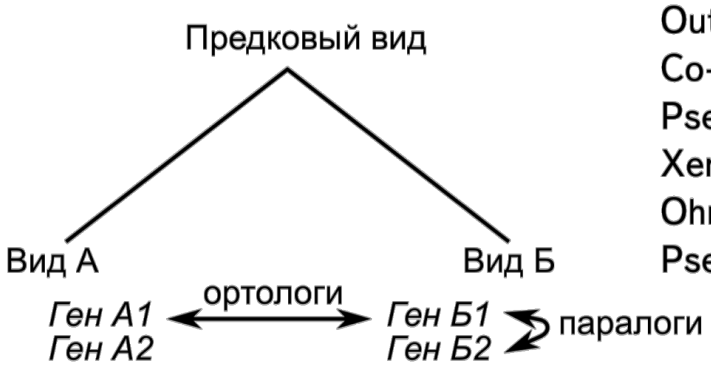


Группы, отходящие из одного узла, называются ***сестринскими***.

Предлагаю задачу на подумать: дано дерево со смайликами (самая последняя картинка) со следующими признаками:
белое/серое/деревянное лицо и грустная/веселая эмоция. Необходимо определить что из предложенного является базальными признаками, а что апоморфными. Так же определите какие признаки являются синапоморфными.



Типы отношений генов (варианты гомологии)



Inparalogs (synparalogs) /
Outparalogs (alloparalogs)
Co-orthologs
Pseudoothologs
Xenologs
Ohnologs
Pseudoparalogs

см. Koonin EV. Orthologs, paralogs, and evolutionary genomics 1. Annu. Rev. Genet. 2005;39:309-38.

Гомология - сходство признаков, обусловленная общим происхождением
Дивергенция – расхождение признаков в ходе эволюции
Гены *ортологи* (orthologs/orthologues) – гомологичные гены в разных геномах
Гены *паралоги* (paralogs/paralogues) – гомологичные гены внутри одного генома

Ортологи приобрели независимость при видообразовании.
Если дупликация генов произошла после видообразования, то есть у предкового вида был один ген, а у дочернего их стало два – такие гены *инпаралоги (синпаралоги).*
Если до – *аутпаралоги (аллопаралоги).*
На схеме гены А1+А2 по отношению к генам Б1+Б2 являются *ко-ортологами*.
Онологи – частный случай паралогов, которые образовались в результате полногеномной дупликации (назван в честь Сусумо Оно).
Ксенологи – гены, образовавшиеся в результате горизонтального переноса.
Также, если на схеме потеряются гены А1 и Б2, то оставшиеся будут называться *псевдопаралогами*, но изучая их, мы возможно сочтем их ортологами

Секвенирование нуклеиновых кислот

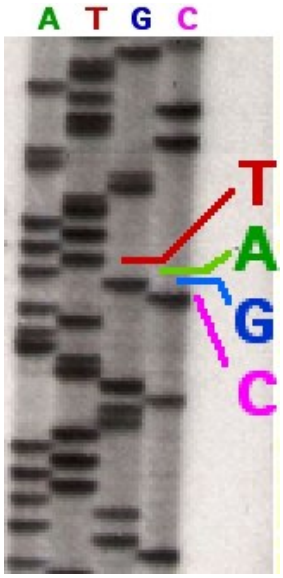
Все вы помните, что кодирование наследственной информации полностью лежит на нуклеиновых кислотах. Общее строение молекул DNA или молекул RNA одинаково, то есть фосфатный остаток, одна из пентоз и азотистое основание. Все вместе это образует нуклеотид, а за саму информацию, необходимую для дальнейшего синтеза белка несут в себе собственно азотистые основания, которые в двойных спиралях образуют комплементарные пары, согласно правилу Чаргаффа. После открытия спирального строения DNA Уотсоном и Криком, была сформулирована центральная догма молекулярной биологии, научились читать последовательности белков, а вот саму наследственную информацию читать не могли. Открыто это все было

сравнительно недавно - во второй половине XX века. Собственно метод секвенирования был придуман Фредериком Сенгером (он единственный в истории получил сразу две Нобелевские премии по химии, что случается невероятно редко: за чтение аминокислотных последовательностей и, собственно, секвенирование). Сегодня же мы с вами разберемся, как вообще можно прочитать геном организма, чтобы в дальнейшем с ним работать. Задумайтесь, что задача примерно такая же как на БАК - повлиять на микромир из нашего макромира.

Для определения первичной структуры DNA используют один из двух методов: метод химического секвенирования, предложенный Максамом и Гилбертом, или ферментативный, разработанный Сэнгером и Коулсоном. Оба метода позволяют секвенировать фрагменты DNA размером до 600 нуклеотидов. Для определения нуклеотидной последовательности более крупных фрагментов DNA необходимо получить набор перекрывающихся фрагментов, в совокупности соответствующих длине полной DNA. Секвенирование целых геномов проводят с использованием автоматических секвенаторов, в которых вместо традиционного радиоактивного мечения применяют флуоресцентную метку. Особое значение имеет сравнение результатов секвенирования с уже имеющимися компьютерными базами данных.

Метод Сэнгера-Коулсона

Для того чтобы получить исследуемый фрагмент DNA в однонитевой форме, его встраивают в вектор, созданный на основе генома фага M13. После выделения одноцепочечной DNA достраивают вторую цепь с помощью фрагмента Клёнова или T7-DNA-полимеразы. В реакционную смесь добавляют dATP, dTTP, dGTP, dCTP и короткий синтетический олигонуклеотид – праймер. Пробу разделяют на 4 части, и в каждую добавляют один из четырех дидезоксинуклеотидов ddATP, ddTTP, ddGTP, ddCTP – для обрыва цепи (после присоединения дидезоксинуклеотида рост цепи прекращается). Концентрацию дидезоксинуклеотидов подбирают таким образом, чтобы полученная смесь представляла собой статистический набор олигонуклеотидов разной длины, заканчивающихся на 3’-конце нуклеотидом, комплементарным соответствующему нуклеотиду в исходной матрице. Разделение продуктов реакции по длине с помощью электрофореза в полиакриламидном геле (ПААГ) позволяет прочитать всю последовательность нуклеотидов исследуемого фрагмента. Для того чтобы результаты секвенирования можно было зафиксировать на рентгеновской пленке, один из дезоксинуклеотидов, вводимых в реакционную смесь, метят радиоактивным изотопом: ^{32}P или ^{35}S



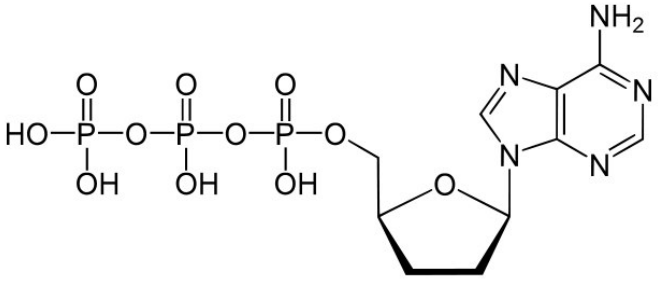
Метод Максама и Гилберта

В современных исследованиях метод химического секевенирования DNA применяется достаточно редко, однако ранее он широко использовался благодаря своей простоте и надежности. Метод заключается в химической модификации в четырех разных реакциях одного из четырех оснований DNA и последующем расщеплении сахарофосфатной цепи в местах модификации. При этом образуется набор меченных молекул разной длины, которые разделяют в ПААГ и читают последовательность нуклеотидов на полученной авторадиограмме. Использование твердофазного процесса и флуоресцентной метки позволило автоматизировать эту методику.

Автоматизация процесса секвенирования

В основе современных автоматических секвенаторов лежит метод ферментативного секвенирования по Сэнгеру, в который внесены следующие изменения:

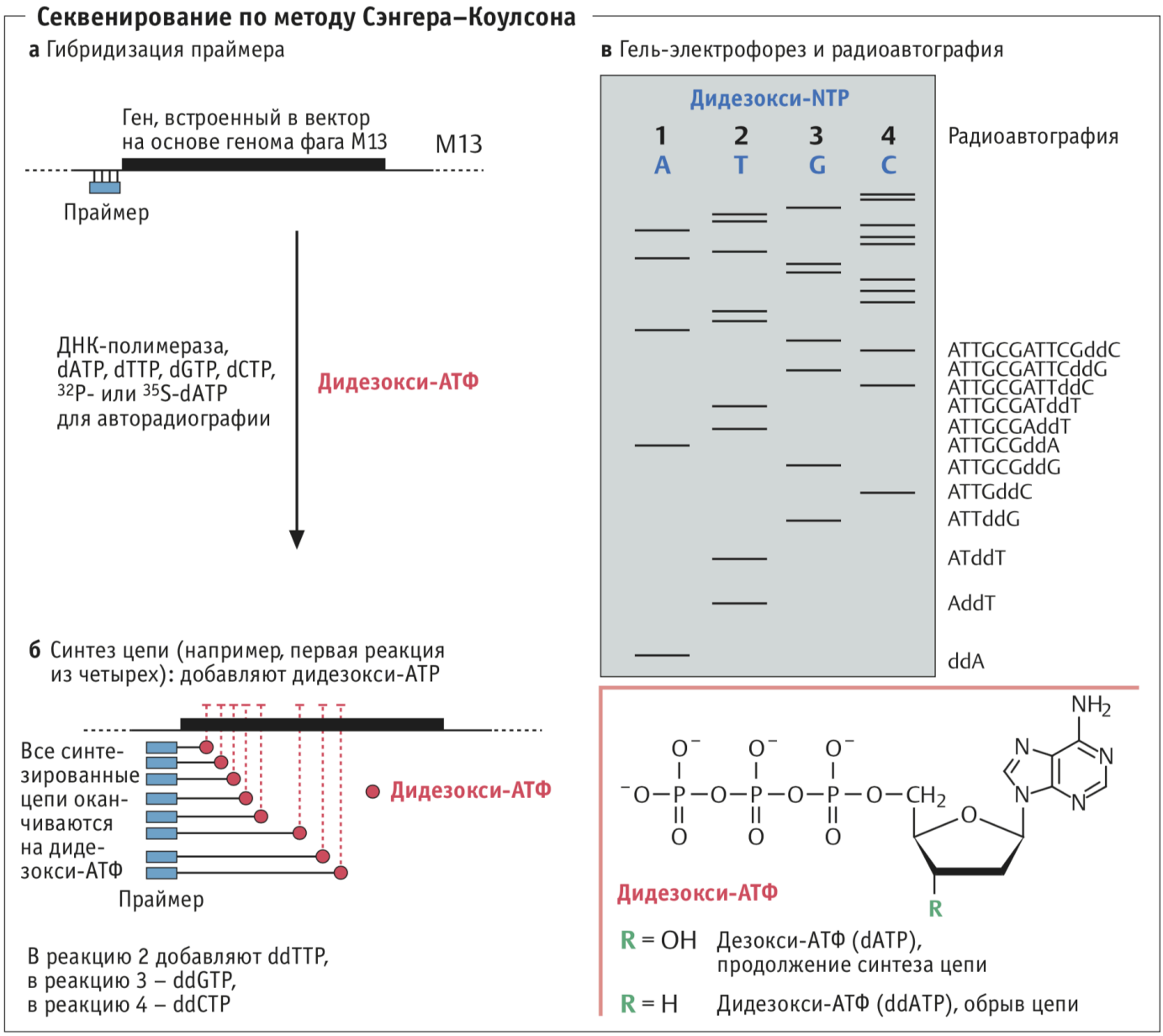
- Для получения однонитевой DNA проводят реакцию, схожую с ПЦР (асимметричная ПЦР)
- Мечение четырех наборов олигонуклеотидов после наращивания праймера осуществляется не с помощью радиоактивной метки, введенной в дезоксинуклеотиды, а посредством флуоресцентной группы, присоединенной непосредственно к дидезоксинуклоетидам

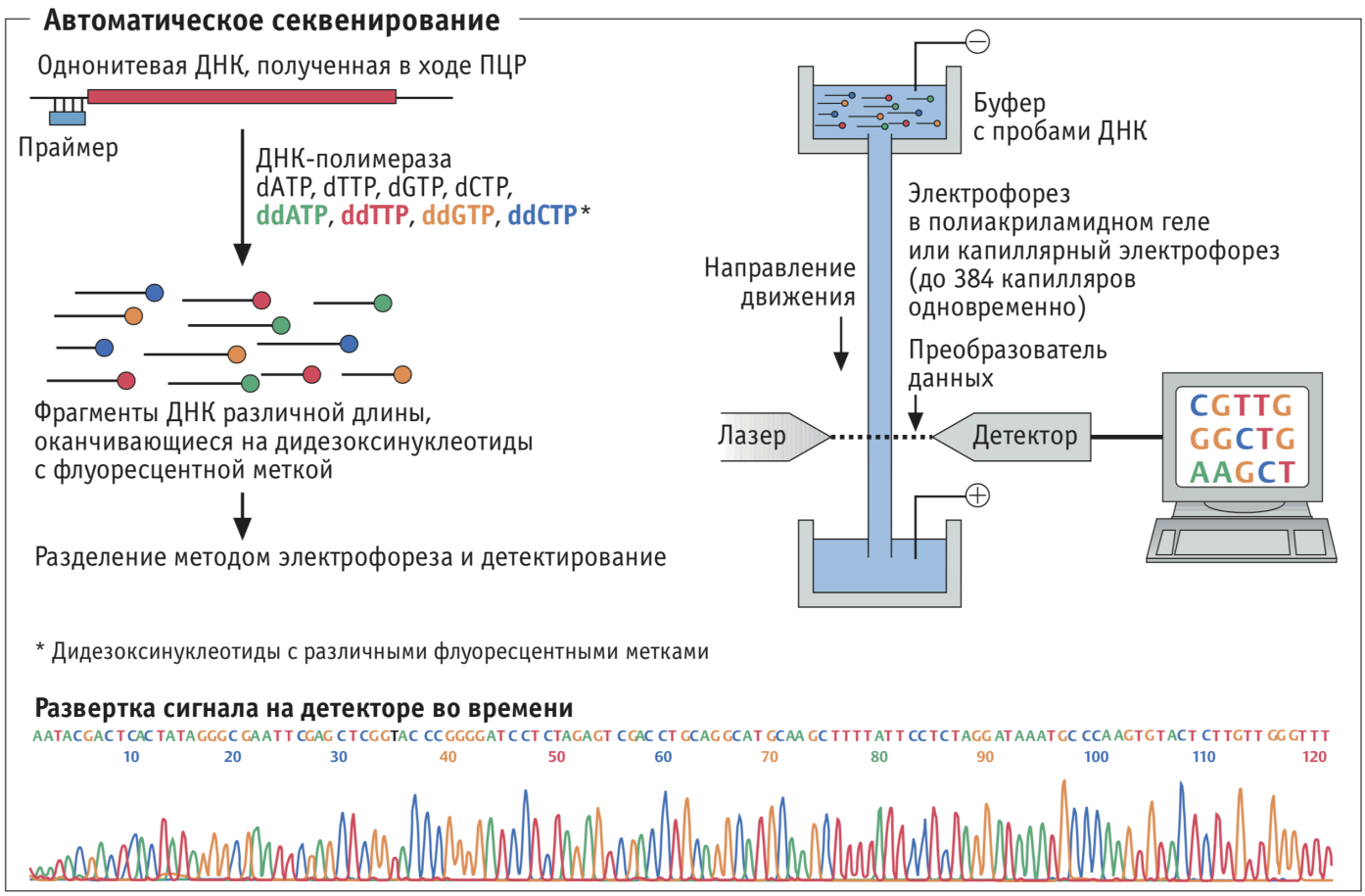


Дидезоксинуклеотид

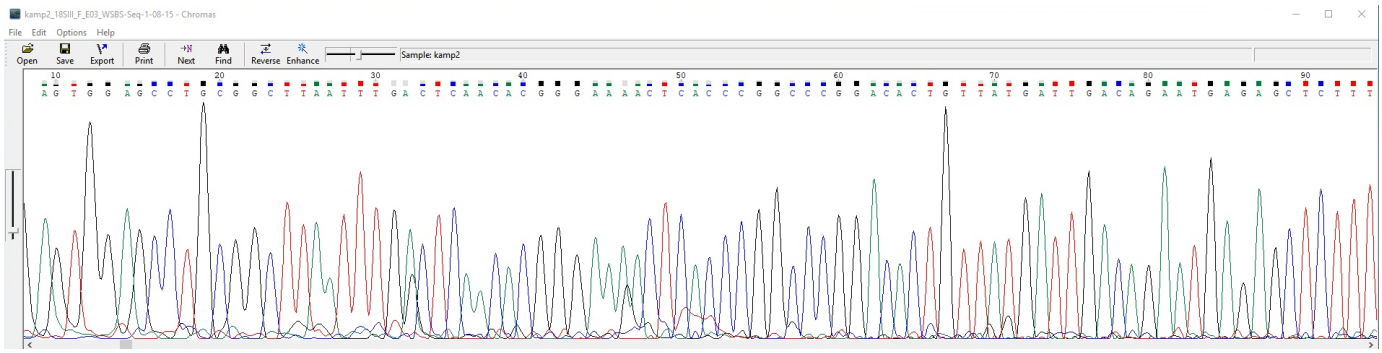
Если дидезоксинуклеотиды несут разные флуоресцентные маркеры, не требуется проведения четырех различных реакций наращивания праймера, и после разделения фрагментов DNA по размерам можно считывать сразу всю последовательность нуклеотидов исследуемой DNA. Размер фрагмента DNA может составлять до 900 пар нуклеотидов, время одного цикла – примерно 13 часов. Если в автоматическом секвенаторе параллельно проводится 96 проб, эффективность процесса достигает 10 000 п.н. за 15 часов. Использование капиллярного электрофореза еще больше увеличивает скорость анализа DNA. Секвенирование 96 проб DNA размером около 650 п.н. в капиллярах занимает всего 3-4 часа и, следовательно, эффективность анализа составляет 400 000 нуклеотидов в день. В настоящее время разрабатываются приборы, содержащие 384 капилляра. Для точного установления нуклеотидной последовательности следует повторить процесс несколько раз, однако даже с учетом многократного повторения секвенирования современные «фабрики», проводящие массовые анализы DNA с использованием автоматических анализаторов и компьютеров, позволяют за короткий срок секвенировать целые геномы. Геном - колоссальной длины текст. У человека геном состоит из 3 млрд нуклеотидов и все это нужно прочесть.

Для интересующихся прилагаю ссылку к Википедии про современные методы секвенирования.
https://ru.wikipedia.org/wiki/Методы_секвенирования_нового_поколения





На выходе такого метода мы получаем хроматограмму, которая выглядит примерно так:



Картинка, кстати взята из реальной жизни. Это фрагмент 18S-рибосомальной RNA Loxosomella murmanica с Беломорской биологической станции МГУ. Чтобы не оставлять все на этой пока что малопонятной хроматограмме, немного поясню, что с ней происходит дальше.

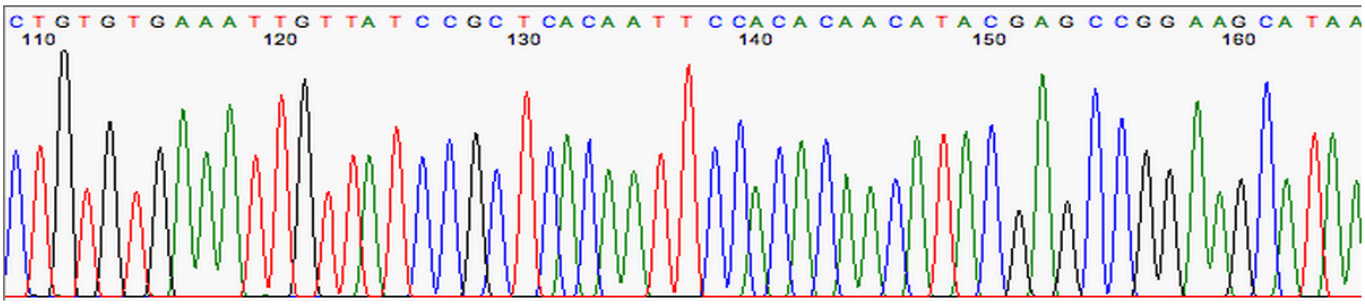
Софт секвенатора выдает последовательность (процедура называется “base calling”). Программы не всегда выдают правильный ответ – человек должен проверить результат и оценить качество хроматограммы в целом, соответствует ли качество решаемой задаче

Задача 1: проверить есть ли определенный ген – тут процент ошибок секвенирования не так важен

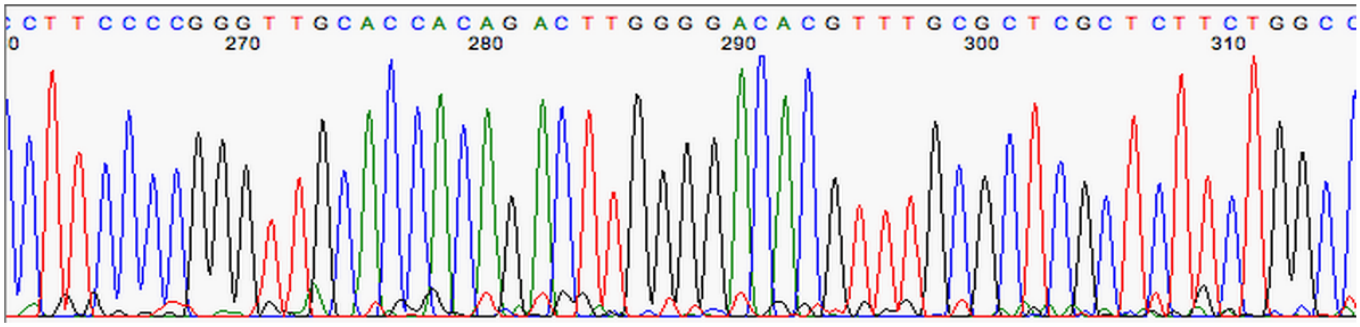
Задача 2: проверить принадлежность организма виду – ДНК штрих-кодирование (barcoding) – требуется отсутствие ошибок

Далее мы должны проверить все места возможных ошибок автоматического определения нуклеотидов и прервать чтение там, где ошибок становится слишком много.

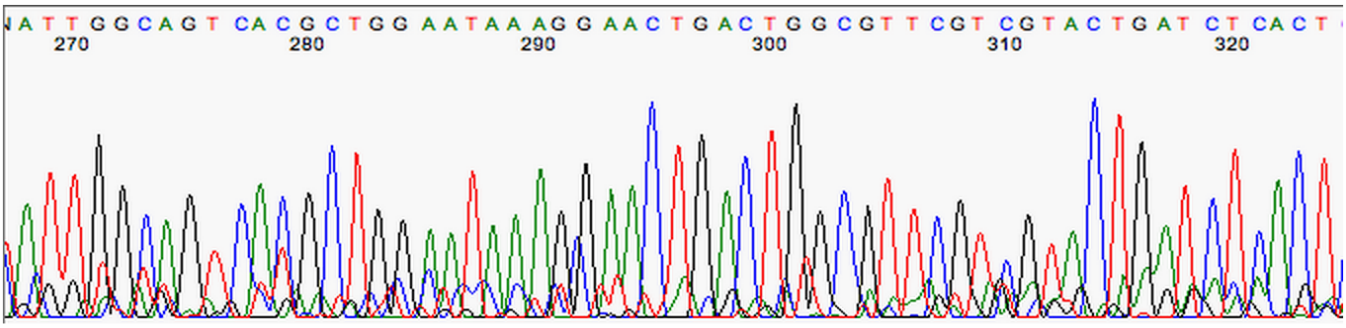
Идеальная хроматограмма



Более реальная хроматограмма со слабым шумом



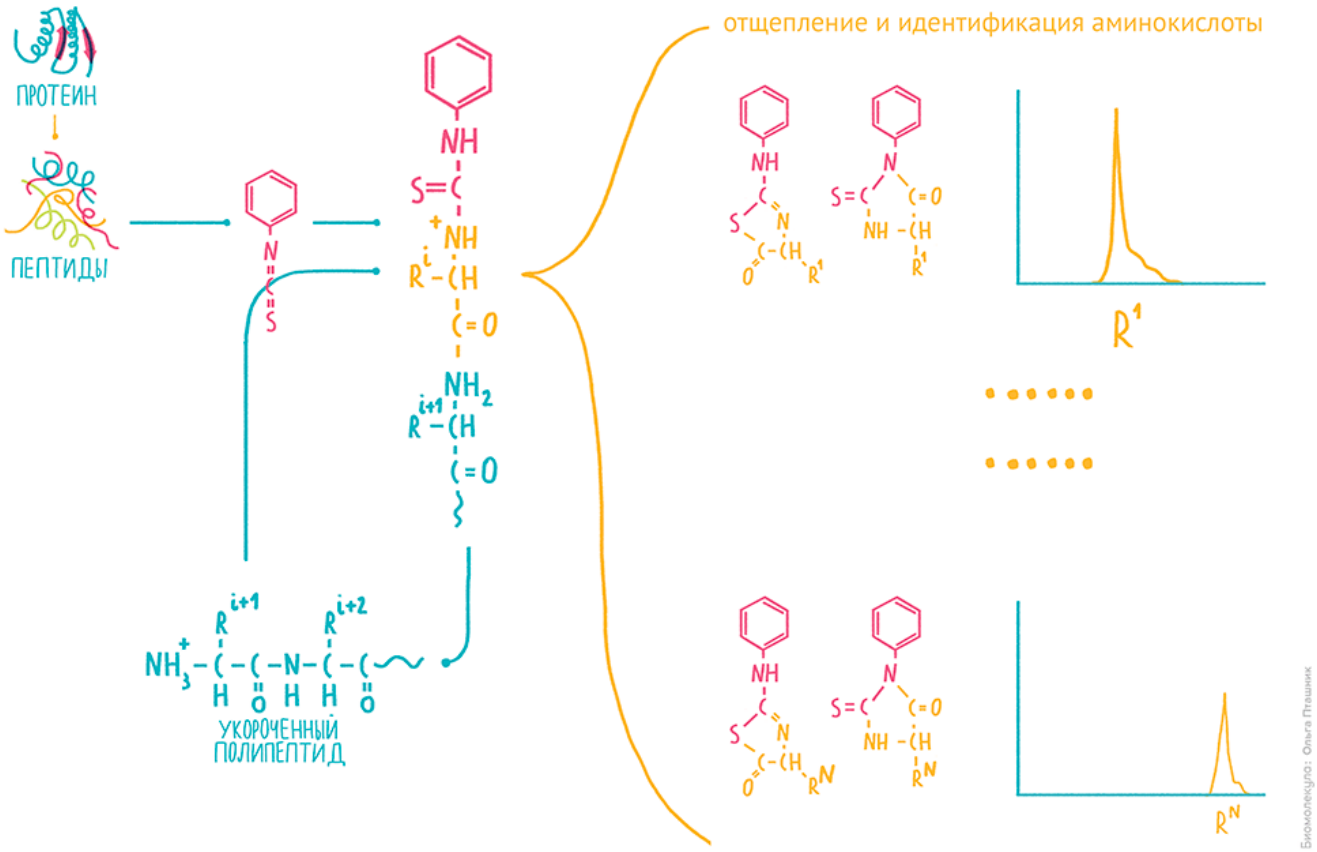
Сильный шум, создающий помехи



Собственно после обработки хроматограммы, программа может выдать нам файл в формате TXT или аналогичном ему.

Секвенирование белков

В 1950-е годы шведский химик Пер Эдман разработал метод секвенирования пептидов. Если обработать пептид изотиоцианатом фенила (ФИТЦ), электрофильный атом углерода на изотиоцианатном радикале при умеренном подщелачивании взаимодействует с нуклеофильным азотом незаряженной аминогруппы. В итоге на N-конце пептида образуется фенилтиокарбомоильный радикал. Если умеренно закислить реакционную смесь, он отщепляется, увлекая с собой N-концевую аминокислоту, с образованием тиазолинона со специфичным радикалом, характеризующим эту аминокислоту. При этом остальная часть аминокислотной цепи остается неизмененной. Особое производное, которое будет отличаться по присущему аминокислоте радикалу, еще раз преобразуют в кислых условиях — для стабилизации — и анализируют хроматографически. Так можно отличить такие производные для всех аминокислот, поскольку из-за характерного радикала они будут характеризоваться своим временем выхода с обращенной фазы. Если белок или пептид, который мы анализируем, присоединен к твердофазному носителю, производное N-концевой аминокислоты можно смыть и анализировать отдельно, а цикл анализа повторить, выстраивая таким образом аминокислотную последовательность.



Метод Эдмана был по тем временам очень прогрессивен. Он с высокой точностью предоставлял последовательность до 30 аминокислотных остатков. Характеризовался достаточно высокой чувствительностью, будучи способным секвенировать пептиды в количестве менее 0,1 нмоль с 99% точностью. Более того, в конце 1960-х его автоматизировали в виде пептидного секвенатора, где робот-раскапыватель поочередно снимал N-концевые производные с полипептидов, закрепленных на специальной бумаге, направляя их затем в хроматограф. Но исследователям опять хотелось большего — их не устраивала необходимость в очистке пептидов и белков перед секвенированием, а также некоторые другие ограничения эдмановского метода, в частности, его неспособность секвенировать продукты с модифицированным N-концом.

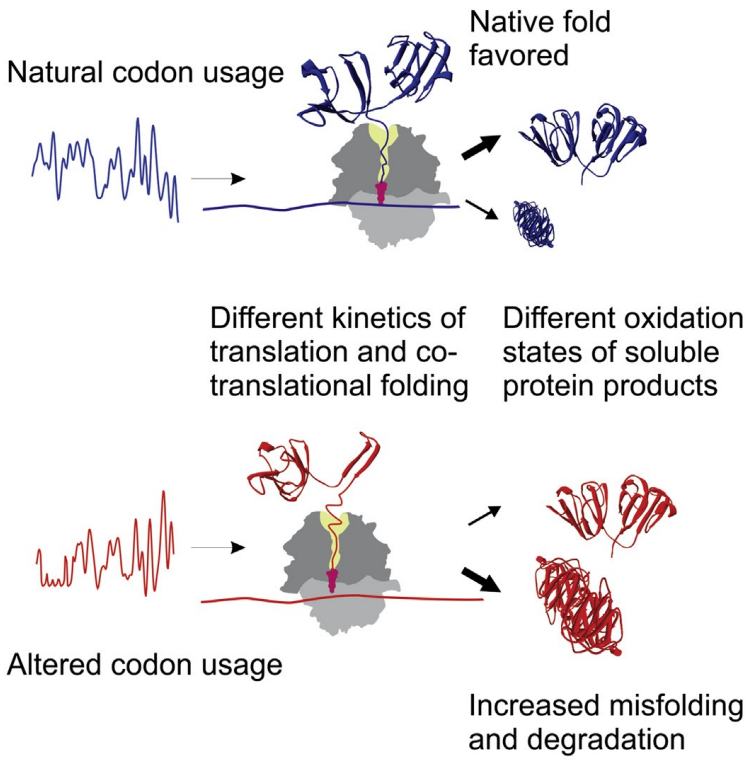
Небольшой интерес к методу Эдмана существует до сих пор, в особенности, для белков и пептидов тех организмов, последовательность которых нельзя предсказать из данных секвенирования нуклеиновых кислот. В этом методе реализуется прямой анализ, где ошибки связаны с технической погрешностью. Последовавшие за ним способы анализа аминокислотной последовательности содержат элементы предсказания, поэтому к техническим ошибкам в них прибавляются алгоритмические (см. ниже).

Появление последовательностей геномов множества организмов, начиная с бактерий и завершая большими геномами многоклеточных эукариот уменьшило пространство поиска при идентификации белков. За исключением ситуации с секвенированием общей ДНК сложной смеси организмов (так называемого *метагенома* почвы, содержимого кишечника, океанских вод и т.д.), биохимики обычно представляют, какой организм они анализируют. И это значит, что белки в исследуемом образце синтезированы при помощи потока информации с кодирующих их генов этого организма. Собственно, так и появился термин «*протеом*» — в 1994 году Марк Уилкинс, австралийский аспирант, предложил его для обозначения белкового, или протеинового дополнения к геному . Прочитанный геном породил остальные «-омы», а технологии, позволяющие их анализировать, в конце 1990-х годов почти гипотетические, составили группу *постгеномных* технологий.

Строго говоря, это анализ продуктов передачи геномной информации, то есть кодирующих и некодирующих РНК и белков. Остальные «-омы», по сути, косвенные. Они не связаны с генетическим кодом прямым

потоком информации и объединяются в группы по химической природе анализируемых соединений. Примечательно, что современные технологии производят одновременный анализ тысяч соединений, например, метаболитов, липидов, гликанов и т.д., и называются, соответственно, метаболомикой, липидомикой, гликомикой и т.д.

Вообще, фундаментальным вопросом применения всего этого является соотношения генотипа и фенотипа. В этом смысле выгоднее изучать последовательности белков, а не нуклеиновые кислоты. Их, во-первых, намного эффективнее выравнивать за счёт числа аминокислот. Во-вторых, у них вряд ли в результате множественных замен аминокислота вернётся на своё место. С другой стороны, их проблема в том, что разрешение не такое большое, как у НК. А также, всё-таки нуклеиновые последовательности могут меняться, а белок при этом оставаться прежним (синонимичные замены). На этом моменте мы просто обязаны вспомнить, что еще более важной является пространственная структура этих молекул. И что бы не говорил Лаймус Поллинг, у одной первичной структуры далеко не одна третичная, более того, даже синонимичные замены могут приводить к иному фолдингу белка в процессе трансляции, что может крайне изменить его функцию. Не смотря на все сложности, видимо, протеом более информативный, чем геном, но все еще не дает нам полной картины мира. Более того, не всегда представляется возможным найти и извлечь белок, особенно для метагеномного анализа.



Выравнивание

Биоинформатика последовательностей занимается изучением нуклеотидных/аминокислотных последовательностей (то есть первичной структуры) основных биополимеров — нуклеиновых кислот и белков. Их сравнительный анализ позволяет соизмерить сходство и установить соответствие между остатками, определить консервативные и вариабельные участки, высказать соображения об эволюционных взаимосвязях.

Наиболее рутинным подходом в этой области является выравнивание последовательностей — базовый биоинформатический метод, основанный на размещении двух или более последовательностей ДНК, РНК или белков друг под другом таким образом, что можно легко определить сходные/различающиеся участки в этих последовательностях.
Чтобы интуитивно понять, зачем это нужно, приведем примеры. Принимаем во внимание, что символы в последовательностях переставлять нельзя (поскольку это будут уже другие последовательности).

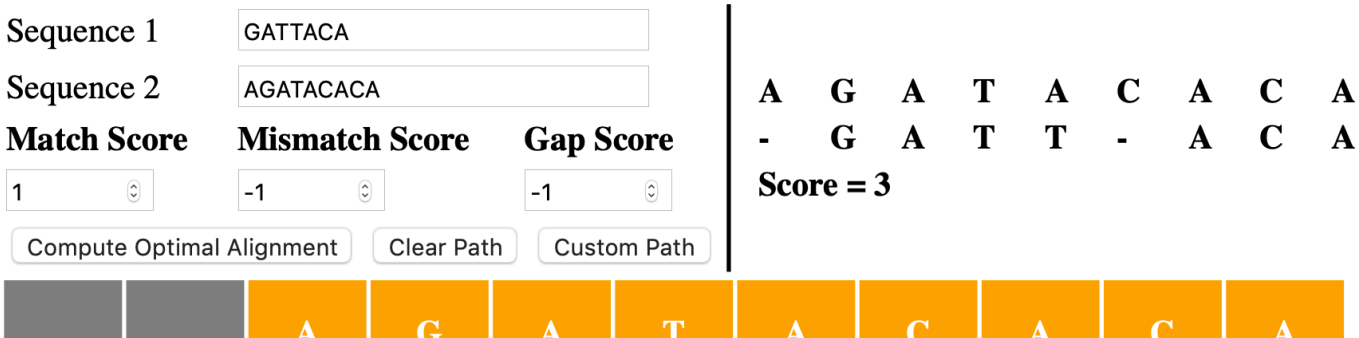
Последовательности	Варианты выравнивания		
ГОРОД ОГОРОД ГОРОДОК	_ГОРОД__ оГОРОД__ _ГОРОДок		
ВООБРАЖЕНИЕ СОДЕРЖАНИЕ	вОобРажение сОдеРжание-	воОбраЖеНИЕ _сОдерЖаНИЕ	вОобРаЖеНИЕ сОдеР-ЖаНИЕ
GTATAGTCTA GTTAGTAGTC	GTaTAGTc_Ta GT_TAGTagTc		GT_A_TAGTCta GTtAgTAGTC__

Становится понятно, что для анализа любых последовательностей с одинаковыми участками, пусть то слова или геномы, нужно провести чисто визуальную операцию выравнивания, где мы подстраиваем друг по дружку эти самые одинаковые части. И тогда между символами возникает три возможных типа отношений: они одинаковы, они различаются, один есть и другой отсутствует. Соответственно в нашем случае это какая-то мутация или ее отсутствие. Кстати, как можно было заметить, вариант с отсутствием показывается пропуском или нижним подчеркиванием.
Однако совершенно не очевидно, какое из этих выравниваний более оптимально (например, в третьем примере, в первом выравнивании больше совпадений, но во втором меньше пропусков). Вообще, количество вариантов выравниваний можно рассчитать с использованием формулы: $N = (a + b)! / (a! \times b!)$, где a и b — длины последовательностей. В случае примера, приведенного выше, $a = b = 10$, а $N = 184\,756$. Если последовательности будут состоять из 20 символов каждая, то N составит уже $1{,}38 \times 1011$, а при длине в 100 символов $N = 9{,}05 \times 1058$. Тут встают два вопроса: во-первых, какое выравнивание считать оптимальным, во-вторых, как его найти среди всех возможных вариантов (ведь очевидно, что простой перебор всех вариантов с использованием существующих вычислительных ресурсов займет годы даже для не очень длинных последовательностей)?

Выравнивание может быть *попарным* или *множественным*. В попарном участвуют две последовательности, а в множественном последовательностей 3 или больше. Поскольку алгоритмы множественного выравнивания основаны на алгоритмах попарного, сначала разберем именно попарное. Представим, что у нас две последовательности:

Все возможные выравнивания можно представить в виде таблицы. Одна последовательность по горизонтали, вторая – по вертикали. В каждой клетке будет возможное соотношение двух букв, а путь из одной клетки в другую будет собирать нам наше выравнивание. Из одной клетки мы можем двигаться в трех направлениях: по горизонтали вправо, во вертикали вниз и по диагонали между ними. Если мы движемся по диагонали, то последовательность выравнивается четко буква под буквой. А если мы вынуждены пойти по горизонтали или вертикали, будет формироваться пропуск, так называемый “gap”.

Какие вообще бывают подходы к выравниванию? Таких подходов три: *глобальное, полуглобальное и локальное*. Первое ищет лучшее выравнивание двух последовательностей от начала и до конца. Подходит, если последовательности сильно похожи и примерно одинаковой длины. Локальное выравнивание ищет похожие участки внутри двух последовательностей, и поэтому используем мы его если последовательности разной длины (например длинный геном, а ищем мы конкретный ген) или если последовательности сильно различаются, но что-то похожее в них все-таки есть. Полуглобальное подразумевает отсутствие штрафа за пропуски в конце. Об этом чуть подробнее потом. Нужен он, когда у вас последовательности перекрываются концом одной и началом другой.
Начнем мы с алгоритма глобального выравнивания, он же алгоритм Нидлмана-Вунша. Нарисуем табличку и введем понятия. Нам нужно объективно хорошее выравнивание, поэтому нужны формальные критерии. Введем бонус за совпадение букв последовательности (match) и штрафы за несовпадение (mismatch) или за пробел (gap). На сайте http://experiments.mostafa.io/public/needleman-wunsch/index.html можно посмотреть наглядную визуализацию этого процесса



		A	G	A	T	A	C	A	C	A
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
G	-1	-1	0	-1	-2	-3	-4	-5	-6	-7
A	-2	0	-1	1	0	-1	-2	-3	-4	-5
T	-3	-1	-1	0	2	1	0	-1	-2	-3
T	-4	-2	-2	-1	1	1	0	-1	-2	-3
A	-5	-3	-3	-1	0	2	1	1	0	-1
C	-6	-4	-4	-2	-1	1	3	2	2	1
A	-7	-5	-5	-3	-2	0	2	4	3	3

Из этой таблицы, чтобы собрать выравнивание, движемся в обратную сторону из правого нижнего угла. В реальности таблиц никто не строит, специалисты же пользуются методами динамического программирования. Также, обычно штраф за пропуски делают гораздо больше, чем -1, чем штраф за несовпадение, так как замена нуклеотида более вероятна, чем инсерция или делеция. Также, штраф за гэп дается только за начало гэпа, но не продолжение.
Основной алгоритм локального выравнивания – алгоритм Смита-Вотермана

http://github.com/haruosuz/books/tree/master/blast

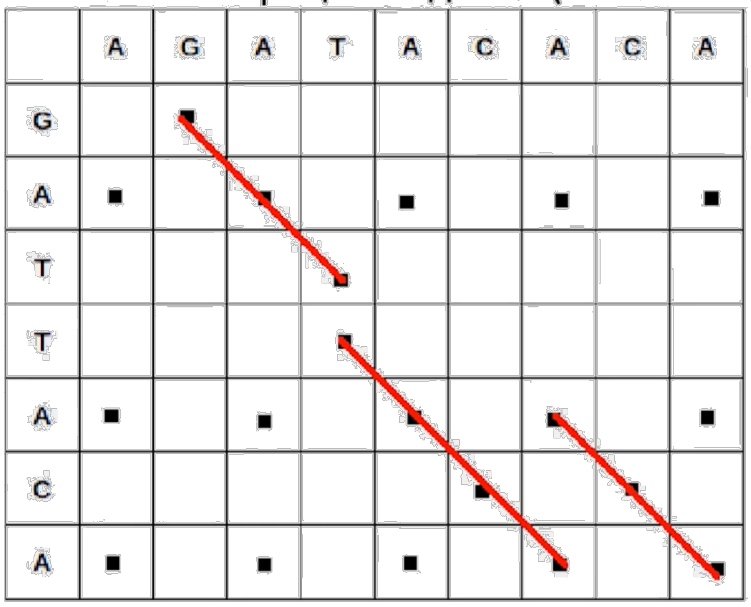
		C	O	E	L	A	C	A	N	T	H
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0	0
L	0	0	0	0	2	1	0	0	0	0	0
I	0	0	0	0	1	1	0	0	0	0	0
C	0	1	0	0	0	0	2	1	0	0	0
A	0	0	0	0	0	1	1	3	2	1	0
N	0	0	0	0	0	0	0	2	4	3	2

Идея та же самая, разница в заполнении таблички. Как только какое-то значение должно стать меньше нуля, оно обнуляется и меньше нуля стать не может. Стоит отметить, что оба этих метода – точные, а значит времязатратные. Существует много эвристических методов, которые позволяют экономить время. Самые популярные: BLAST и fastA

fastA = fastP + fastsN (fastAny = fastProtein + fastNucleotide)

Прежде чем подойти к fastA разберемся в том, что такое точечная матрица сходства. Берем таблицу и ставим точки там, где буквы в столбцах и строках совпадают. Далее соединяем по диагонали точки, если это возможно. Собственно алгоритм выбирает 10 лучших диагоналей, переоценивает их и соединяет в локальное выравнивание с пропусками.

Точечная матрица сходства (dot matrix)



http://www.gen.tcd.ie/molevol/fasta.html
http://www.ebi.ac.uk/Tools/sss/fasta/

Поговорим немного о BLAST (Basic Local Alignment Search Tool). Здесь важно понимать, что это самостоятельный алгоритм, а не средство поиска внутри NCBI

BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Варианты BLAST

Вариант BLAST	Последовательность в запросе	Последовательности в базе
blastn	Нуклеотидная	Нуклеотидные
blastp	Белковая	Белковые
blastx	Транслированная	Белковые
tblastn	Белковая	Транслированные
tblastx	Транслированная	Транслированные

NCBI Blast / WU-Blast

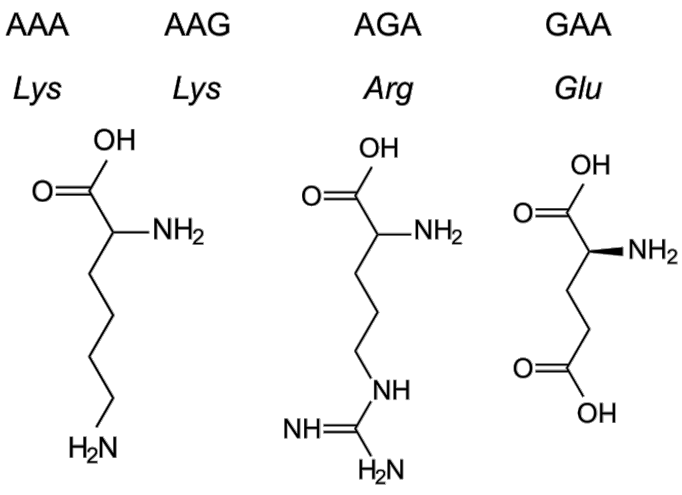
Как он работает? BLAST берет из последовательности, данной на вход какое-то слово. По умолчанию это 11 нуклеотидов или 3 аминокислоты. Далее он ищет это слово в базе, а затем увеличивает длину фрагмента влево и вправо, сужая круг поиска, пока качество выравнивания не упадет ниже определенного уровня. В конце нам выдается локальное выравнивания. Как объективно оценить качество? У обоих методов есть метрики качества выравнивания.

- Bit score — нормализованное (не зависящее от базы) значение score.
  $S' = \frac{\lambda S - \ln K}{\ln 2}$, где K и λ — параметры, позволяющие описать объём пространства поиска и систему оценки качества выравнивания.
- E-value — ожидаемое число случайных находок.
  $E = m \cdot n \cdot 2^{-S'}$, где m и n — длины сравниваемых последовательностей.

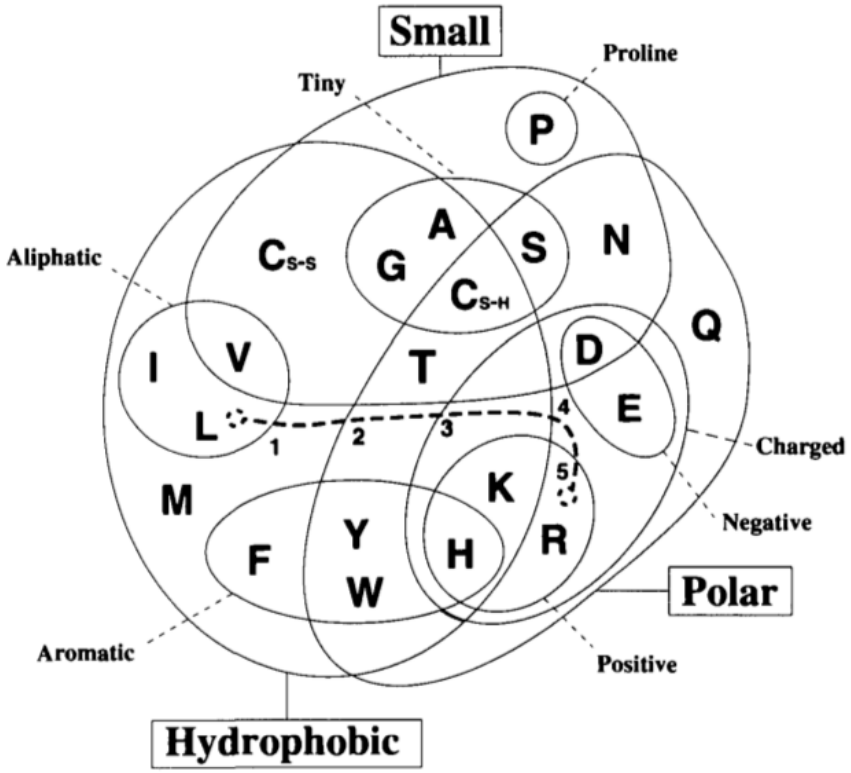
https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html
http://faculty.virginia.edu/wrpearson/fasta/fasta_guide.pdf

Понятно, основное что можно сказать о выравнивании – это его счет. Это штрафы и бонусы. Выравнивание можно считать хорошим, если E-value меньше 10^{-6}

Если мы работаем с протеомом, важно помнить, что не все аминокислотные замены одинаковые.



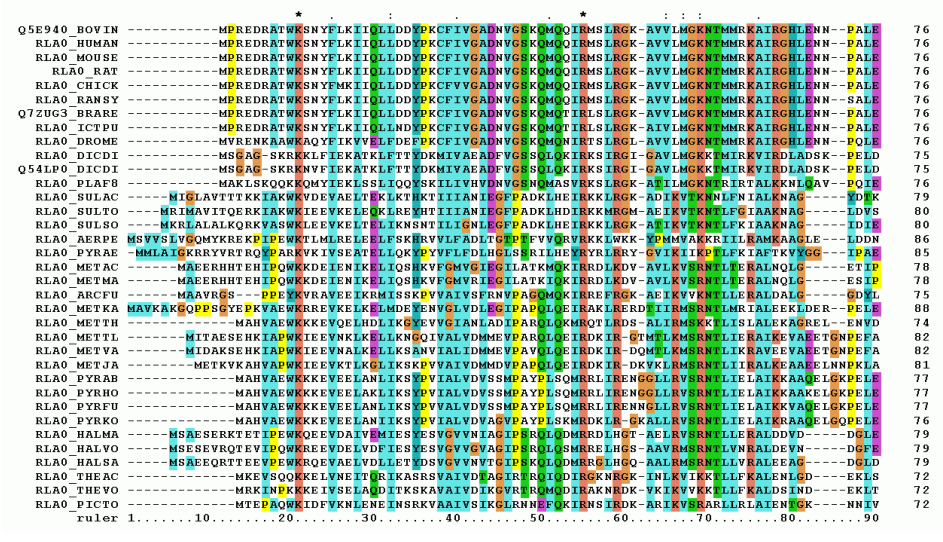
Здесь приведена классификация аминокислот по их важных физико-химическим свойствам, которые влияют на третичную структуру белка



Не вдаваясь в подробности, стоит сказать, что существуют алгоритмы выравнивания аминокислотных последовательностей, сравнивающие физическую значимость замен. Это, например PAM (point accepted mutation) – в нем чем больше номер, тем больше замен. Есть еще BLOSUM (BLOcks SUbstitution Matrix) – в нем чем меньше номер, тем больше замен.

Множественное выравнивание

При решении задач современной молекулярной биологии зачастую необходимо знать, насколько похожи/различны более чем две последовательности. В таких случаях используют не *парное выравнивание*, о котором речь шла выше, а *множественное*. Такое визуальное представление первичных структур белков и НК не только позволяет оценить степень их сходства, но и сделать предположения о том, каким путем шла молекулярная эволюция. Являет собой сумму многих попарных выравниваний, с некоторыми дополнениями. Более подробно будет рассмотрено в следующих версиях.



Базы данных

Нуклеотидные:

1. EMBL http://www.ebi.ac.uk/ena
2. NCBI (GENBANK) <= **E-utilities**
   https://www.ncbi.nlm.nih.gov/genbank/
3. DDBJ http://www.ddbj.nig.ac.jp/

Белковые:

- UniProt = Swiss-Prot + TrEMBL
  http://www.uniprot.org/

Базы для конкретных организмов, типов последовательностей *etc*:

- *Drosophila* http://flybase.org/
- rRNA https://www.arb-silva.de/
- SGD http://yeastgenome.org
- 1KITE http://www.1kite.org/
- Ensembl http://www.ensembl.org/index.html

Методы построения филогенетических деревьев

Наконец, мы добрались до построения деревьев. Задачка-то оказалась совсем математической. Чистой воды комбинаторика.
Прежде чем строить дерево, необходимо провести фильтрацию выравнивания, поскольку от его качества зависит точность нашего дерева. Поэтому нужно убрать часть данных из выравнивания – те, что плохо получились или сомнительные

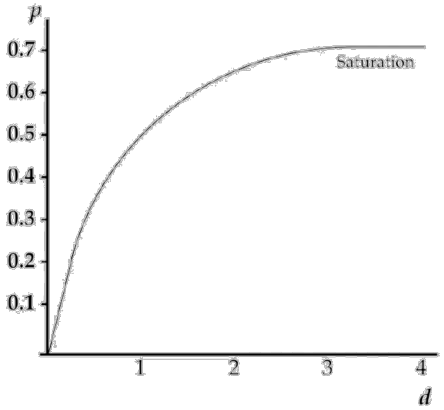
Фильтрация выравнивания перед построением дерева

Gblocks: популярный, но давно не обновленный и консервативный
*trimA*l: можно несколько выравниваний, лучше для большого объема данных
GUIDANCE2: только веб-форма, сам выравнивает
Aliscore: локальное качество, счет каждой позиции
AL2CO: только белки, чаще используют для оценки консервативности, а не для фильтрации
Zorro aka Probmask: тоже оценивает каждую позицию, более сложная модель оценки

Итак, мы получили хорошее множественное выравнивание. Как мы будем строить дерево? Вероятно, стоит определить, насколько последовательности похожи. Самое простое, что приходит в голову – посчитать количество сайтов с различиями и разделить на общую длину последовательности. Метрика довольно очевидная, а называется она p-distance

Определение расстояния между последовательностями

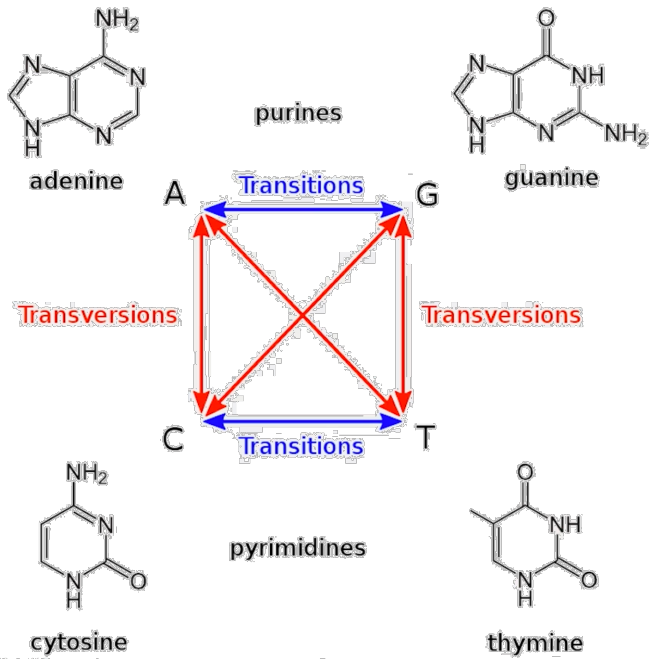
- p-distance (наблюдаемое расстояние)



Все это конечно хорошо, если бы не одна проблема – множественные замены. Если в последовательности DNA происходит замена, а потом здесь же происходит еще одна замена. В лучшем случае, мы узнаем только об одной из замен, в худшем (например A -> T -> A) не узнаем вообще о том, что замена была. График выходит на плато. То есть при повышении частоты замен наблюдаемое расстояние не будет значительно меняться. Как с этим бороться? С чем из статистики может быть связана частота замен? Распределение

редких событий (распределение Пуассона). Это несколько улучшает картину, но не полностью. Видимо, чистая математика нас не спасает. Обратимся к биологии. Каковы есть возможные варианты замены нуклеотидов?

Модели эволюции нуклеотидов



Эволюцию нуклеотидных последовательностей можно представить как марковский процесс, поскольку каждое последующее состояние зависит от текущего, но не от предыдущего. Замены не оставляют никаких меток на цепи DNA, влияющих на следующие замены. Появляется 10 параметров: частота четырех нуклеотидов и 6 возможных переходов между ними (2 транзиции и 4 трансверсии).

Модели эволюции нуклеотидов

Автор(ы), название	Частоты нуклеотидов	Частоты переходов	Свободных параметров
Jukes-Cantor (JC69)	равные	равные	0
Kimura (K80=K2P)	равные	$Ts \neq Tv$	1
Felsenstein (F81)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	равные	3
Felsenstein (F84) & Hasegawa-Kishono-Yano (HKY85)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts \neq Tv$	4
Tamura-Nei (TN93)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts_R \neq Ts_Y \neq Tv$	5
Tavaré (GTR)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	6 переходов	8

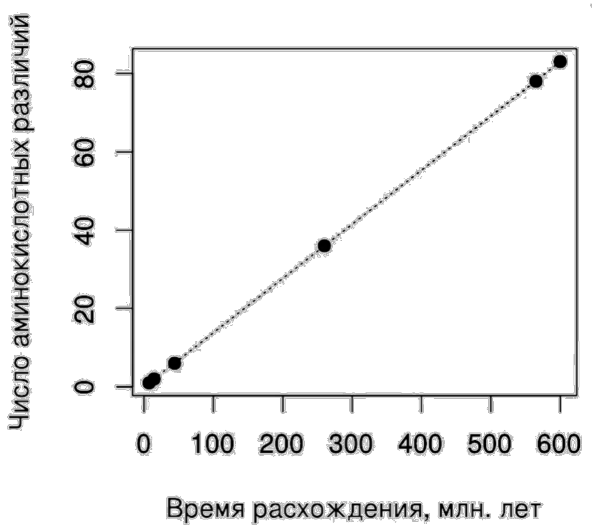
Пойдем далее. Какие вообще бывают методы построения деревьев? Глобально их можно разделить на 2 группы:

- Методы расстояний (дистанционные)
    - UPGMA (невзвешенного попарного среднего) – эвристический, предполагает молекулярные часы
    - NJ (присоединения ближайшего соседа) – эвристический, не предполагает молекулярные часы
    - LS (наименьших квадратов) – переборный, может предполагать молекулярные часы
    - ME (минимальной эволюции) – переборный, может предполагать молекулярные часы
    - Фитча-Марголиаша – переборный, может предполагать молекулярные часы
- Дискретные методы (символьно-ориентированные)
    - MP (максимальной парсимонии) – переборный, не предполагает молекулярные часы
    - ML (максимального правдоподобия) – переборный, может предполагать молекулярные часы
    - BI (байесовский подход) - переборный

Методы расстояний каждую пару последовательностей характеризуют конкретным числом – тем самым расстоянием между последовательностей, о котором мы говорили выше. Более точными методами являются дискретные методы, которые обсчитывают современные компьютеры. Они используют максимум возможной информации о последовательности, фактически информацию о каждой позиции.

Датировка

Вероятно, стоит поговорить о датировке, привязке топологии дерева к реальному историческому времени. Представьте, что у нас есть какая-то ветвь. Мы знаем, сколько замен на ней произошло. И эта длина ветви будет численно равно произведению настоящему времени дивергенции (времени появления от узла) до текущего времени на скорость накопления замен. Эту скорость накопления нам и нужно знать, чтобы определить дату появления группы. Тут либо мы обойдемся относительными значениями, а если нам нужны абсолютные значения, то тут пригодятся, например, палеонтологические данные. Как определить скорость накопления замен? В некоторых случаях она постоянна, и именно это привело к появлению *гипотезы молекулярных часов (molecular clock).* Теория говорит о том, что в целом скорость накопления мутаций постоянна, то есть время дивергенции будет линейно зависеть от числа накопленных различий.



И все бы хорошо, только вот это не всегда правда. Мы рассмотрели так называемые *строгие часы*, но придумали и модели, позволяющие определить разную скорость для разных ветвей.

Модели распределения скорости накопления замен

- Constant-rates (global clock) <= strict clock («строгие» часы)
- Variable-rates <= relaxed clock («ослабленные») + random (случайные)
    - Autocorrelated
        - Continuous: ACLN, ACG, AOUP, ACIP
        - Episodic: ACE, ACPP
    - Independent: UCLN, UCG, IGR, UCE
    - Local Clock: AHLC, RLC
    - Mixture: FMM, DPP

По Chen, M. H., Kuo, L., & Lewis, P. O. (Eds.). (2014). Bayesian Phylogenetics: methods, algorithms, and applications. Chapman and Hall/CRC.

Мы можем сформулировать свои априорные представления не только о скорости накопления замен, но и возрасте узлов. Такие группы моделей будут разделяться на специфичные для каждого узла и общие для каждой модели. Есть общие математические, а есть биологические модели. И биологические представления о видообразовании более важны в данном случае. Они подразделяются на микроэволюционные (популяционные) и работающие на уровне видов и выше.

Модели распределения возраста узлов

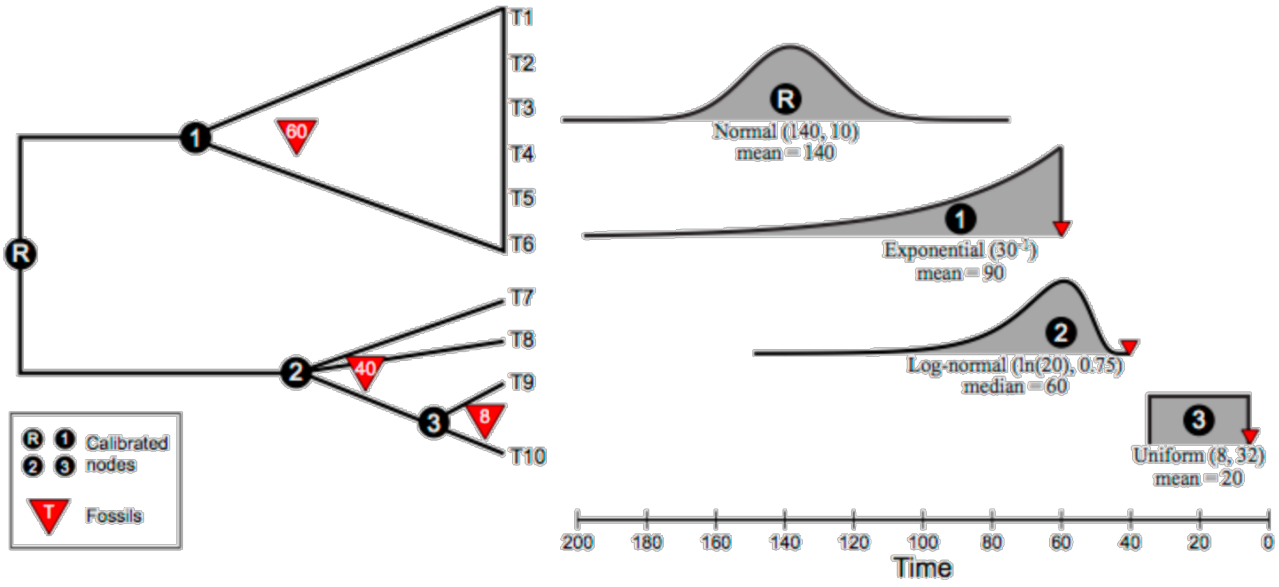
- Tree-Wide
  - Generic: uniform, Dirichlet
  - Biological
    - Population-Level: constant, exponential, logarithmic, skyline
    - Species-Level: Yule, birth-death, general birth-death
- Node-Specific

Chen, M. H., Kuo, L., & Lewis, P. O. (Eds.). (2014). Bayesian Phylogenetics: methods, algorithms, and applications. Chapman and Hall/CRC.

Если у нас есть сторонние данные, мы можем сказать и об абсолютном времени существования вида. Например, это могут быть:

- Биогеографические события
  Пример: был остров с двумя популяциями, потом их разделило некоторое географическое препятствие, например непроходимая гора или остров раскололся. Здесь мы можем геологически выяснить дату этого события. Однако так мы выясним лишь минимальный возраст узла, но не фактический, так как необязательно виды дивергировали вместе с появлением горы.
- Серии образцов
  В основном работает на вирусах
- Ископаемые останки
  Допустим у нас есть образец, который мы датировали и он является общим предком всей клады, то мы можем использовать эти данные для получения минимального возраста этой группы.

Также мы можем ввести некое априорное распределение:



https://sites.google.com/site/eeob563/computer-labs/lab-7

Построение деревьев методом UPGMA

Самым первым мы рассмотрим метод построения филогенетических деревьев, который довольно несложно применять на блокнотике. Метод называется UPGMA (Unweighted Pair Group Method with Arithmetic Averages) — математический метод невзвешенного попарного среднего. Корни метода уходят далеко в математику, а придумывался он вообще для классификации любого заданного множества объектов.

Как метод пришел в биологию

В экологии это один из самых популярных методов классификации единиц выборки (например, участков растительности) на основе их попарного сходства в соответствующих переменных (например, состав видов). В молекулярной биологии впервые был применен для исследований электрофореза белков. Но самую большую ценность метод имеет в биоинформатике, где используется для выравнивания, поскольку дает всего один его результат. В филогенетике строит дерево, нацеленное на группировку наиболее сходных последовательностей, независимо от их эволюционной скорости или филогенетической афинности. Он независим от скорости, потому что предполагает, что она постоянна (гипотеза молекулярных часов) и не является хорошо расцененным методом для определения отношений, если это предположение не было проверено и не оправдано для используемого набора данных.

Как метод работает

UPGMA является методом расстояний, то есть он оперирует попарным генетическим расстоянием между каждым секвенированным геномом. Пусть требуется провести классификацию заданного множества объектов методом невзвешенного попарного среднего.

Перед началом работы алгоритма рассчитывается матрица расстояний между объектами. На каждом шаге в матрице расстояний ищется минимальное значение, соответствующее расстоянию между двумя наиболее близкими кластерами. Найденные кластеры u и v объединяются, образуя новый кластер k. Строки и столбцы, соответствующие кластерам u и v, выбрасываются из матрицы расстояний, и добавляется новая строка и новый столбец, соответствующие кластеру k. В результате матрица сокращается на одну строку и один столбец. Эта процедура повторяется до тех пор, пока не будут объединены все кластеры. Пусть задана следующая матрица расстояний:

Пусть кластеры u, v и k содержат T(u), T(v) и T(k) объектов, соответственно. Кластер k образован путем объединения кластеров u и v, тогда T(k)=T(u) + T(v). Необходимо рассчитать удаленность кластера k от некоторого кластера w. Расстояние между этими кластерами определяется согласно формуле:

$$D\left(\left(u,v\right),w\right) = \frac{T_u D_{u,w} + T_v D_{v,w}}{T_u + T_v}$$

Здесь мы не будем строить дерево по геномным последовательностям, а попробуем метод на вкус с помощью сторонней задачи. Пусть у нас есть ряд абстрактных организмов (AB,C,D,E), у которых признаки «1,2,…,9» либо есть (1), либо нет (0). И таких признаков много. Для начала составляем таблицу эволюционных расстояний (произведем расчёт матрицы попарных дистанций). Здесь мы просто получаем дистанцию попарно между каждым видом суммой единиц и нолей, иными словами принимаем каждое различие в признаках за 1.

Признаки	A	B	C	D	E
1	1	1	0	1	1
2	0	1	0	1	0
3	1	1	0	1	1
4	0	0	0	0	0
5	0	1	1	0	0
6	0	0	1	1	0
7	1	0	1	1	1
8	0	0	1	0	0
9	0	1	1	0	1

Таксоны	A	B	C	D	E
A	0				
B	4	0			
C	6	6	0		
D	2	4	5	0	
E	1	3	5	3	0

Теперь алгоритм такой:

1. Находим два самых близких значения в матрице
2. Объединяем соответствующие узлы в группу
3. Перерасчитываем расстояния до остальных узлов как среднее расстояние до первого и второго членов группы
4. Строим новую матрицу (-1 строка и -1 столбец)
5. Если в матрице >1 значения, возвращаемся к первому пункту

Производим первое вычисление. На первом шаге, когда каждый объект представляет собой отдельный кластер (от A до E). Согласно критерию классификации, объединение происходит между кластерами, расстояние между которыми наименьшее. Т.е. на этом шаге объединяются кластеры A и E. Расстояние объединения — 1. Необходимо произвести перерасчет матрицы расстояний с учетом нового кластера:

Таксоны	A	B	C	D	E
A	0				
B	4	0			
C	6	6	0		
D	2	4	5	0	
E	1	3	5	3	0

Таксоны	(AE)	B	C	D
(AE)	0			
B	3,5	0		
C	5,5		0	
D	2,5			0

А) Относительный возраст кластера (AE)

$$\frac{d(A,E)}{2} = \frac{1}{2}$$

Б) Расстояния от кластера (AE) до остальных таксонов

$$d[(AE),B] = \frac{d(A,B)+d(E,B)}{N} = \frac{4+3}{2} = 3,5$$

Таксоны	A	B	C	D	E
A	0				
B	4	0			
C	6	6	0		
D	2	4	5	0	
E	1	3	5	3	0

Таксоны	(AE)	B	C	D
(AE)	0			
B	3,5	0		
C	5,5	6	0	
D	2,5	4	5	0

А) Относительный возраст кластера (AE)

$$\frac{d(A,E)}{2} = \frac{1}{2}$$

Б) Расстояния от кластера (AE) до остальных

$$d[(AE),B] = \frac{d[(AE),B)]+d[AE),B]}{N} = \frac{4+3}{2} = 3,5$$

Произведем второе вычисление. Приведем пример расчета расстояния между кластерами k=AE и w=D. Кластер k образован путем объединения кластеров u=A и v=E. Расстояния D(u,w) и D(v,w) берем из начальной матрицы расстояний. Подставив полученные значения в формулу, получим:

Таксоны	(AE)	B	C	D
(AE)	0			
B	3,5	0		
C	5,5	6	0	
D	2,5	4	5	0

Таксоны	((AE)D)	B	C
((AE)D)	0		
B	3,66	0	
C	5,33	6	0

А) Относительный возраст кластера ((AE),D)

$$\frac{d[(AE),D]}{2} = \frac{2,5}{2} = 1,25$$

Б) Расстояния от кластера ((AE),D) до остальных таксонов

$$d[((AE)D),B] = \frac{d[A,B]+d(E,B)+d(D,B)}{N} = \frac{3,5+3,5+4}{3} = 3,66$$

И наконец, третье вычисление. На последнем шаге объединяются два оставшихся кластера и ищем расстояние объединения.

Таксоны	((AE)D)	B	C
((AE)D)	0		
B	3,67	0	
C	5,33	6	0

Таксоны	(((AE)D)B)	C
(((AE)D)B)	0	
C	5,4975	0

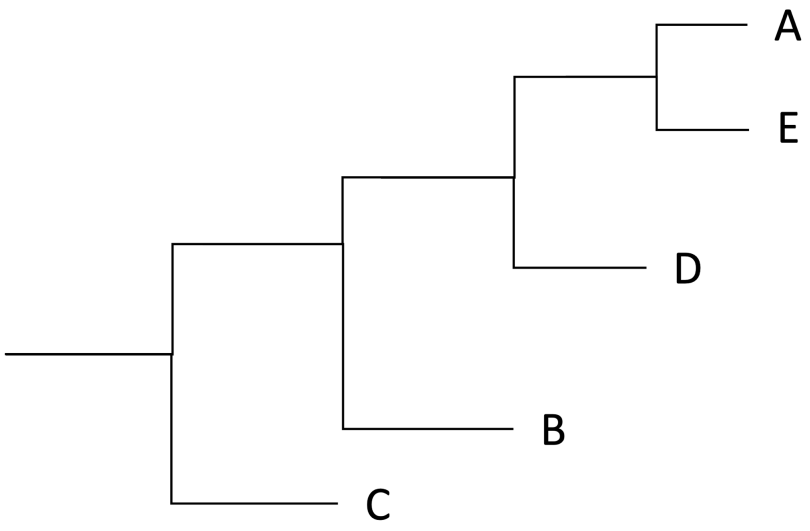
Относительный возраст кластера (((AE)D)B)

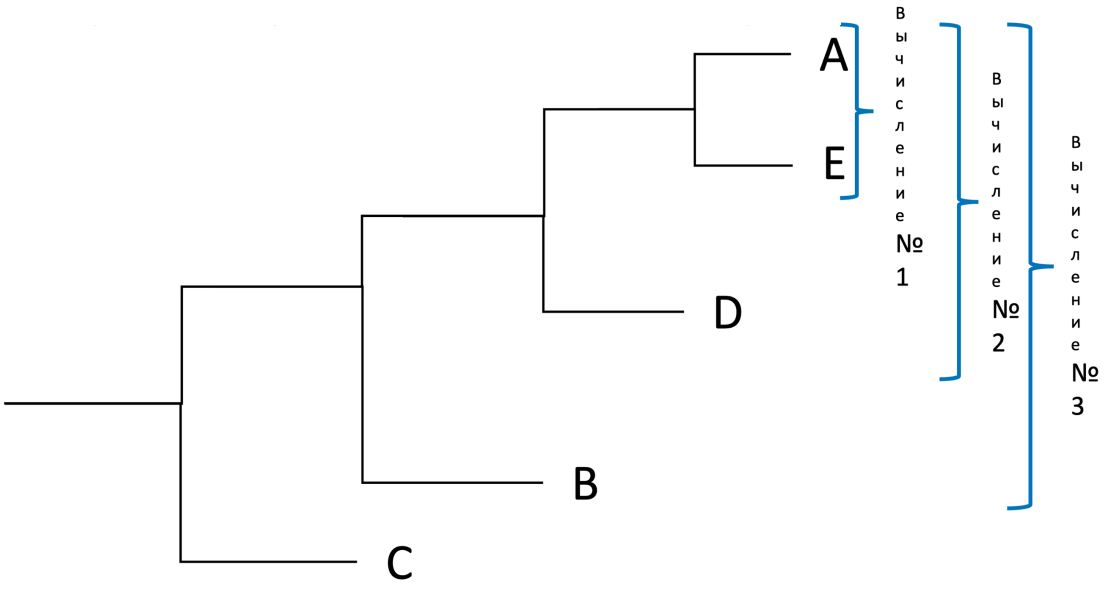
$$\frac{3,67}{2} = 1.835$$

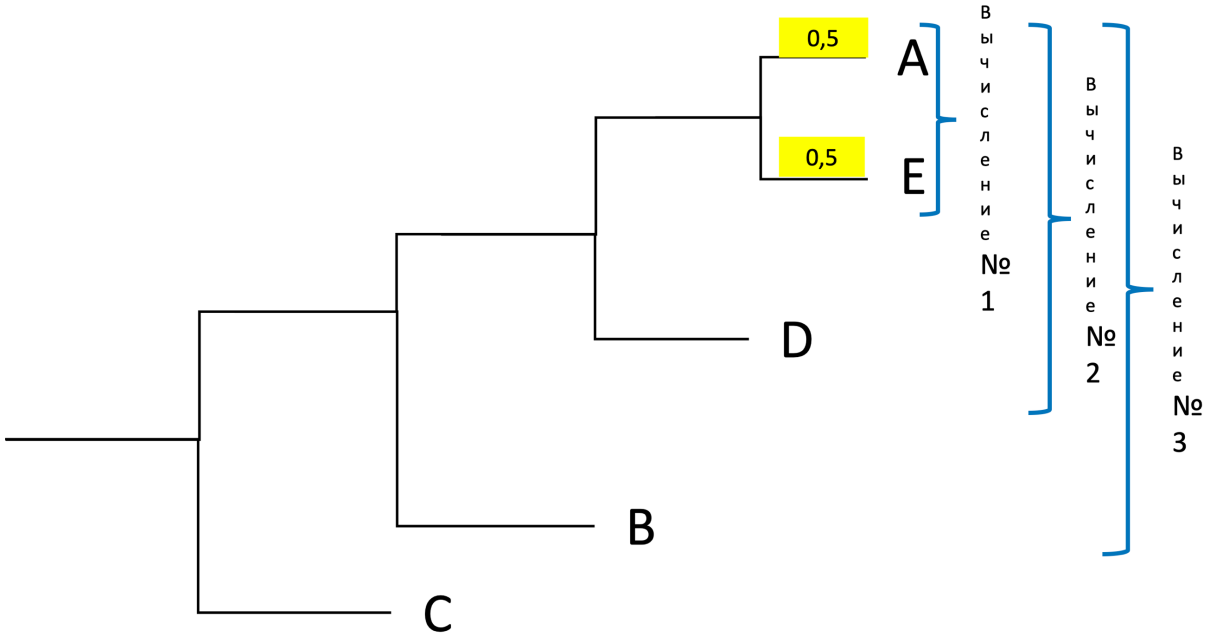
Относительный возраст кластера ((((AE)D)B)C)

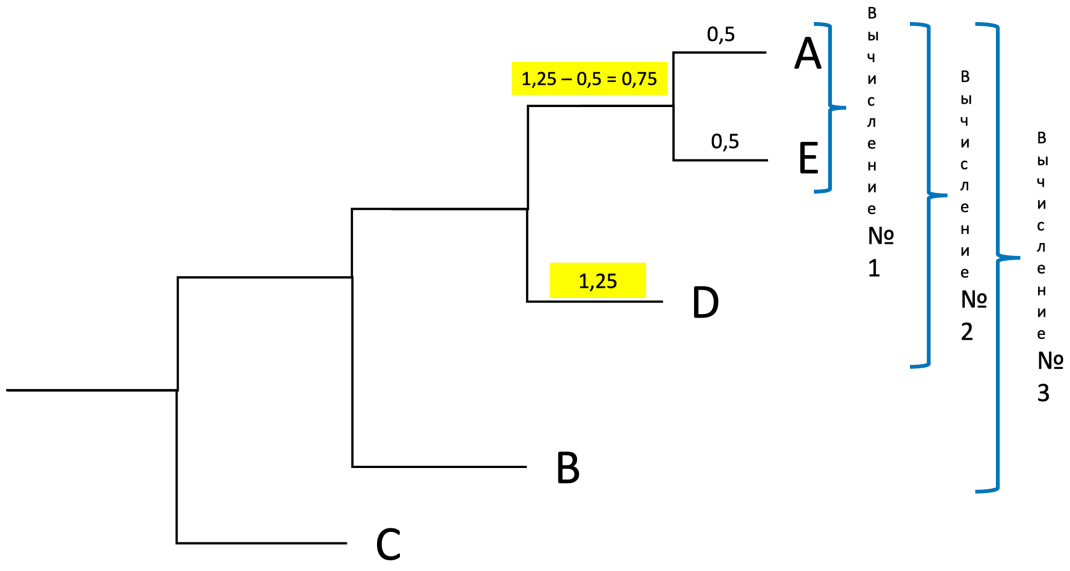
$$\frac{5,4975}{2} = 2,746$$

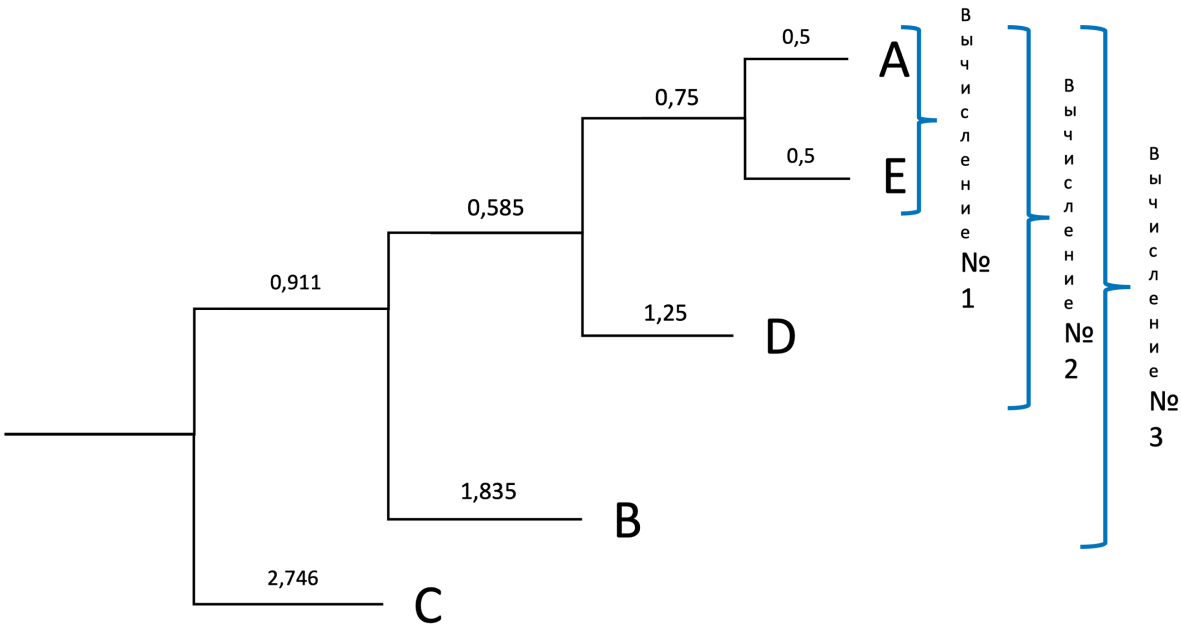
Результат работы алгоритма представляется в виде дендрограммы:

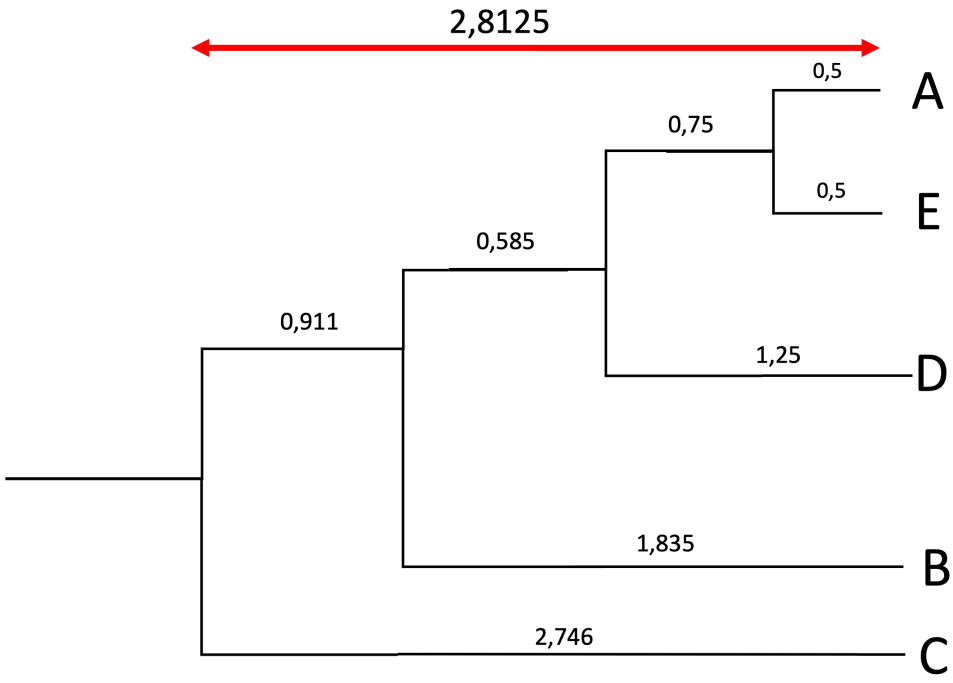



A
E
D
B
C
вычисление №1
вычисление №2
вычисление №3


0,5
0,5
A
E
D
B
C
вычисление №1
вычисление №2
вычисление №3


0,5
1,25 – 0,5 = 0,75
0,5
1,25
A
E
D
B
C
вычисление №1
вычисление №2
вычисление №3



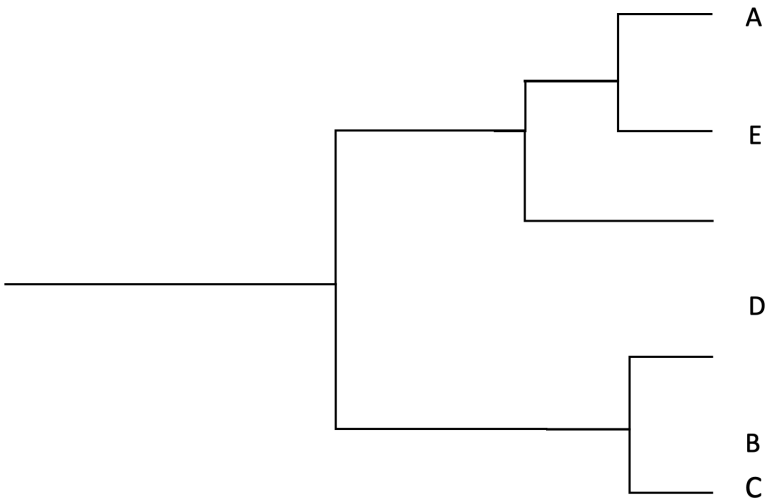


Что же делать, если в таблице сразу два наименьших значения? Сформируем любой из кластеров в первом вычислении, затем во втором объединим два оставшихся вида в другой кластер.

Таксоны	A	B	C	D	E
A	0				
B	4	0			
C	6	1	0		
D	2	4	5	0	
E	1	3	5	3	0

Таксоны	(AE)	B	C	D
(AE)	0			
B	3,5	0		
C	5,5	1	0	
D	2,5	4	5	0

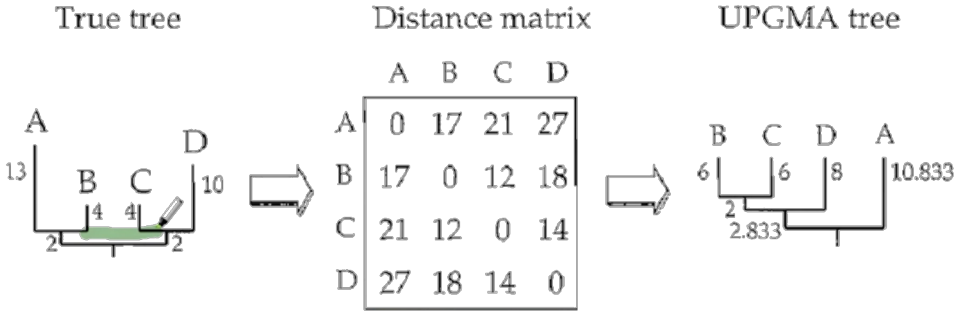
Таксоны	(AE)	(BC)	D
(AE)	0		
(BC)	4,5	0	
D	2,5	4,5	0



Метод далеко не хорош, поэтому в биоинформатике его используют, но почти никогда для построения филогенетических деревьев. Почему? Дело в том, что если у нас в дереве есть несколько очень коротких ветвей, на которых произошло немного изменений, то метод выдаст нам неправильное дерево. Короткие ветви (на рис. ниже выделены зеленым) он склеит между собой. Поэтому метод не подходит, если в разных ветвях разная скорость эволюции, то есть в одной ветви произошло много замен нуклеотидов, а в другой мало. Метод, однако, зашит в некоторые программы для выравнивания, а также для построения первого чернового дерева. Грубо говоря, построить дерево кластеризацией быстрее, чем перебором.

Мы рассмотрели пример, несвязанный с геномами, но суть там такая же, просто различия ищутся между последовательностями.

Иногда UPGMA ошибается. Метод не подходит, если скорость эволюции в разных ветвях разная.

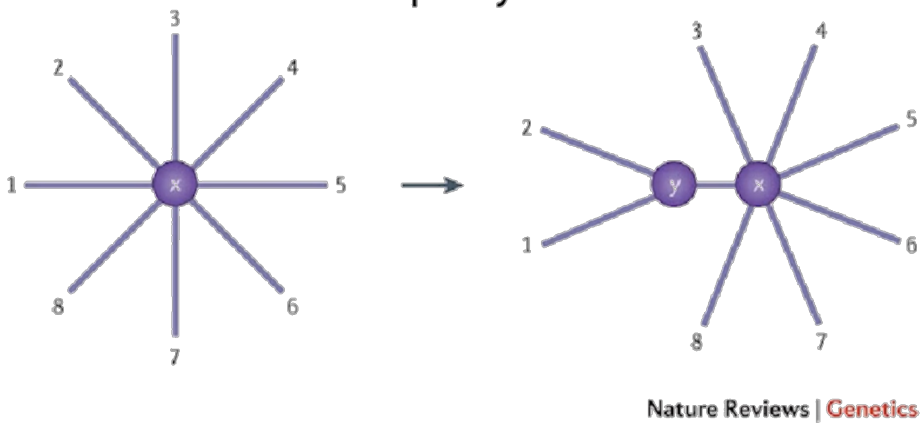


Felsenstein, Joseph. Inferring phylogenies. Vol. 2. Sunderland: Sinauer associates, 2004

Построение деревьев методом Neighbor joining

Здесь мы поговорим о методе присоединения ближайшего соседа (Neighbor joining — NJ). Мы выбираем на нашем полностью неразрешенном дереве (в виде звездочки) два случайных узла, объединяем их в группу, а все остальные узлы считаем второй группой. Определяем расстояние, далее проводим такой же анализ со всеми остальными возможными вариантами, то есть со всеми остальными группами из двух узлов. Такую версию, в которой расстояния будут минимальными мы фиксируем, строим новую матрицу расстояний (на одну строку и столбец меньше) и если в матрице все еще не одно значение, а больше, то повторяем. Таким образом, в отличие от метода UPGMA, мы получаем неукорененное дерево, потому что тут мы не строим иерархические кластеры.

1. Выбираем два случайных узла, объединяем их, всё остальное — второй узел.



Nature Reviews | Genetics

2. Определяем расстояния в матрице.
3. Повторяем со всеми остальными вариантами.
4. Минимальное расположение фиксируем.
5. Строим новую матрицу (-1 строка, -1 столбец).
6. Если в матрице >1 значения, см. п. 1.

На выходе мы имеем одно дерево и не факт, что оно правильное. Неукорененные деревья, как говорилось ранее, можно сделать укорененными.

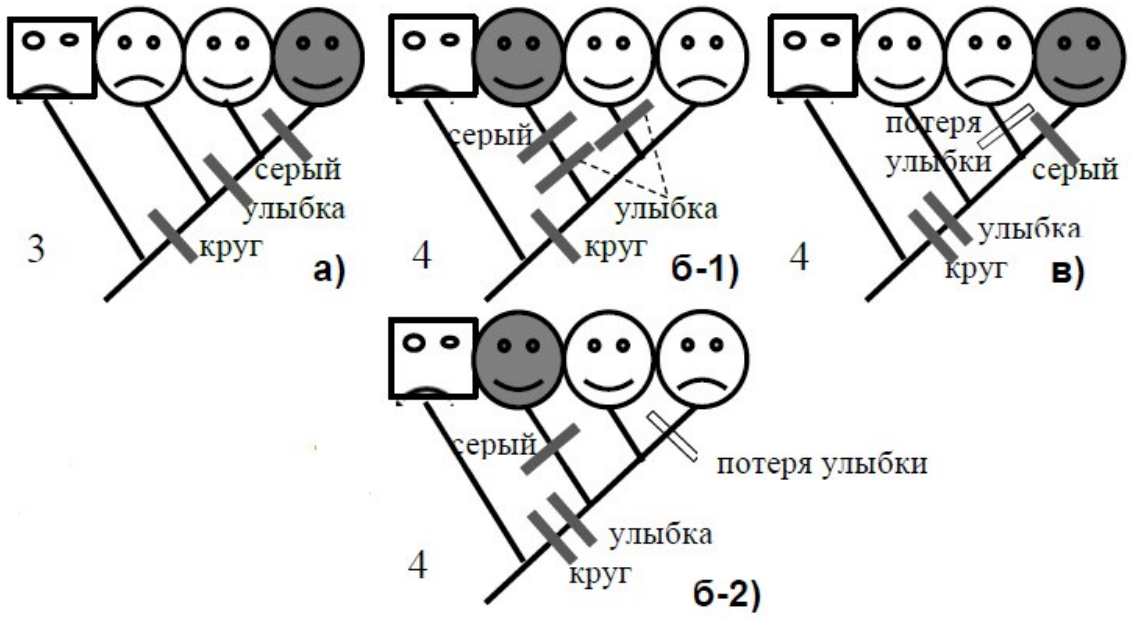
У метода NJ есть много модификаций, связанные с тем, как мы, собственно, определяем расстояние, корректируем ли мы на то, сколько последовательностей мы объединили, сколько вообще последовательностей. Наиболее используемой модификацией является BioNJ. Лучше всего работает с биологическими последовательностями. Из плюсов у NJ высокая скорость работы. Нужен для быстрого построения примерного дерева или если у вас очень много последовательностей и мало времени. Однако самый хороший результат у вас будет, если последовательности хорошо выровнены и между ними мало пропусков, иначе вероятность ошибки крайне высока.

Построение деревьев методом максимальной парсимонии

Maximum Parsimonia (MP) является в чистом виде кладистическим методом. Другое название - метод максимальной экономии. Теперь уже все более очевидно. Основное предположение кладистики заключается в том, что члены группы имеют общую эволюционную историю. Но важно помнить, что филогенетическое дерево - это практически всегда гипотеза. Филограмма, полученная дистанционными методами, не отражает эволюционного процесса, а только демонстрирует конечную степень дивергенции таксонов.

Метод парсимонии заключается в нахождении среди всех возможных вариантов филогенеза таких, которые достигаются за наименьшее число эволюционных событий, т.е. изменений состояний признаков. Еще более понятно - число мутаций Такие варианты являются наиболее экономными.

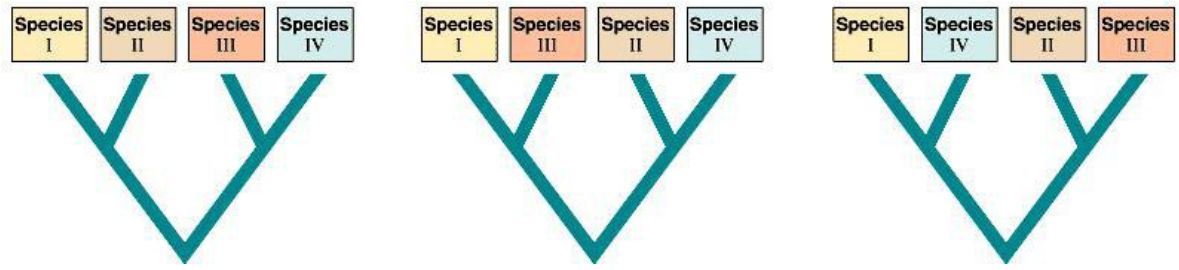
Рассмотрим филогенез смайлов, которых мы уже задействовали ранее. Они бывают 3 видов: грустные белые, веселые белые и серые веселые. Какие состояния признаков считать плезиоморфными, а какие апоморфными? Для этого введем в рассмотрение квадратиков, внешнюю группу для всех смайлов. Квадратики – белые и грустные ребята. Соответственно, эти состояния и являются плезиоморфиями. Построим все возможные варианты филогенеза.



Наименьшее число эволюционных событий наблюдается на дереве наблюдается под буквой а. Значит это и есть наиболее вероятный вариант филогенеза согласно методу парсимонии. Стоит также отметить, что различные варианты ложных синапоморфий часто дают одинаковое число эволюционных событий (рис. б-1 с конвергенцией и б-2 с реверсией). Это различные эволюционные сценарии, по сути не различимые в кладистике.

Что лежит в основе метода?
– Общие, видоизмененные признаки как основа для
классификации
– Принцип экономичности (Бритва Оккама)
Критично: Проблема выявления гомологий. Эволюция не имеет цели и не всегда экономична.
Проблема сетчатой эволюции не учитывается, потому-что кладистика может анализировать только дивергирующие таксоны. Т.е. теоретически нельзя использовать кладистический метод для внутривидового анализа и при наличии гибридогенных видов.

Процесс преобразования данных в филогенетические деревья сопряжен с серьезными проблемами: если мы хотим выяснить родственные связи между четырьмя видами, мы должны будем сделать выбор между несколькими деревьями. Это мы увидели на смайлах.



При всех недостатках, принцип парсимонии помогает систематикам реконструировать филогению
При включении в анализ все больше и больше деревьев, количество возможных деревьев будет чудовищно увеличиваться (экспонента). Например для дерева в 50 видов найдется порядка 3000 деревьев.
Даже используя компьютер, анализ подобного обьема данных для поиска наилучшего дерева, будет длится очень долго.
Систематики используют принцип парсимонии (бережливости), чтобы выбрать среди множества возможных деревьев одно дерево, которое наилучшим образом отражает анализируемые данные.

Для желающих прилагаю название программы, где можно с этим поиграться (попробуйте найти сами файлы каких-нибудь цитохром-оксидаз)

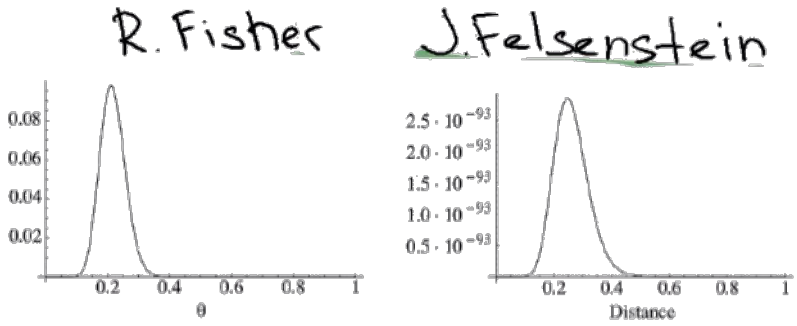
PAUP: Phylogenetic Analysis Using Parsimony (and
Other Methods) 4.0 Beta

Построение деревьев методом максимальной правдоподобия

Maximum Likelihood (ML) – дискретный метод

В чем суть этого метода? Нам нужно объяснение эволюционного пути, которое делает наблюдаемое решение наиболее правдоподобным. Но правдоподобие в данном случае будет не нашим откликом в разуме об адекватности того или иного варианта, а конкретной математической величиной. Более формально, если имеются некоторые данные D и гипотеза H, вероятность получают при LD=Pr (D|H)
Какова вероятность D при H?
В контексте молекулярной филогенетики:
D - это сравниваемые сиквенсы
H - это филогенетическое дерево
Мы хотим получить наиболее вероятное дерево на основе полученных данных (матрица сиквенсов). Наиболее вероятное дерево, которое получается на основе полученных данных является максимально правдоподобным вариантом филогении. Важно различать правдоподобие и вероятность: все вероятности в сумме дают единицу, правдоподобие нет. В случае дерева и модели: исследуется вероятность получения дерева при включении всех возможных наборов данных. Сумма этих вероятностей будет равна 1. Но мы заинтересованы только в одном наборе данных, который мы получили.
Примечание: правдоподобие (likelihood) это не вероятность, что полученное дерево — это правильное дерево, а просто максимальная вероятность дерева на базе полученных данных.

Немного о правдоподобии с точки зрения математики.
В случае расчета вероятности, имеем функцию, зависящую от события, а в случае с правдоподобием — от параметра при фиксированном событии.
С помощью методов ML 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



Функция правдоподобия для монетки и выравнивания (JC69).

$$L(\tau, \Sigma) = Pr(Data|\tau, \Sigma) = Pr(alignment|tree, model)$$

Schmidt HA, von Haeseler A. Phylogenetic inference using maximum likelihood methods. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. 2009.

$$L(\tau, \Sigma) = Pr(Data|\tau, \Sigma) = Pr(alignment|tree, model)$$

Достоинства:

- наиболее биологически оправданный.

Недостатки:

1. время- и ресурсоёмкость;
2. сильно зависит от выбранной модели.

Приведем пример. Пусть монетку подбросили 100 раз - из них 80 выпало орлом. Вероятность, соответственно: 0,8. Рассмотрим задачу наоборот - зная, что монетка выпала орлом 80 из 100 раз, какова вероятность того, что эта монетка честная?
В деревьях мы работаем с условной вероятностью, при такой-то топологии дерева и такой-то модели эволюции. Мы не считаем вероятность правильности нашей гипотезы. Мы считаем, насколько вероятно получить такое-то выравнивание при таком-то дереве. И ищем мы такое дерево из возможных, которое максимизирует вероятность того, что эволюция шла именно так. Максимальное правдоподобие, как и любой другой метод зависит от выбранной модели эволюции. В конечном счете любой метод зависит от допущений, сделанных при формулировке гипотезы, но здесь это проявляется сильнее всего.

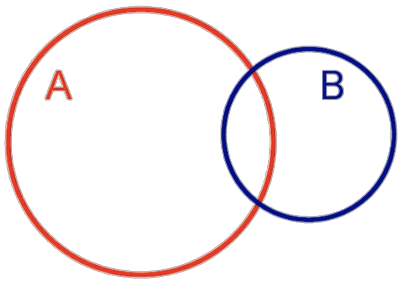
Итак, это один из самых сложных, но хороших методов - с ним могут соперничать только байесовские.

Байесовские методы построения деревьев

Перед тем, как пойти дальше, немного погрузимся в математику. Поговорим о байесовской статистике.

Условная вероятность

Условная вероятность события A, при условии, что произойдет событие B – это отношение вероятности того, что произойдут оба события к вероятности события B. Из этого следует, что вероятность того, что произойдут оба события равна произведению вероятности A при условии B и вероятности собственно B. Путем нехитрых преобразований можем получить обратное: B при условии A. Но так как левые части выражений очевидно равны, мы можем приравнять и правые. Приравняв правые, поделим обе части на вероятность одного из событий и получим знаменитую теорему Байеса. Условная вероятность некоторого события равняется произведению его вероятности на обратную условную вероятность, отнесенную к вероятности второго события.



$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad <=> \quad P(A \cap B) = P(A|B) \cdot P(B)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad <=> \quad P(A \cap B) = P(B|A) \cdot P(A)$$

$$P(A|B) \cdot P(B) = P(A) \cdot P(B|A)$$

Условную вероятность A от B принято называть апостериорной (posterior), вероятность первого события априорной (prior). Кстати, в этом случае вероятность события B будет равняться сумме произведений вероятности A на B при условии A и вероятности «не A» на B при условии «не A». Событий может быть больше, чем два.

Теорема Байеса (Bayes' theorem)

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

$P(A|B)$ — апостериорная (posterior) вероятность;

$P(A)$ — априорная (prior) вероятность;

$P(B|A)$ — вероятность B при условии A;

$P(B)$ — вероятность события B.

$$P(B) = P(A) \cdot P(B|A) + P(\overline{A}) \cdot P(B|\overline{A})$$

Какое собственно отношение все происходящее имеет к филогенетике. Дело в том, что мы можем для нашего дерева, для его топологии, длины ветвей, разных параметров эволюционной модели написать такое же выражение: мы можем посмотреть какова апостериорная вероятность, получить вот такое дерево при условиях X, скажем, нашего выравнивания. Вероятность будет равна произведению некоторой априорной вероятности гипотезы на правдоподобие данных, отнесенное к полной вероятности наших данных. В данном случае это нормализующая константа, нужная чтобы вероятность в итоге оставалась в промежутке от 0 до 1.

$$P(T, \beta, k|X) = \frac{P(T, \beta, k) \cdot P(X|T, \beta, k)}{P(X)}$$

T — топология;

β — длина ветвей;

k — параметры эволюционной модели.

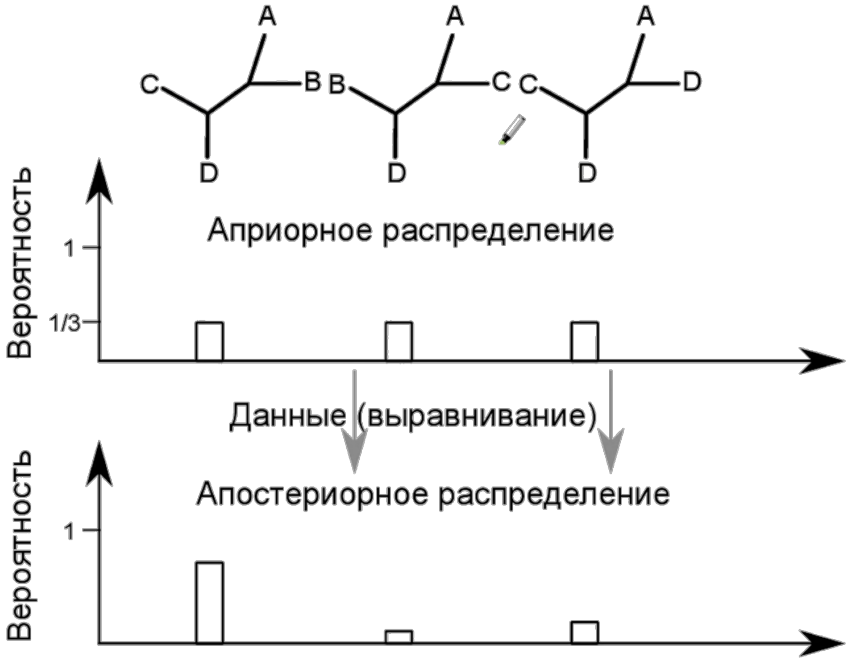
$P(T, \beta, k|X)$ — апостериорная вероятность гипотезы;

$P(T, \beta, k)$ — априорная вероятность гипотезы;

$P(X|T, \beta, k)$ — правдоподобие данных;

$P(X)$ — нормализующая константа.

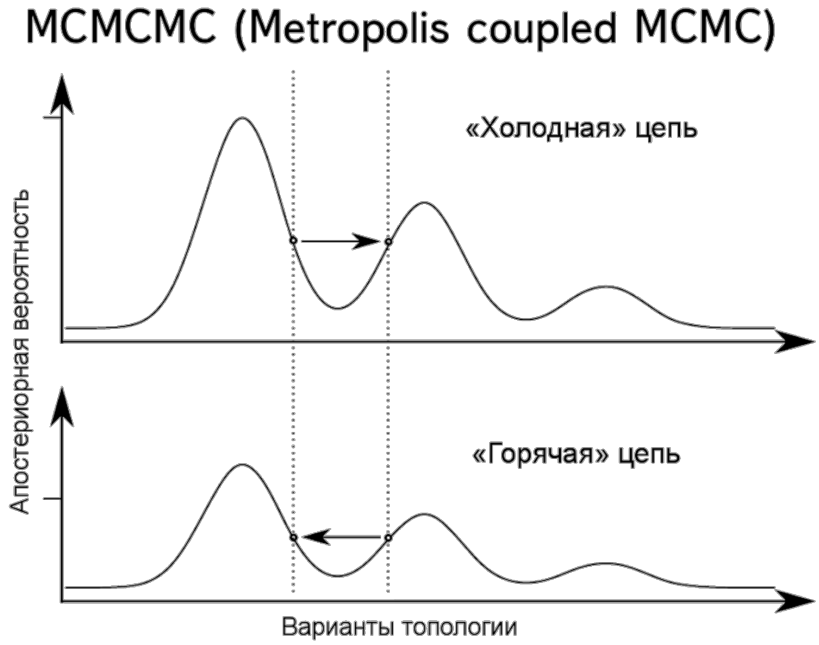
Для примера представим себе неукорененное дерево из 4 таксонов, у него есть всего 3 возможных топологии. Априорное распределение их равномерное (1/3), то есть вероятность получить кажждую из топологий составляет 1/3, потом мы можем пересчитать эту вероятность при условии наших данных, на основании выравнивания и получить апостериорное распределение, которое нам скажет, что допустим вероятность первого распределения больше, чем у остальных топологий.



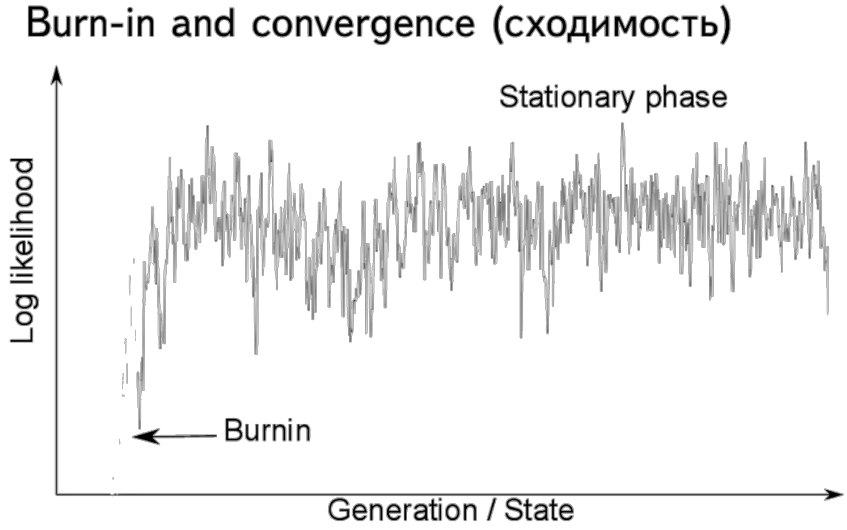
Все предыдущее звучит очень неплохо. Но у нас есть проблема. При работе с реальными данными у нас будет очень много деревьев, разные возможные параметры моделей эволюции и так далее, у всего этого есть распределения. Более того, нам не поможет случайная выборка. Поэтому для работы с апостериорными вероятностями используют алгоритм Монте-Карло по схеме марковской цепи (MCMC). Суть метода легко объяснить на визуальном примере: наш человечек высадился на поверхности с вероятностью (для удобства представим ее одномерной). И вот человечек начинает шагать, делать шаги заранее зафиксированной длины. Если он делает шаг вверх (вероятность текущая будет больше, чем на предыдущем шаге), то он этот шаг всегда сделает. Если человечек пытается сделать шаг вниз (апостериорная вероятность), то есть апостериорная вероятность станет ниже, то он либо совершит, либо не совершит этот шаг с вероятностью, равной соотношению апостериорных вероятностей, которые были до того и которые получились после. То бишь вверх мы идем всегда, а вниз с некоторой вероятностью. И рано или поздно человечек шагая доберется до вершины (максимума). Ему просто не будет смысла уходить оттуда – он будет шагать и возвращаться на вершину.

По Lemey, P, Salemi, M. and Vandamme, A. M. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press, 2009.

Это хороший случай, но не всегда может повезти. Робот может забраться на вершину маленького холма – локальный максимум, но не глобальный. Чтобы выбраться из локального максимума и дойти до нужного нам глобального используют метод MCMCMC (Metropolis coupled MCMC). Его суть состоит в следующем: у нас есть привычные цепи, которые ходят по нашему ландшафту (холодных). Мы можем запустить еще несколько «горячих» цепей. Они будут ходить по несколько уплощенному ландшафту, которые мы создадим, возведя наши апостериорные вероятности в дробную степень. Соответственно горячие цепи будут совершать более смелые шаги и перепрыгивать с одного максимума на другой. И время от времени горячие и холодные цепи будут меняться местами. И соответственно горячие цепи будут прыгать по максимумам, а холодные цепи будут обследовать каждый холм и в конце концов дойдут до глобального максимума.



Следить за происходящим можно с помощью следующих графиков. Их обычно делают в логарифмическом масштабе, так как в начале правдоподобие сильно возрастает (burn-in участок), а деревья, полученные на этом этапе, даже не записывают, потому что они нам не нужны, в значительной степени они случайны. После мы видим, как наши «человечки» совершают шаги, находясь плюс минус у одного места, наверное, они нашли уже максимум вероятности. И деревья на этом этапе программа и будет записывать. И когда шаги становятся совсем маленькими, и в течение долгого времени разброс не увеличивается и не уменьшается, анализ можно прекращать – это стационарная фаза.

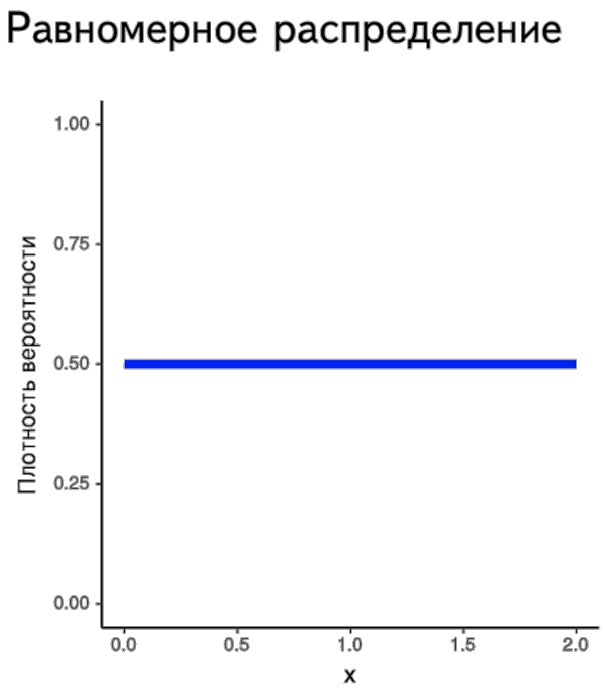


www.molecularevolution.org/resources/activities/beast_activity/viruses
см. также Lemey, P, Salemi, M. and Vandamme, A. M. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press, 2009.

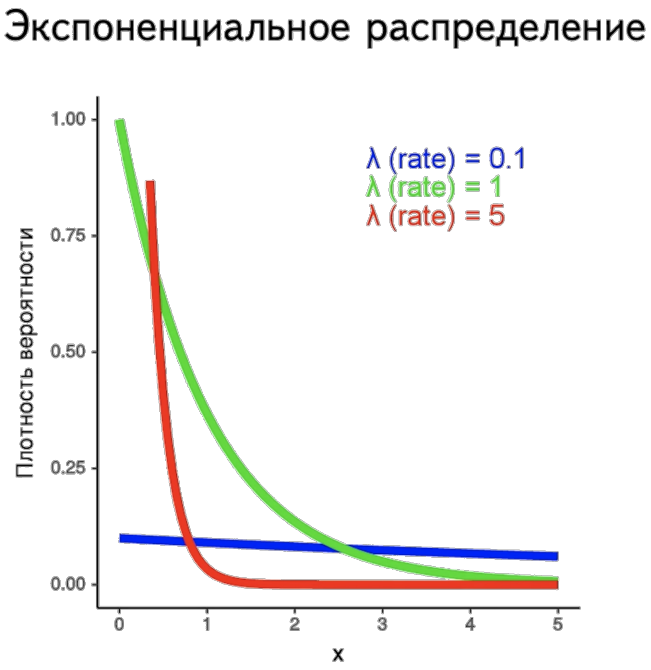
Ура! Мы подошли к байесовским методам восстановления филогении.

Обсудим более подробно априорную вероятность. В филогенетике ее мы не знаем, но мы можем задать ее распределение. Наиболее часто для этой цели используют равномерное, экспоненциальное, гамма- и бета-распределения и распределение Дирихле.

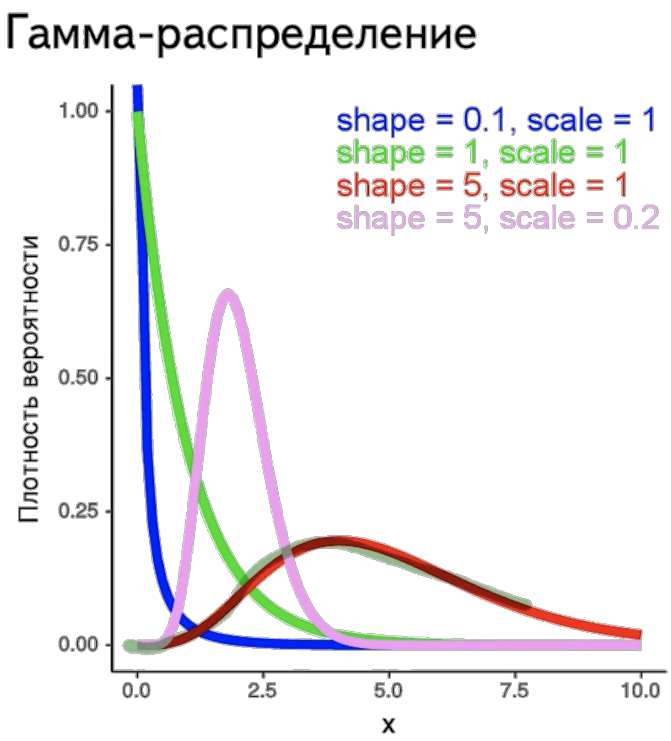
Равномерное распределение используют для топологии дерева, так как исходно у нас нет оснований предпочесть какую-то одну из топологий, поэтому лучше всего равномерно распределить вероятность между ними.



Экспоненциальное распределение хорошо показывает распределение длин ветвей. У него один параметр, обычно называемый лямбдой, и он регулирует форму этого распределения.

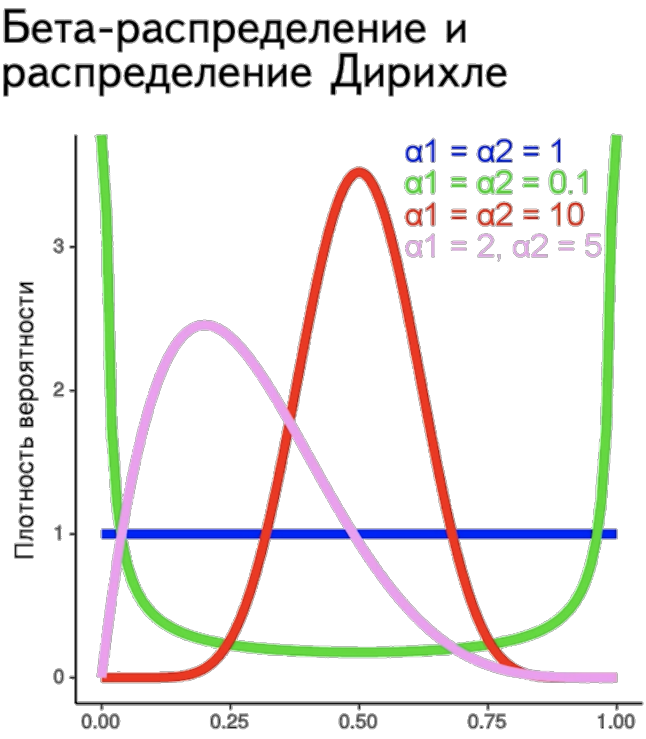


Гамма-распределение имеет два параметра, и, в зависимости от их взаимоотношения, оно сильно различается по форме. Годится для определения длины ветвей, потому что является суммой экспоненциальных распределений. Можно также использовать для определения скорости накопления замен, распределения скоростей накопления замен между разными сайтами.



Бета-распределение имеет два параметра и в зависимости от их взаимоотношения также имеет разную форму. Хорошо для всего, что является долями, например для частот замен, в данном случае два параметра задают частоты транзиции и трансверсии.

Распределение Дирихле похоже на бета-, но у него больше параметров.



Все вышеперечисленное можно применить в программах MrBayes и Beast.

В итоге у нас есть два явных мастодонта в области кладистики – метод максимального правдоподобия и байесовские методы

Немного о других способах построения деревьев

Метод наименьших квадратов, метод Фитча

Метод измерения расстояний, основанный на минимизации разностей между расстояниями на филогенетическом дереве и в соответствующей матрице расстояний. По точности и эффективности примерно равен методу объединения ближайших соседей. Считается неподходящим для исчерпывающего филогенетического анализа, но используется для построения предварительных филогенетических деревьев для метода наибольшего правдоподобия

Метод минимальной эволюции (Minimal Evolution)

По сути, делает то же что и NJ, но не с помощью кластеризации, а с помощью сравнения топологий деревьев. Основная идея – сделать расстояние между ветвями минимальное. Лучшее дерево – дерево с меньшей суммой ветвей. На вход подается матрица расстояний. Хорошо работает с большим количеством длинных последовательностей. В обратной ситуации NJ справляется лучше.

Заключение

В данном пособии мы попытались максимально широко рассмотреть современный подход к систематике живых организмов, обозначении и формализации эволюционных процессов. Мы рассмотрели понятие деревьев, затронули теорию графов, поговорили об их топологии. Мы узнали о процессе секвенирования, выравнивания последовательностей и о различных методах построения филогенетических деревьев. Стоит, однако подвести пессимистичный итог. Во-первых, само пособие, вероятно, не является полным, не все вопросы были освещены, хотя это в принципе невозможно в формате пособия. Возможно, здесь также остались некоторые неточности в формулировках и не поясненные термины. Но это, с другой стороны, перспектива для улучшения. Во-вторых, самый внимательный читатель мог заметить, что филогенетика и кладистика, хоть и признаны на сегодняшний день наиболее правдоподобными в реконструкции эволюционных процессов и используются для классификации, не являются истиной в последней инстанции. Биоинформатики, возможно, излишне математизируют процесс и забывают о реальном мире, вводя многие допущения для моделей. А ведь именно на допущениях рушатся модели и теории. Морфология и филогенетика должны не жить изолированно друг от друга, а совместно строить картину мира. Похоже на то, что никто никогда не узнает всей эволюционной истории органического мира с точностью до нуклеотида, а значит, мы располагаем лишь тенью процесса и не более. В-третьих, остается очень сложным вопрос соотношения фенотипа и генотипа. Одни лишь последовательности первичной структуры биополимеров не очень много говорят об их функции. При всем этом, скорее всего, изучать белки более перспективно, но не всегда возможно. Не смотря на подобное заключение, автор хочет верить в то, что данное пособие было полезным.

Использованная литература

1. Futuyma, Douglas J. Evolution / 2005
2. Sneath, P. H. A.; Sokal, R. R. Numerical taxonomy. The principles and practice of numerical classification. / 1973 pp. xv + 573 pp.
3. Кунин Е.В. Логика случая. О природе и происхождении биологической эволюции/Пер. с англ. – М.: ЗАО Издательство Центр-полиграф, 2014 – 527 с.
4. Павлинов И.Я. Кладистический анализ (методологические проблемы). – М.: Изд.- во МГУ, 1990. – 160 с.: ил.- ISBN 5-211-00918-5
5. Бизяев Н.С. Пособие по филогенетике. Киров, 2016
6. Павлинов. И.Я. Введение в современную филогенетику (кладогенетический аспект). М.: изд-во КМК, 2005.