Методы филогенетического анализа: процедура выравнивания, филогенетические деревья и методы их построения

Задача молекулярной филогенетики состоит в преобразовании информации содержащейся в семантидах (ДНК) в эволюционное дерево....

Для **преобразовании информации** содержащейся в ДНК, необходимо подготовить полученные сиквенсы (или данные RAPD, AFLP и др.) для анализа - **Матрица** или **Alingnment**

Alignment
Процедура выравнивания

Alignment - это гипотеза о Гомологии между нуклеотидами

- Гомология: Сходство это результат Наследства от одного общего предка

- установление и сравнение гомологичных признаков это центральный принцип филогенетического анализа

Процесс выравнивания (**alignment**) подразумевает выстраивание последовательностей наиболее близких к исследуемому объекту видов друг под другом с таким расчетом, чтобы они совпадали как можно более полно, тогда одни консервативные участки будут располагаться под другими.

ДНК Алфавит

Все ДНК молекулы состоят из нуклеотидов, которые содержат сахар, фосфатную группу и одну из четырех нуклеидных кислот: Adenin, Cytosin, Guanin и Thymin. Из их начальных букв и состоит алфавит ДНК (A, C, G; T)

Однобуквенные сокращения для ДНК алфавита

A	Adenine
C	Cytosine
T	Thymine
G	Guanine
W	Weack bonds (A, T)
S	Strong bonds (C, G)
R	Purines (A, G)
Y	Pyramidines (C, T)
K	Keto (T, G)
M	Amino (A,C)

B	C,G or T
D	A,G or T
H	A,C or T
V	A,C or G
N	A,C,T or G

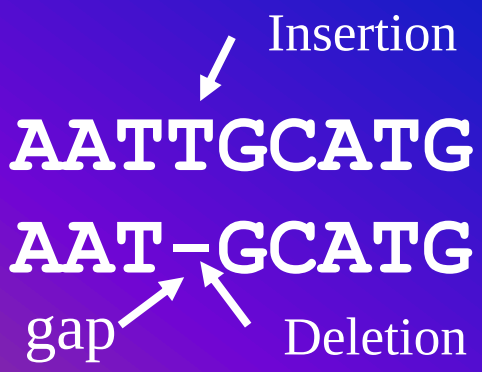
Мутации

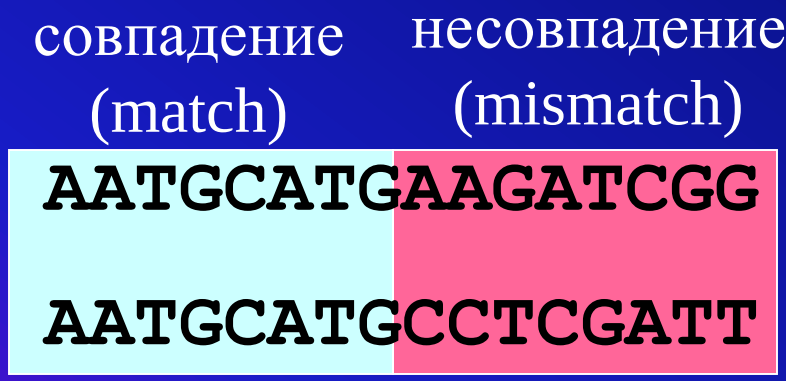
1) **Характеризуемые по длине сиквенса**
   - **a) Точечная мутация** (только один нуклеотид)
   - **b) длинная мутация** (затрагивает 2 или многих соседних нуклеотидв

2) **Характеризуемые по способу замены нуклеотидов**

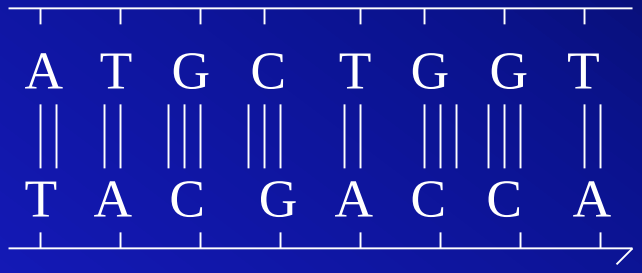
- **a) Substitution** (замена одного нуклеотида на другой)
- **b) Inversion** (переворот 2 или нескольких нуклеотидов одной части ДНК на 180°)
- **c) Insertion** (вставка одного или нескольких нуклеотидов)
- **d) Deletion** (потеря одного или нескольких нуклеотидов)

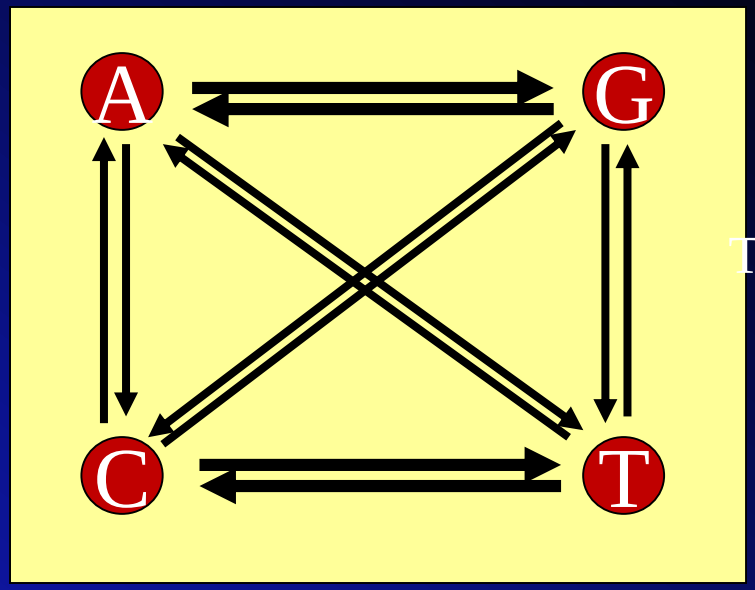
} **InDels**





Предковая ДНК

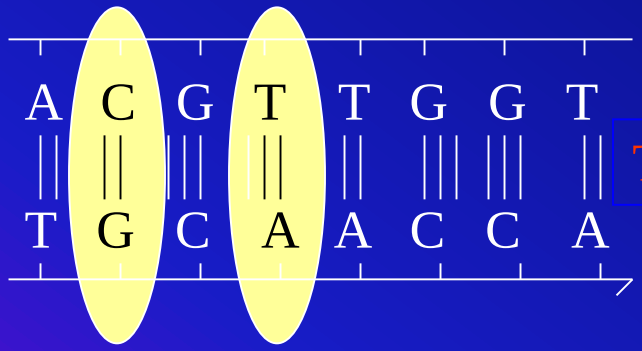




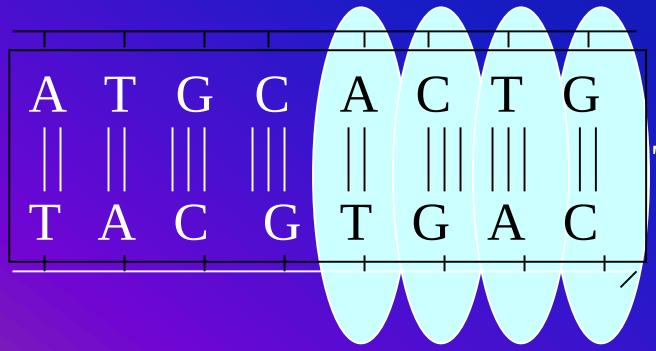
Transitions

Transversions

Transitions



Transitions



Transversions

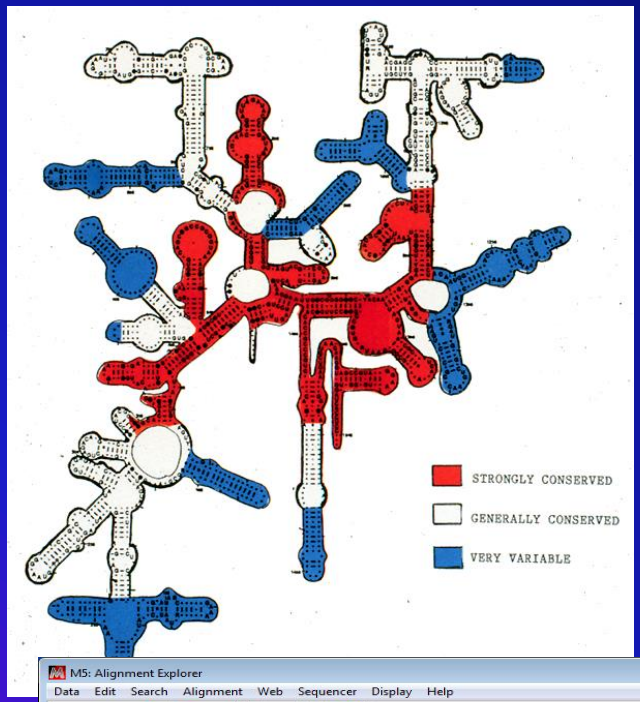
Выравнивание может быть простым или сложным

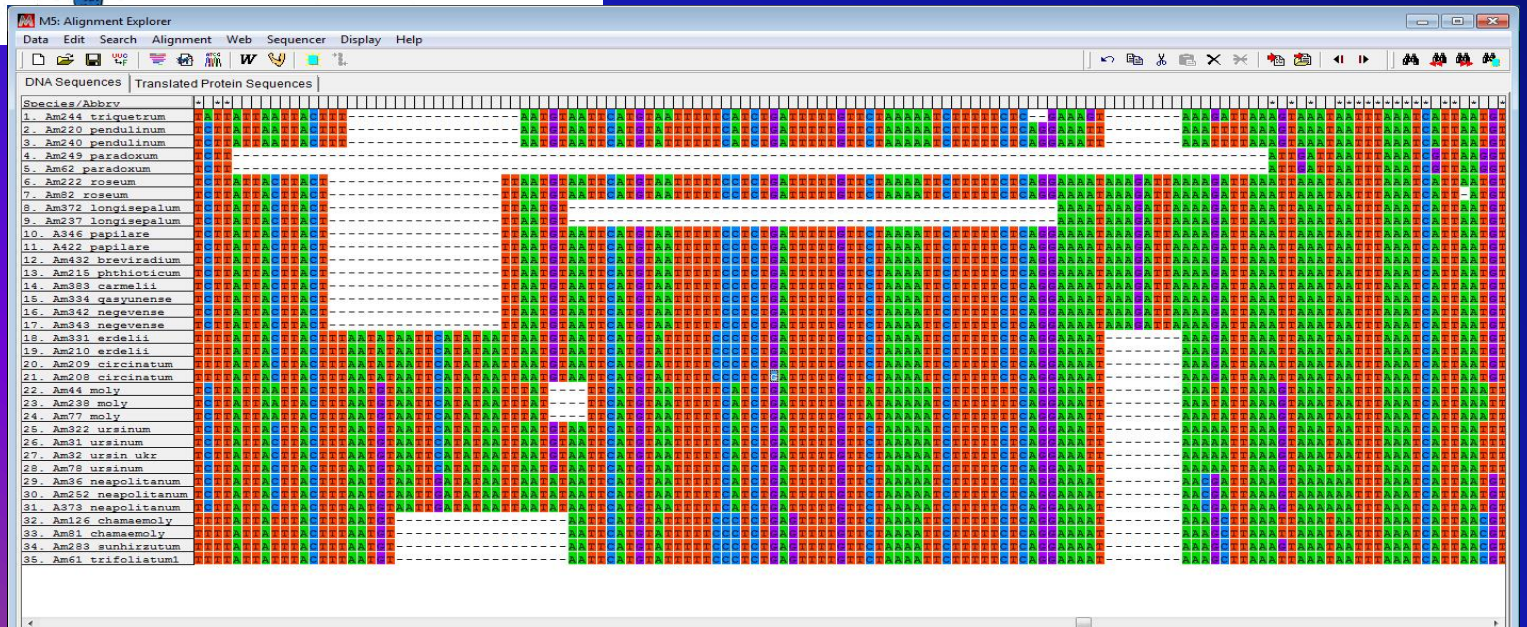
```
GCGGCCCA TCAGGTAGTT GGTGG
GCGGCCCA TCAGGTAGTT GGTGG
GCGTTCCA TCAGCTGGTT GGTGG
GCGTCCCA TCAGCTAGTT GGTGG
GCGGCGCA TTAGCTAGTT GGTGA
******** ********** *****

TTGACATG CCGGGG---A AACCG
TTGACATG CCGGTG--GT AAGCC
TTGACATG -CTAGG---A ACGCG
TTGACATG -CTAGGGAAC ACGCG
TTGACATC -CTCTG---A ACGCG
******** ?????????? *****
```

Простое

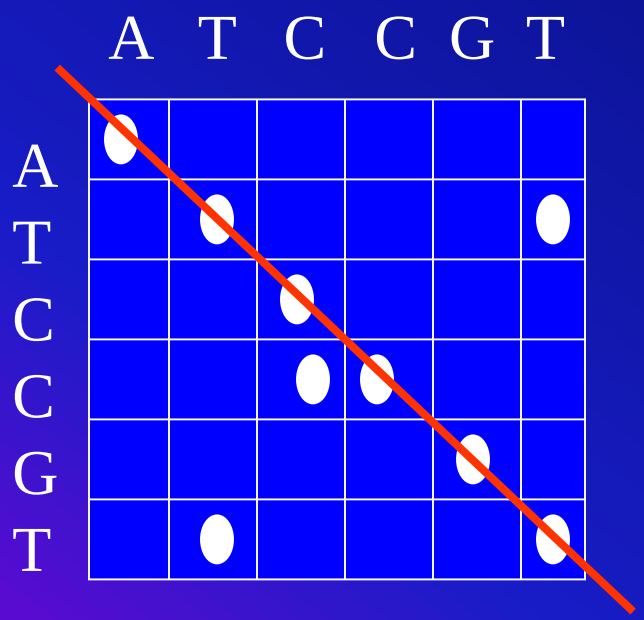
Сложное из-за вставок или делеций (indels)

Структупная формула части молекулы хлоропластной ДНК

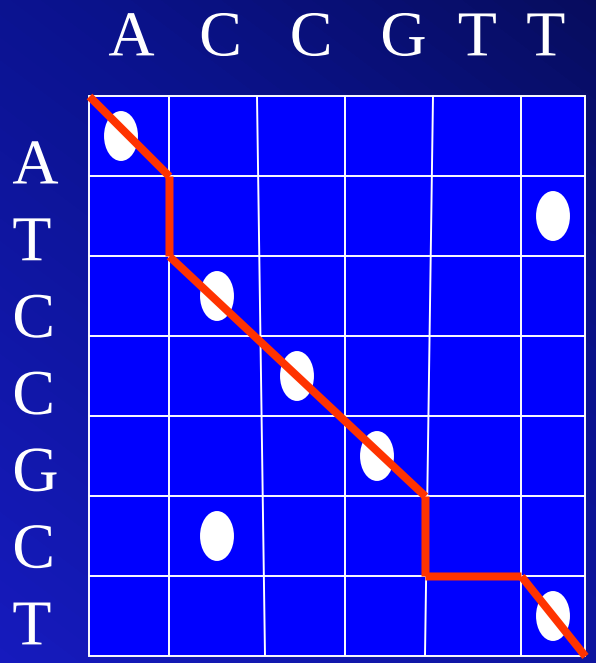




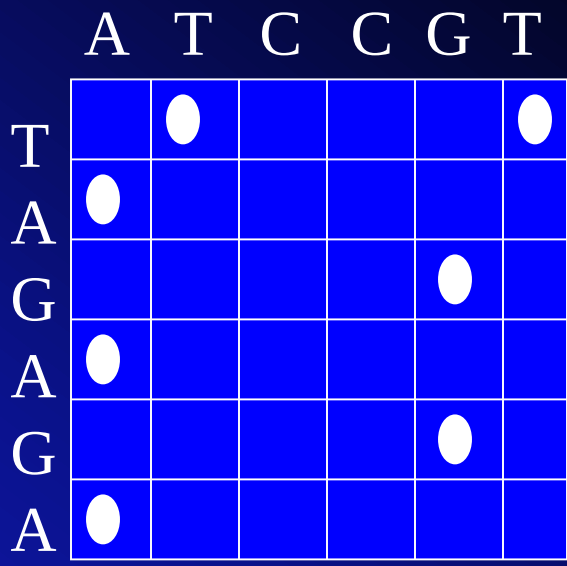
Метод точечной матрицы



ATCCGT
ATCCGT



A-CC-GTT
ATCC-GCT



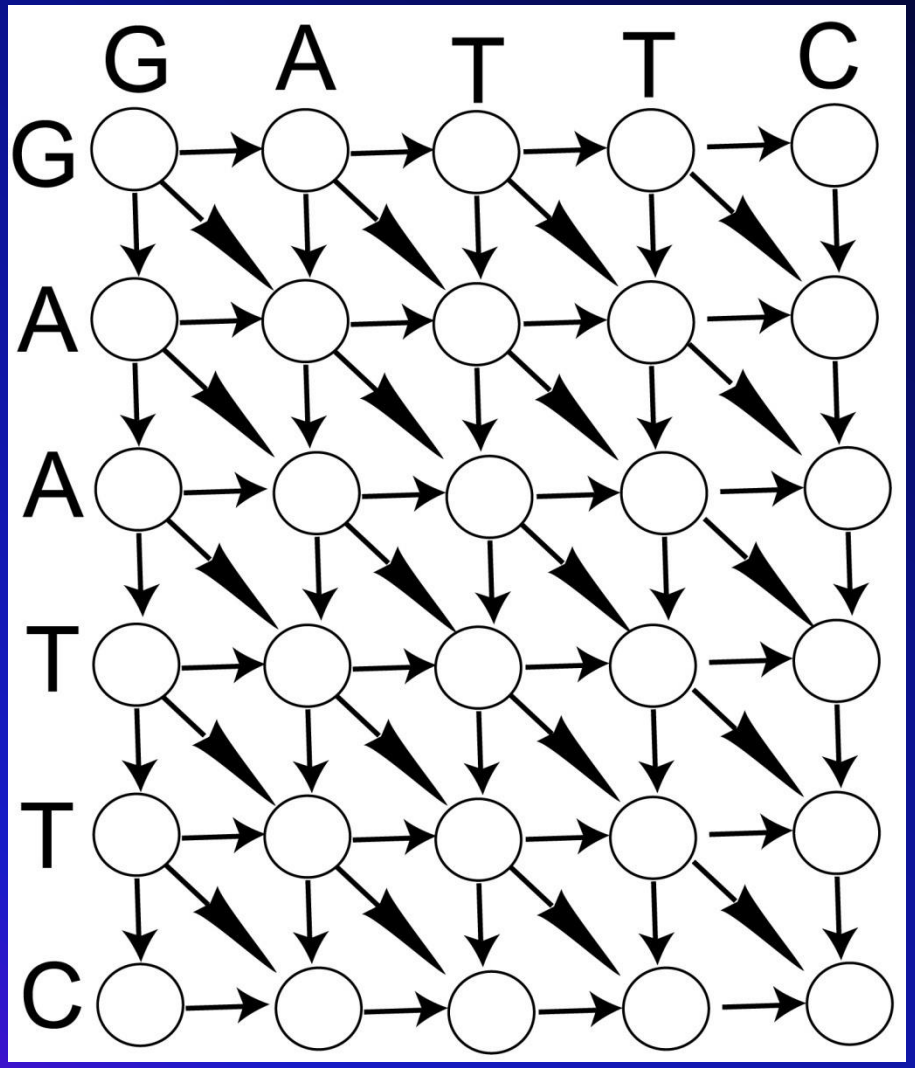
ATCCGT

?

TAGAGA

Графический метод путей следования для выравния двух сиквенсов (Path Graph).

1 - GATTC
2 - GAATTC



Возможное выравнивание (alignment)

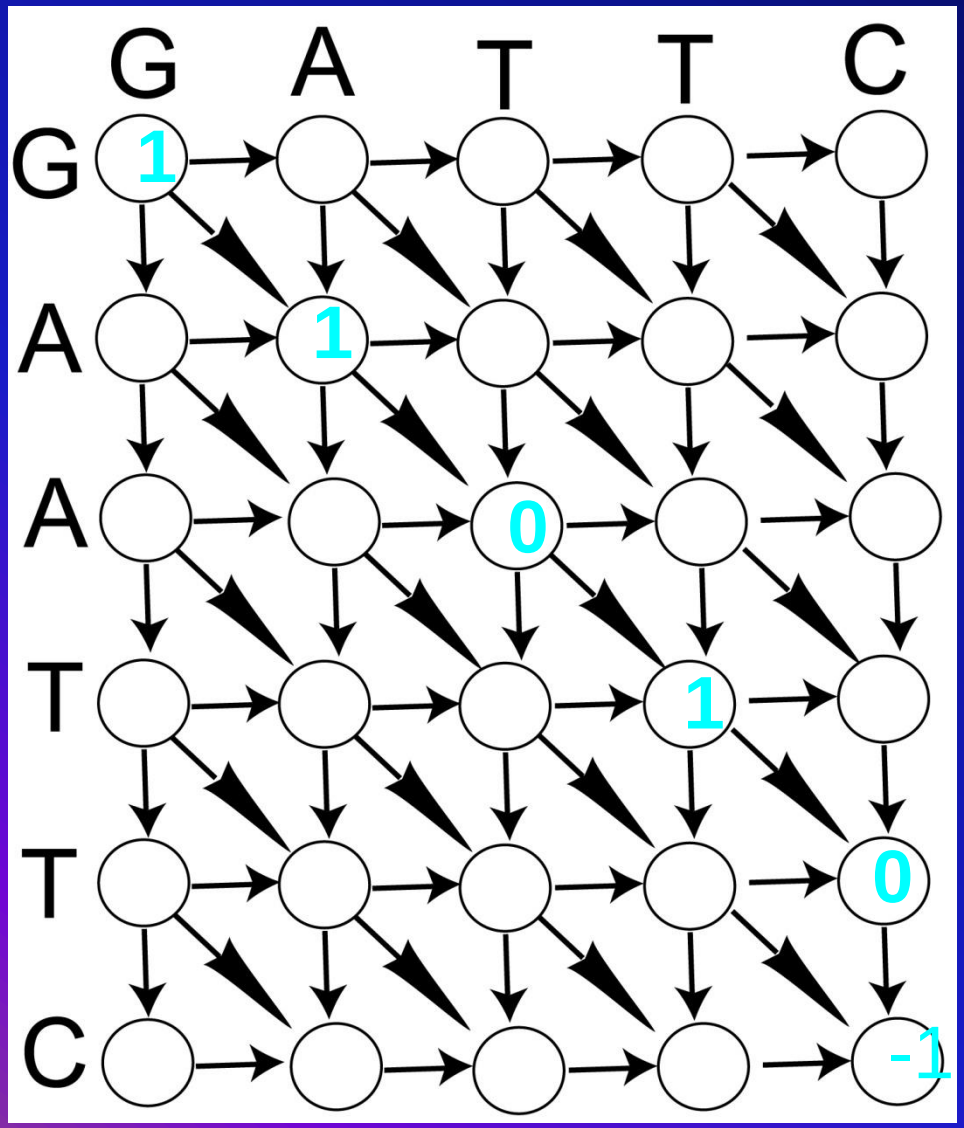


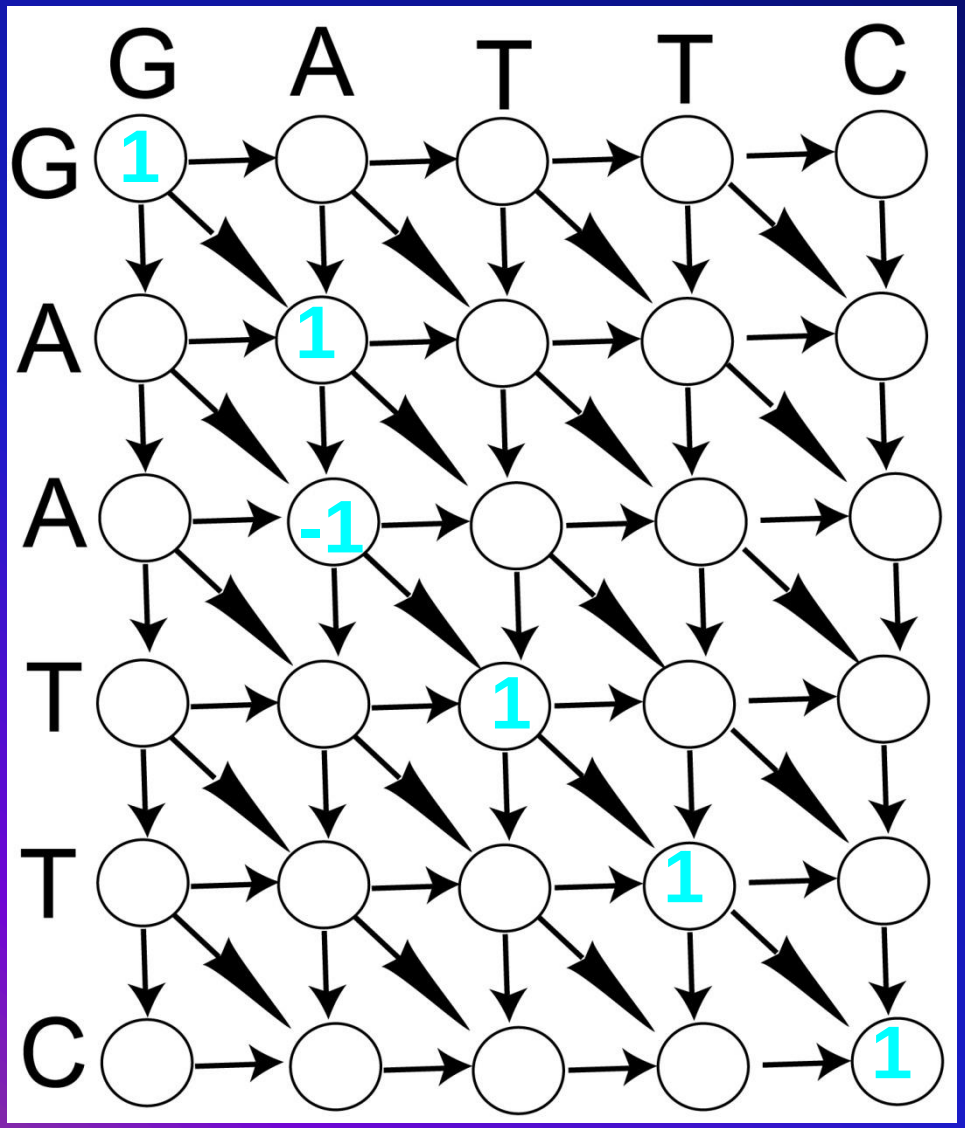
Схема анализа:

- **Match: +1**
- **Mismatch: 0**
- **Indel: -1**

Сумма для этого пути= 2

GATTC-
GAATTC

Оптимальное выравнивание 1

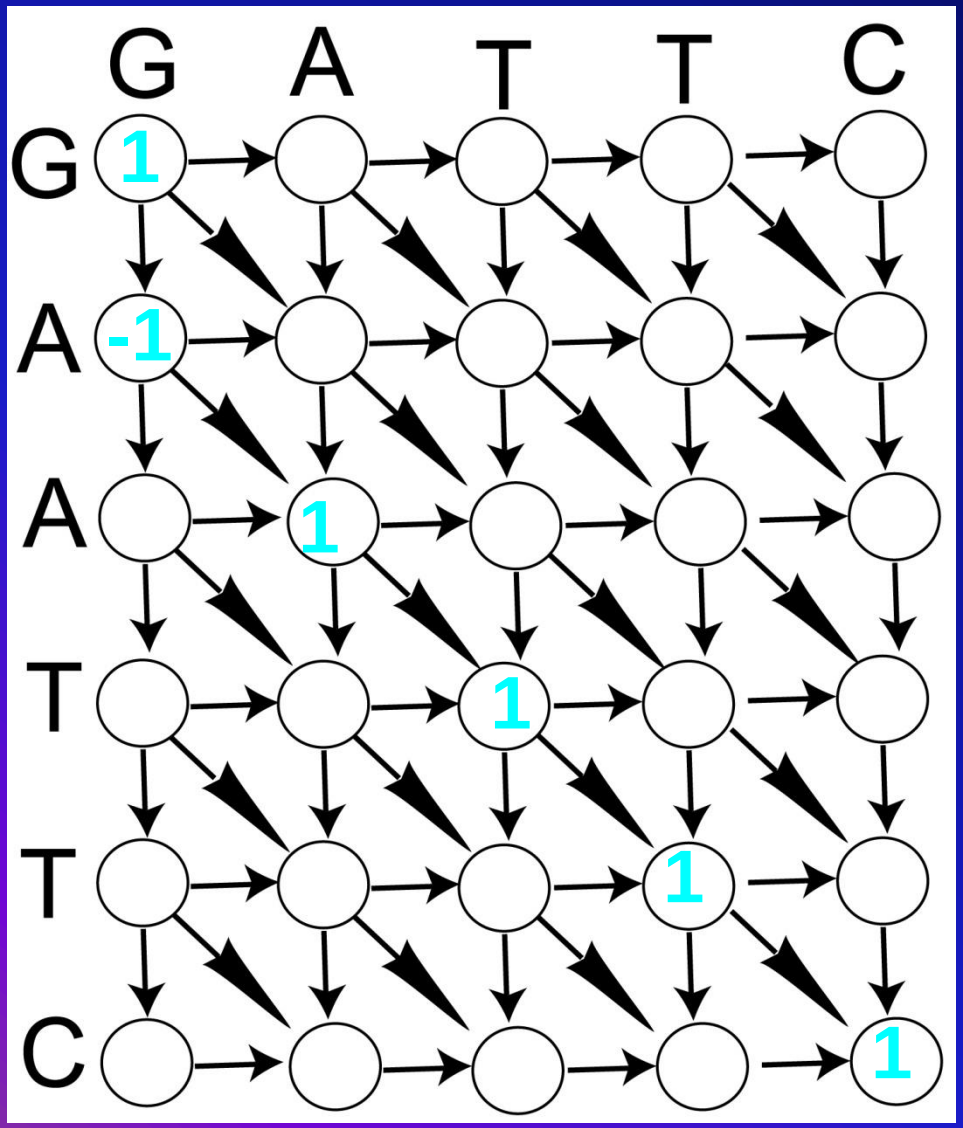


Выравнивание при использовании этого пути следования

GA-TTC
GAATTC

Сумма для этого пути= 4

Оптимальное выравнивание 2



Выравнивание при использовании этого пути следования

G–ATTC

GAATTC

Сумма для этого пути= 4

Alignment

TCAGACGATTG n=11
TCGGAGCTG n=9

TCAG-ACG-ATTG

TC-GGA-GC-T-G 6 gaps, 0 mismatches

TCAGACGATTG

TCGGAGCTG-- 1 gap, 5 mismatches

TCAGACGATTG

TCGGA-GC-TG 2 gaps, 2 mismatches

Задания для выравнивания (**Alignment**)

1)

Seq.1 AGGCTTGCGATGATCGGGTTAG

Seq. 2 AGGGCTATGATCGTGTTAG

2)

Seq. 1. GTCCTAATTTTTTTGACTGGATCTC

Seq. 2 GTCCTAATTTTTTGACCGGATCTC

3)

Seq. 1 ATCCGGTATAGTAGTAGGTCTAGCTGCTAT

Seq. 1 ATCCGGTATAGCTGATAT

4)

Seq. 1 TGCTAAGCCGTGTGATCTAGTCAAATGCGTGTGTAT

Seq. 2 TGCTAATGTGCCTGATCTGATCGCGTGTGTTT

Задания для выравнивания (**Alignment**)

```
1)
Seq.1       AGGCTTGCGATGATCGGGTTAG
            AGGGCT---ATGATCGTGTTAG
Seq. 2      AGGGCTATGATCGTGTTAG

2)
Seq. 1.     GTCCTAATTTTTTTGACTGGATCTC
            GTCCTAATTTTTT-GACCGGATCTC
Seq. 2      GTCCTAATTTTTTGACCGGATCTC

3)
Seq. 1      ATCCGGTATAGTAGTAGGTCTAGCTGCTAT
            ATCCGGTATAG----------CTGATAT
Seq. 1      ATCCGGTATAGCTGATAT

4)
Seq. 1      TGCTAAGCCGTGTGATCTAGTCAAATGCGTGTGTAT
            TGCTAA----TGTGCCTGATCTGATCGCGTGTGTTT
Seq. 2      TGCTAATGTGCCTGATCTGATCGCGTGTGTTT
```

Алгоритмы выравнивания

Искоммый алгоритм должен выбрать выравнивание с наименьшей дистанцией (наибольшим сходством) из всех возможных вариантов. Учитывая, что два сравнительно коротких сиквенса в 300 бп дает 10^{88} различных вариантов, то анализ всех этих вариантов является не самым экономным путем.

Методы выравнивания сиквенсов

–3 главных метода:

- Ручной (Manual)
- Компьютерный
- Комбинированный

Ручное выравнивание Почему?

- Может применяться:
  - Выравнивание простое.
  - Имеется дополнительная информация о структуре ДНК
  - Компьютерное выравнивание имеет локальные проблеммы
  - Компьютерное выравнивание можно проверить и откорректировать

ClustalW ПРИНЦИП

1. Каждые две ДНК последовательности выравниваются при помощи Needleman-Wunsch коэфициента сходства
2. Отсюда получают генетические дистанции.
3. На основе генетических дистанций расчитывается NJ-дерево
4. На основе этого модельного дерева составляется множественное выравнивание (multiples Alignment) ДНК последовательностей. Последовательно составляются вначале из соседних последовательностей в NJ-дереве блоки последовательностей и затем из блоков конструируется полное выравнивание всех последовательностей.

Филогенетические Деревья
– способ графического выражения эволюционных взаимоотношений

Элементы филогенетического дерева.

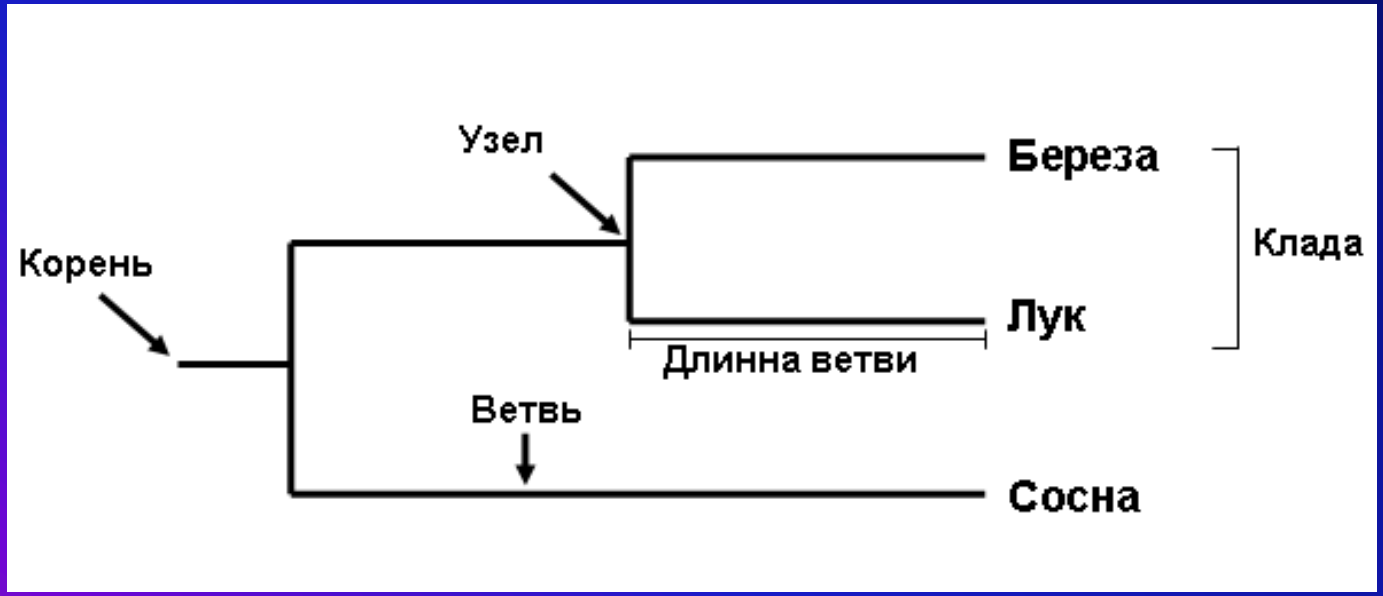
Филогенетическое дерево представляет собой математическую структуру, используемую для моделирования эволюционой истории группы последовательностей или организмов.

Филогения или эволюционное дерево графически отражают состояние исторических связей

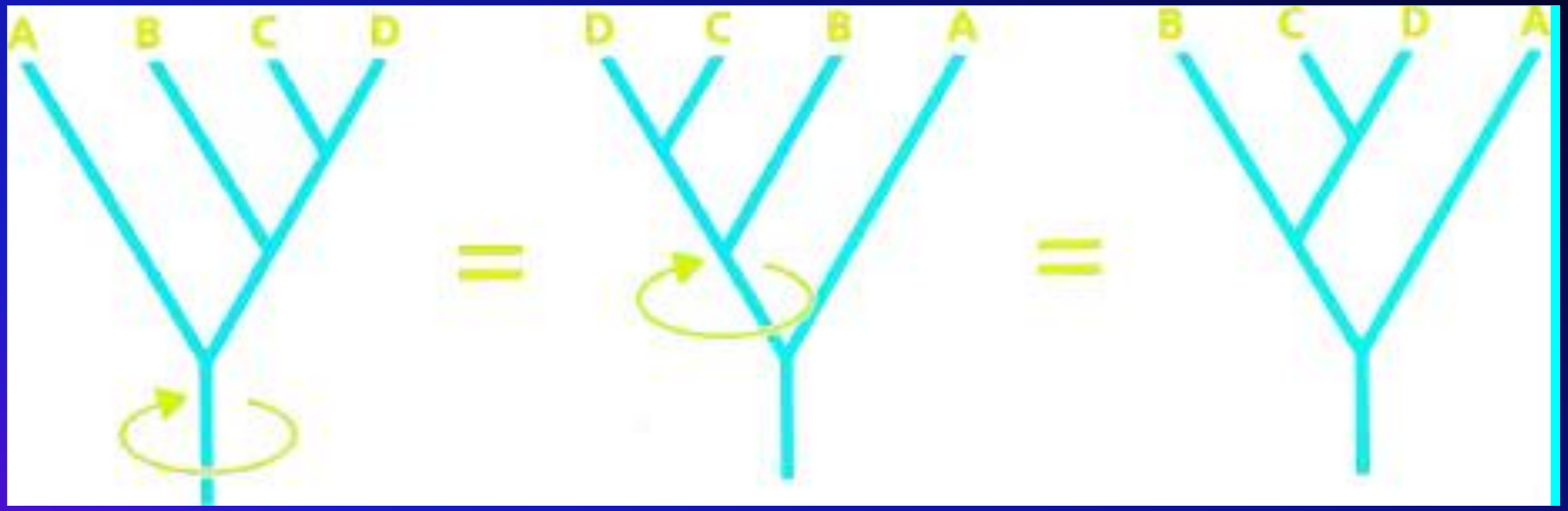
Филогенетическое дерево - это практически всегда
Гипотеза

Терминология филогенетического дерева

Дерево состоит из узлов (nodes), связаных между собой ветвями (Branches)
Терминальные узлы (Terminal nodes) (также именуемые OTU, оперативной таксономической единицей, или терминальным таксоном), представляют собой последовательности (sequences) генов или организмов, по которым мы располагаем данными. Они могут быть вымершими или живущими в настоящее время.
Внутренние узлы (Internal nodes) представляют собой гипотетических предков
Корень (root) является родоначальником всех последовательностей, составляющих дерево.
Клада (clade) - Комплекс видов, которые включают все таксоны происшедшие от одного общего предка.



У филогенетических деревьев надо правильно считыват информацию

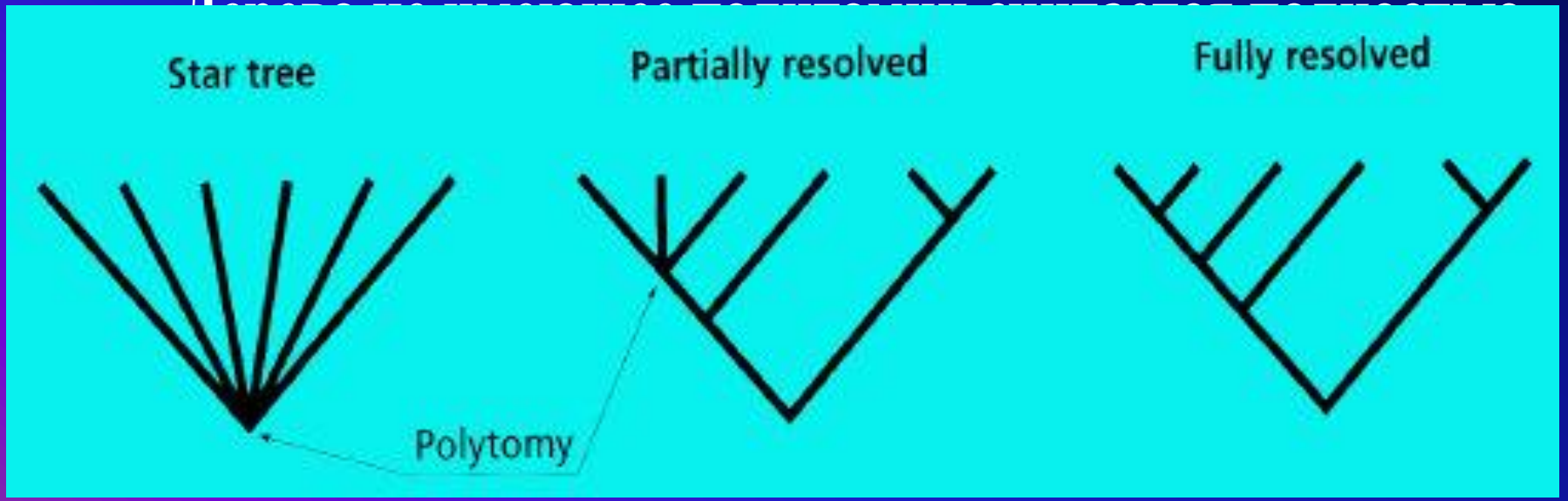


Все три дерева содержат одинаковую топологическую информацию

Узлы и Ветви дерева содержат различные типы информации ..

От числа ветвей отходящих от внутреннего узла зависит степень разрешения узла:

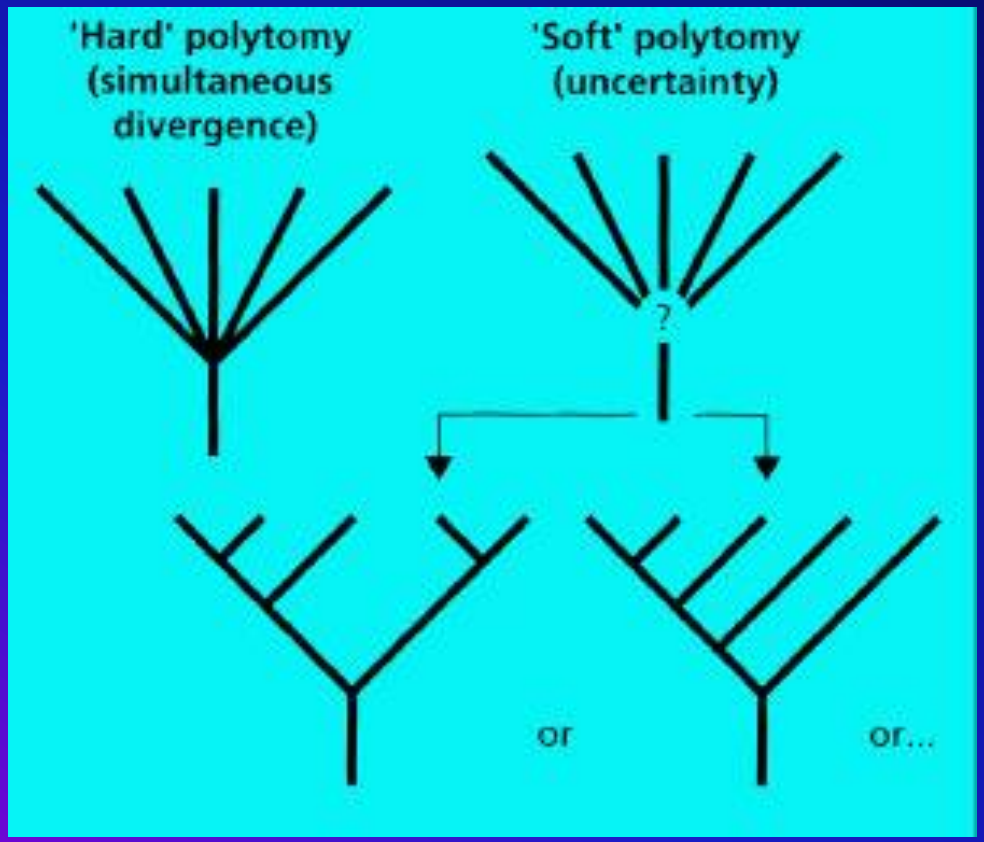
Если узел имеет степень больше, чем 3 (т. е. один предок и два ближайших потомка), это называется политомией (polytomy)



Имеется два типа политомий:

Жесткая политомия – все линии произошли одновременно от одного предка

Мягкая политомия – является отражением неуверенности. Все линии не обязательно произошли одновременноо, но мы не уверены в порядке расхождения.



Понятия моно- поли- и парафилии - ключевые термины в таксономи и филогенетике.

Монофилия (др. греч. μόνος — один и φυλή — семейный клан) — происхождение таксона от одного общего предка

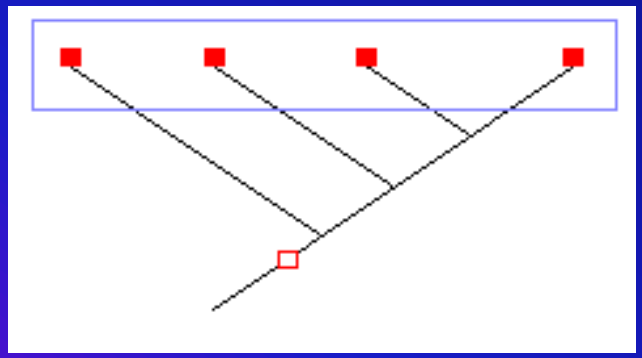
Полифилия (др. греч. πολύς — многочисленный и φυλή — семейный клан) — происхождение таксона от разных предков

Парафилия - Парафилетическая группа включает ближайшего общего предка, но в отличие от монофилетической, не всех потомков

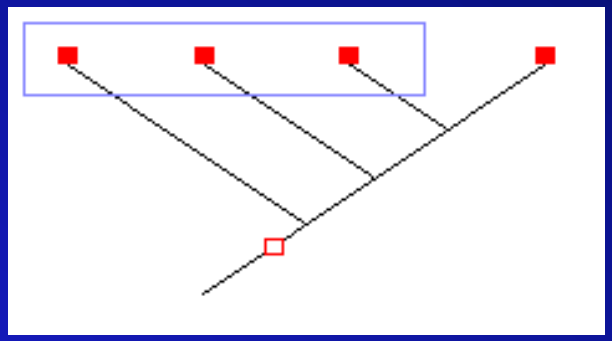
Парафилетическую группу невозможно охарактеризовать уникальными синапоморфиями. Все общие свойства, которые можно указать для её представителей относятся к симплезиоморфиям (унаследованы от более отдаленных предков, чем ближайший общий предок представителей группы) или гомоплазиям (возникли у разных представителей исследуемой группы независимо).

A: monophyletische Gruppe
B: paraphyletische Gruppe
C: polyphyletische Gruppe

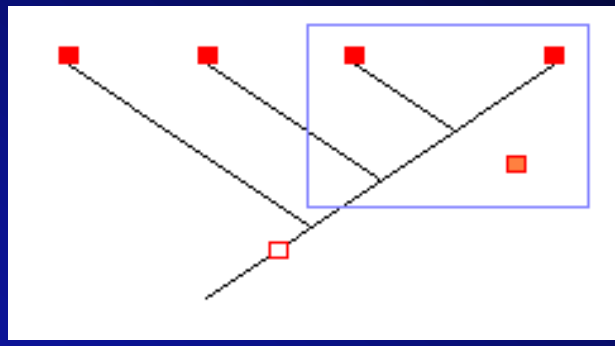
A



B



C



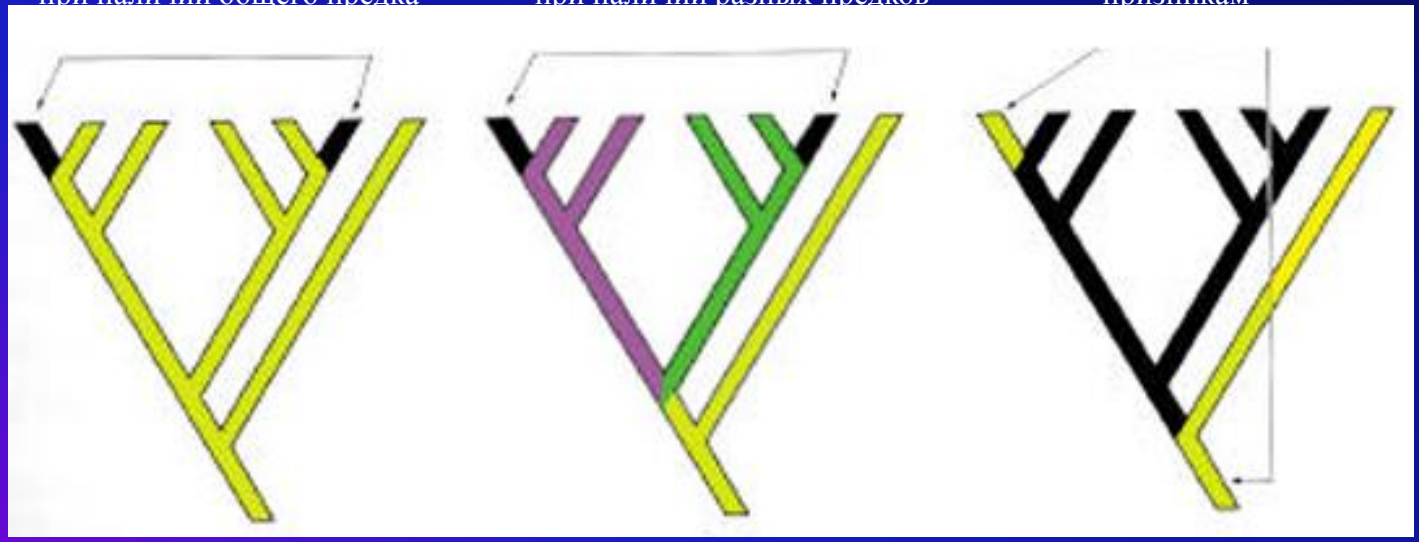
Три возможных типа гомологии признаков (Гомоплазии - Homoplasy).

(независимая эволюция одинаковых изменений)

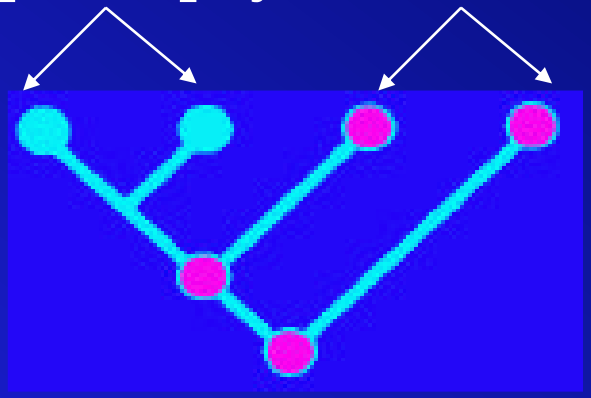
Паралельная эволюция
независимая эволюция одинаковых изменений при наличии общего предка

Конвергентная эволюция
независимая эволюция одинаковых изменений при наличии разных предков

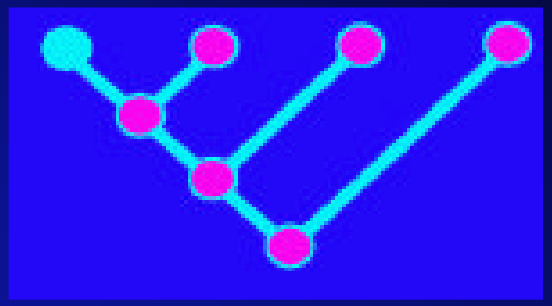
Повторная потеря
Возврат к предковым признпкам



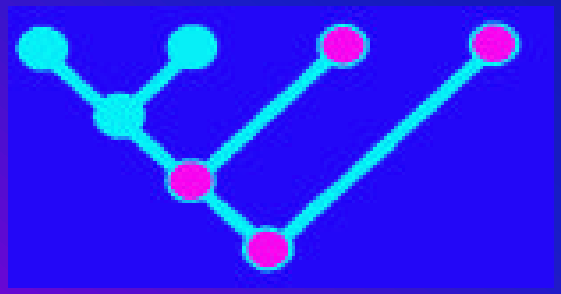
Apomorphy **Plesiomorphy**



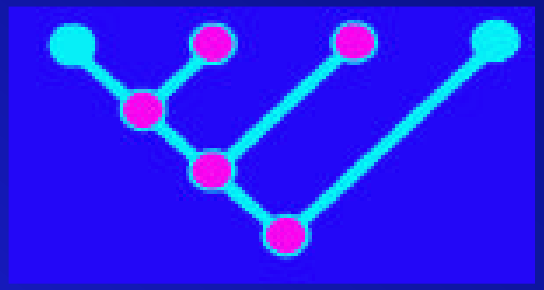
Autapomorphy



Synapomorphy

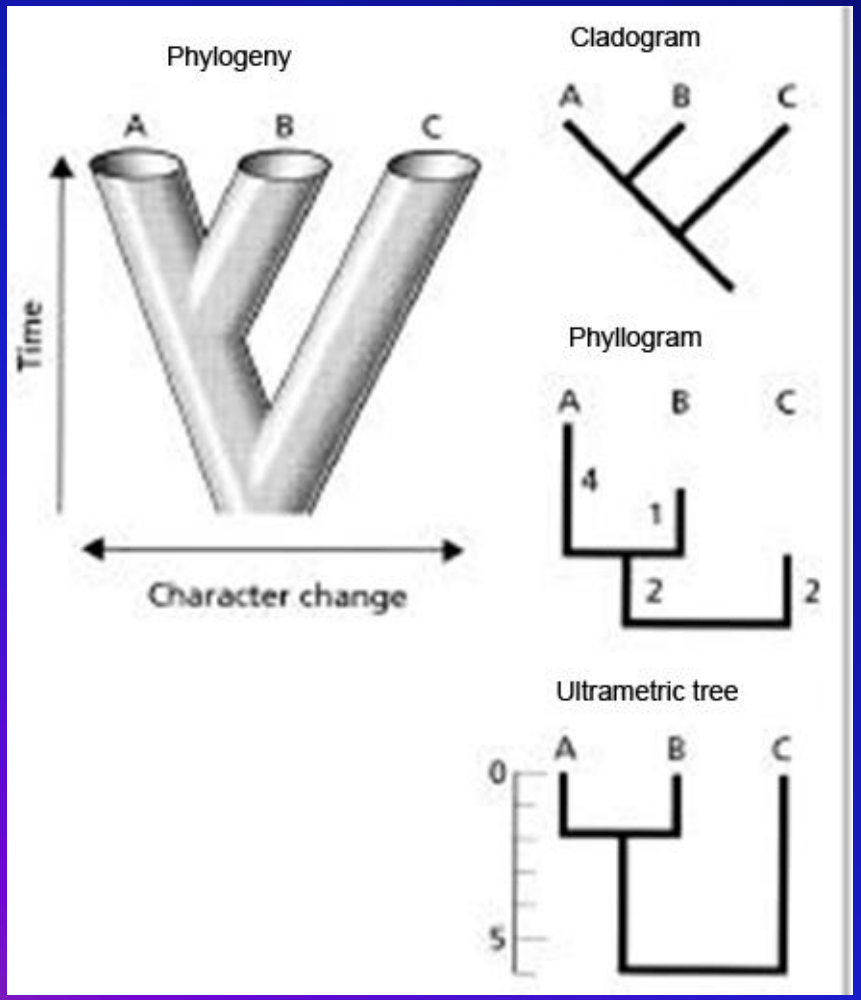


Homoplasy



Предковый статус – Розовый
Продвиннутый статус- голубой

Кладограммы, Филлограммы и Ультрамерные дереья

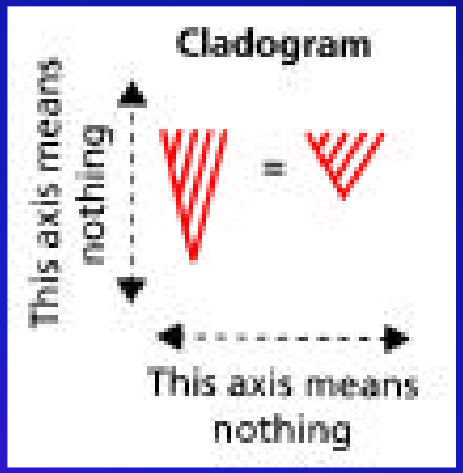


Кладограмма:
Показывает путь происхождение от одного предка

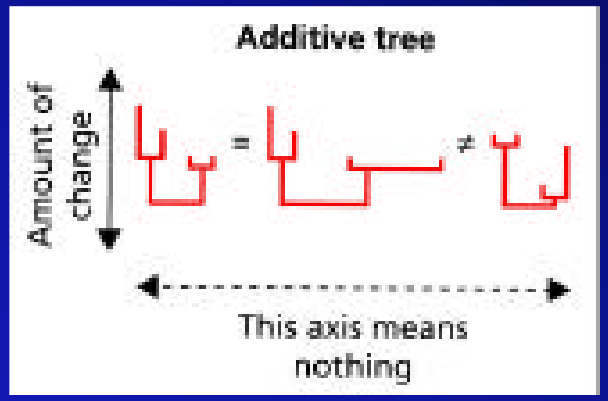
Филлограмма:
Дает дополнительную информацию о присхождении: длина ветвей.
Числа отражают количество еволюционных изменений на ветвях.

Хронограмма
Особый вид дерева, в рамках которого все еволюционные изменения пересчитаны (откалиброваны) на эволюционное время. Эволюционное время выражено либо в виде дивергенции последовательностей или непосредственно в годах.

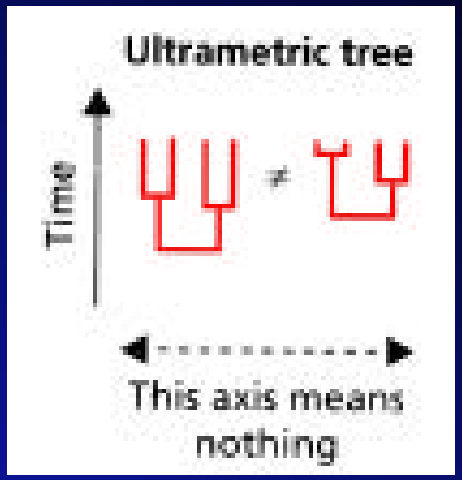
Что отражают горизонтальные и вертикальные оси у раличных типов деревьев?



Только последовательность дивергенции



Число еволюционных изменений на ветвях.

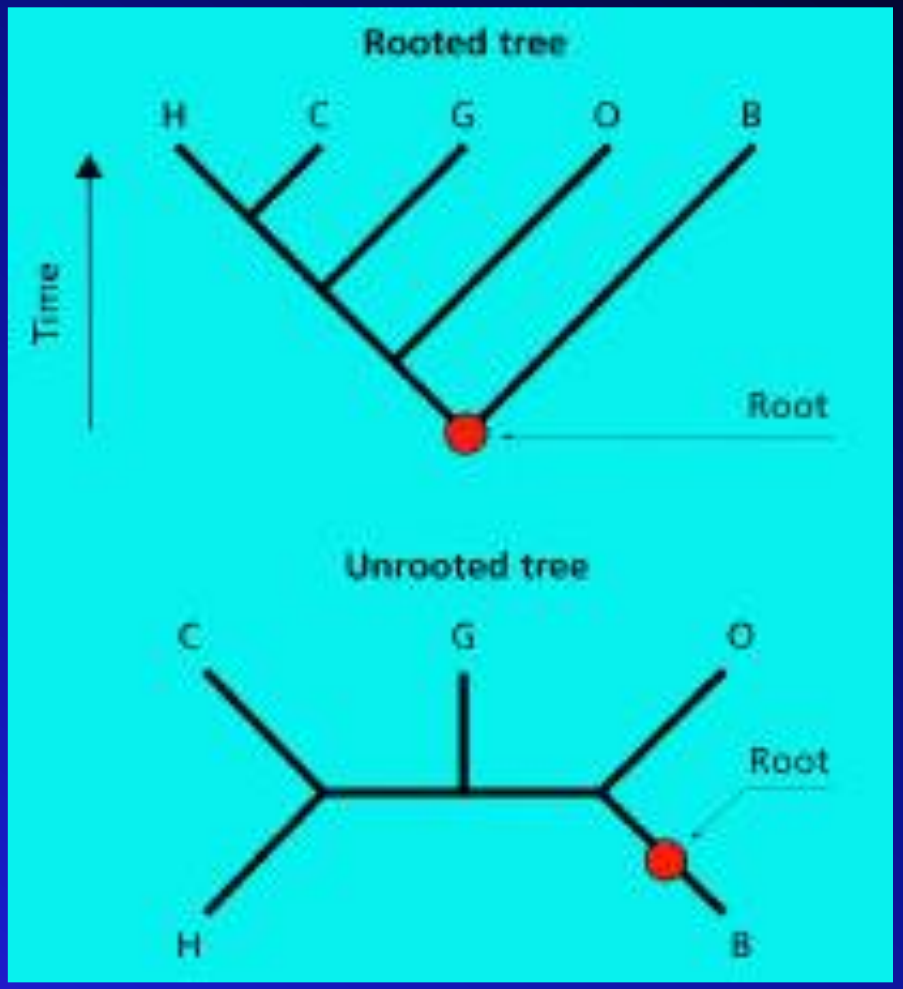


Число еволюционных изменений на ветвях хорошо корелирует с временем

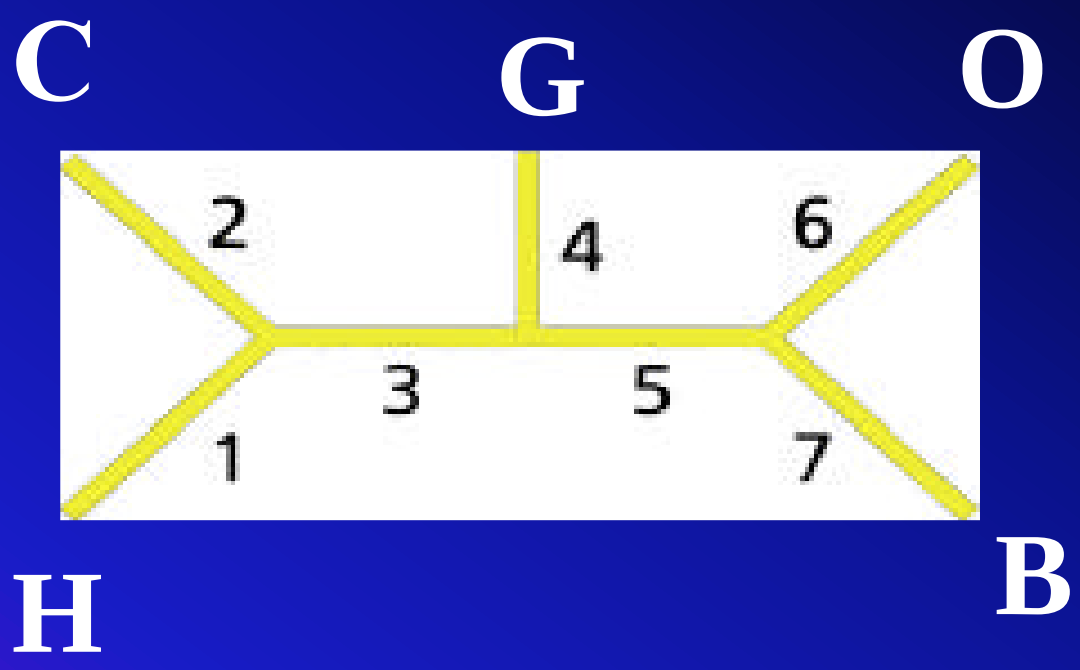
Укорененные (rooted) и неукорененные (unrooted) деревья

Укорененные деревья: показывают направление от предка к потомкам

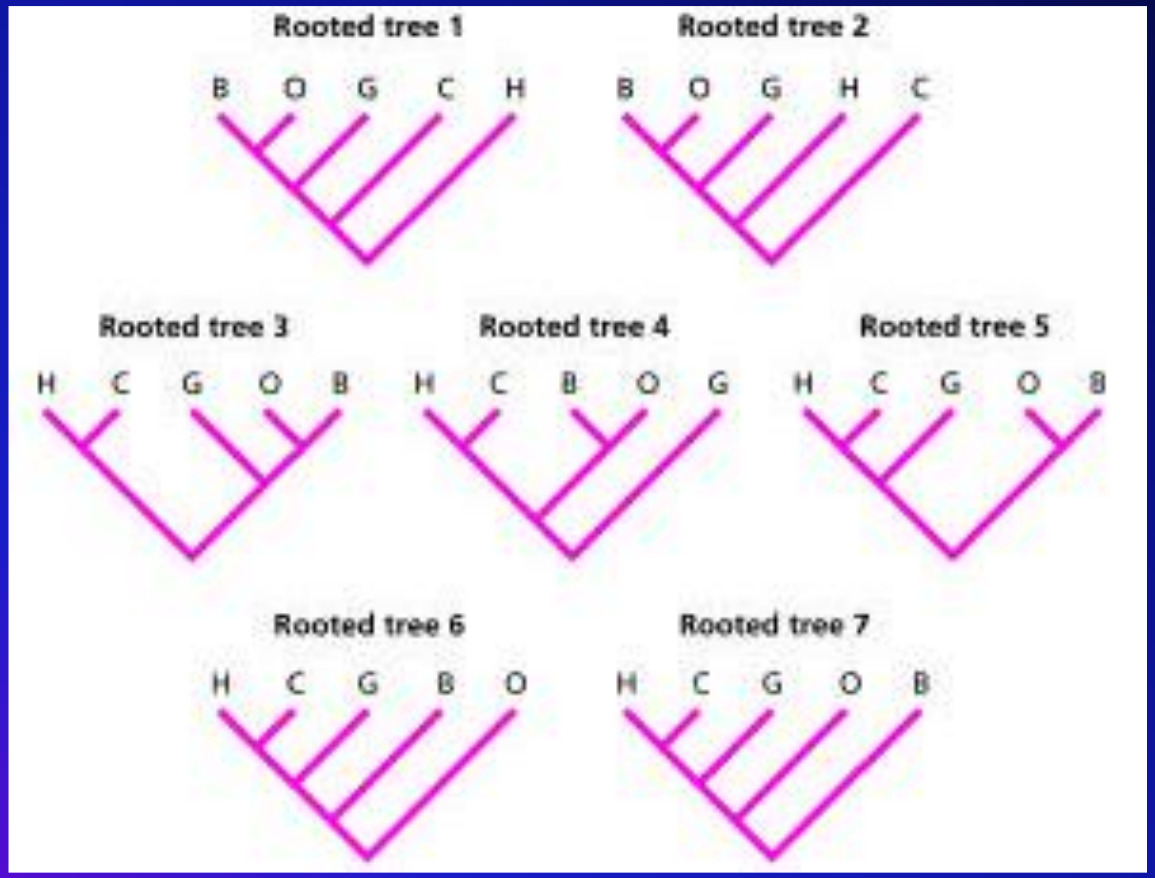
Неукорененные деревья: Мы не можем говорить о предках и потомках, а только о дивергенции



Например: Для этого неукорененного дерева.....



есть 7 отличающихся соответствующих укорененных деревьев!



Очень важно различать укорененные и неукорененные деревья

Поскольку многие филогенетические методы реконструкции генерируют **неукорененные деревья** и не могут самостоятельно выявить различие между этими семью **укорененных деревьев**.

Количество возможных неукорененных деревьев **Un** для **n** сиквенсов

$$Un = (2n-5)\ (2n-7).....(etc)$$

Количество возможных укорененных деревьев **Rn** для **n** сиквенсов

$$Rn = (2n-3)\ (2n-5).....(etc)$$

Количество возможных деревьев возрастает в геометрической прогресии при возрастании количества сиквенсов

Количество Сиквенсов	Количество неукорененных деревьев	Количество укорененных деревьев
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	344594425

Методы построения деревьев

Фенетические или дистанционные методы

Деревья строятся на основе определенных генетических дистанций

Кладистические методы

Деревья строятся на основании анализа дискретных признаков - апоморфий

Основное предположение кладистики заключается в том, что члены группы имеют общую эволюционную историю.

Фенетические методы

UPGMA - (**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean)
МЭ - метод минимальной эволюции
NJ - метод ближайшего связывания

Филограмма, полученная дистанционными методами, не отражает эволюционного процесса, а только демонстрирует конечную степень дивергенции таксонов.

UPGMA

UPGMA является простейшим методом оценки филогенетических взаимоотношений на основе генетических дистанций. Шаг за шагом, похожие виды суммируются к новым единицам (OTUs, operational taxonomic units)

UPGMA предполагает постоянную эволюционную изменчивость (молекулярные часы).

МЭ - метод минимальной эволюции

-Общая длина всех ветвей на филогенетическом дереве должна быть минимальной

- Длина ветвей оценивается в генетической дистанции (Число замещений на нуклеотид), расчитанные с помощью определенных формул.

-Кимура (Kimura, 1980) предложил метод, который учитывает химические особенности мутационного процесса. А именно, неравную вероятность транзиций.

NJ - метод ближайшего связывания

Метод **NJ** основан на алгоритмической апроксимации дерева минимальной эволюции.

Из полностью неразрешенного дерева (звезда-дерево), шаг за шагом суммируются пары OTUs до тех пор, пока не будет построено дихотомическое дерево. При каждом конструкционом шаге используется апринцип минимальной эволюции (Minimum-Evolution-Prinzip).

Кладистические методы

MP - метод максимальной экономии (Maximum Parsimony)

ML - метод максимального правдоподобия (Maximum Likelihood)

Метод Баесовой вероятности (Bayesian probability)

Maximum Parsimony
Принцип максимальной экономии

– Общие, видоизмененные признаки как основа для классификации

– Принцип экономичности (“Ockham’s razor”)

Критично: Проблема выявления гомологий

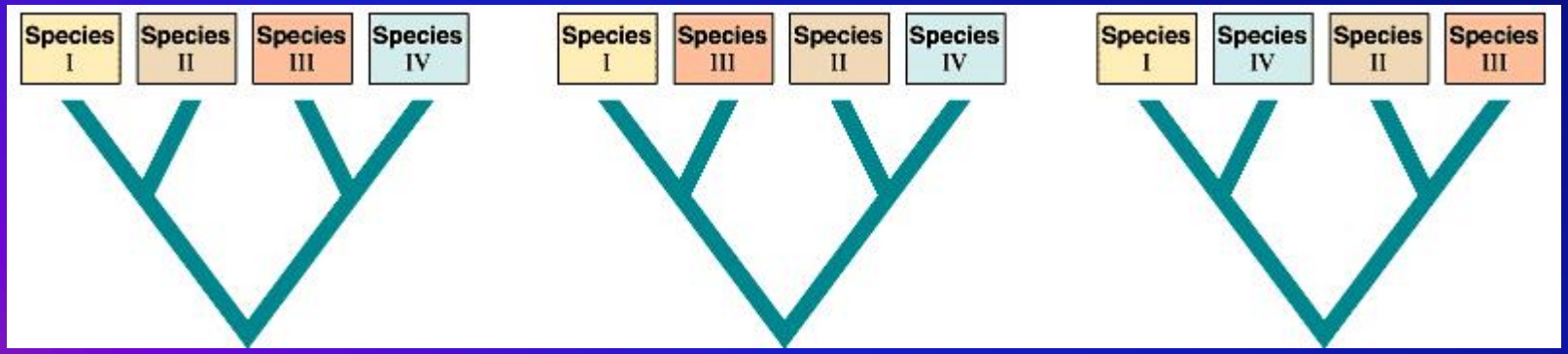
Эволюция не всегда имеет цель и не всегда экономична

Проблемма 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

Принцип парсимонии помогает систематикам реконструировать филогению

- 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.



Maximum Parsimony

- При включении в анализ все больше и больше деревьев, количество возможных деревьев будет увеличиваться драматически.
  - для группы из 50 видов имеются 3 x 10^{76} возможных филогенетических деревьев

Даже используя компьютер, анализ подобного обьема данных для поиска наилучшего дерева, будет длится очень долго.

Maximum Parsimony

- Систематики используют принцип парсимонии (бережливости), чтобы выбрать среди множества возможных деревьев одно дерево, которое наилучшим образом отражает анализируемые данные.
- В филогенетическом анализе, парсимония исползуется для выбора дерева, для получения которого было необходимо наименьшее количество эволюционых изменений.
- Принцип парсимонии (“Occam’s Razor”) назван в честь английского философа 14 века - William of Occam.

Maximum Parsimony
Принцип максимальной экономии

-Анализирует все возможные топологии деревьев

-эволюционный путь признака должен быть таким, который требует наименьшего числа его преобразований

- Выбирает дерево с найменьшим числом изменений

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

Maximum Likelihood

Ме́тод максима́льного правдоподо́бия

Имеются два объяснения для особого решения, который мы должны выбрать?

- Объяснение, которое делает наблюдаемое решение найболее правдоподобным...
- Более формально если имеются некоторые данные D и гипотеза H, вероятность которой 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).**
- **Найболее вероятное дерево, которое получается на основе полученных данных является максима́льно правдопода́бным вариантом филогении.**

Важно раличать между правдоподобием и вероятностью

- **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 =1**
- **Но мы заинтересованны только в одном наборе данных, который мы получили**

Примечание: правдоподобие (likelihood) это не вероятность, что полученное дерево - это правильное дерево, а просто махимальная вероятность дерева на базе полученных ДАННЫХ.

Другой путь для понимания likelihood

Пример трех кубиков

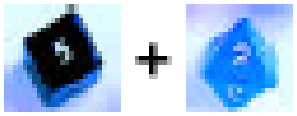
«Кубик» с 6 сторонами с 8 сторонами с 12 сторонами

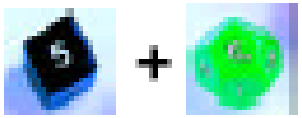
Бросаются два «кубика» и получают сумму “14”

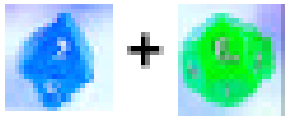
Какая пара кубиков максимально правдоподобна для получения этого результата?

Эквивалентно: какое дерево максимально правдоподобно при полученных сиквенсах

Сколько возможных вариатов для получения суммы “14” для каждой пары?







6 + 8

1

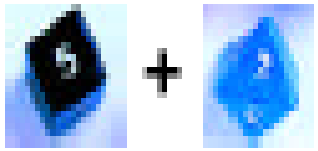
2 + 12
3 + 11
4 + 10
5 + 9
6 + 8

5

2 + 12
3 + 11
4 + 10
5 + 9
6 + 8
7 + 7
8 + 6

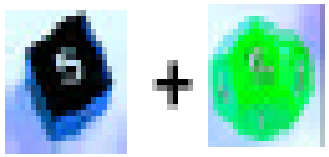
7

Какова вероятность каждого отдельного результата для каждой пары?



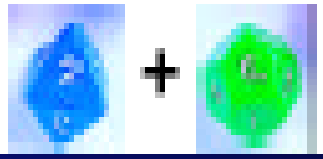
1/6 x 1/8

= 1/48



1/6 x 1/12

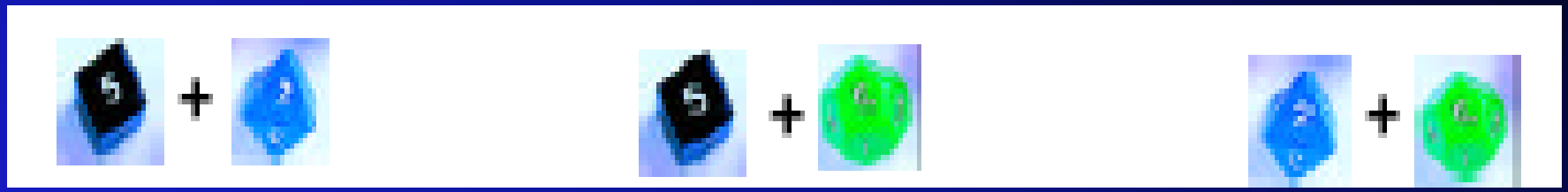
= 1/ 72



1/8 x 1/12

= 1/96

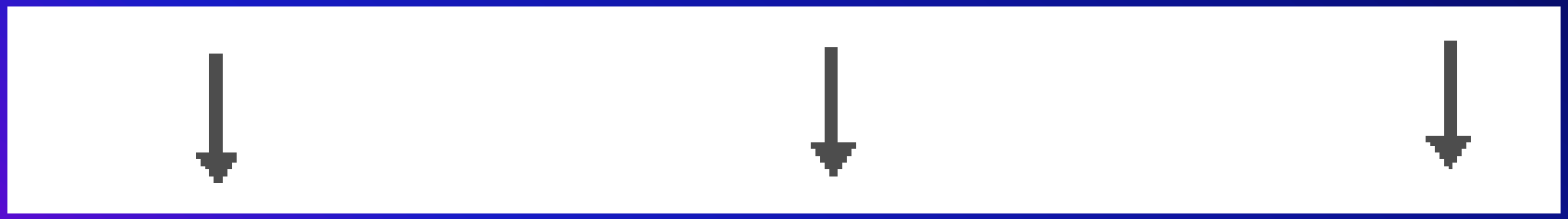
Теперь умножаем варианты для получения суммы “14” на количество вероятных сочетаний в каждой паре для получения правдоподобия (likelihood).



1/48 x 1

1/ 72 x 5

1/96 x 7



0.0729

0.0694

0.0208

maximum likelihood

Баесова Вероятность
Bayesian Probability

вероятности представляет собой математические алгоритмы для разрешения, или аргументации, используя вероятности.

Одним из важнейших элементов Баесовой теории, это мнение, что вероятность переводится в гипотезу.

Баесова теория предусматривает стандартный набор процедур и формул для выполнения такого расчета.

MrBayes

Метод Байесовой оценки филогении – Оценка степени правдоподобия филогенетической реконструкции относительно априори заданной "эволюции" признаков на основе байесовых вероятностей

Последующее распределение вероятности деревьев невозможно рассчитать аналитически; вместо этого программа MrBayes использует Метод Моделирования, который называется Марков цепи Монте-Карло (или
mcmc) для оценки степени правдоподобия филогенетической реконструкции .

Как и **Ме́тод максима́льного правдоподо́бия** MrBayes требует модель эволюции ДНК, которая определяется с помощью программы **Modeltest**

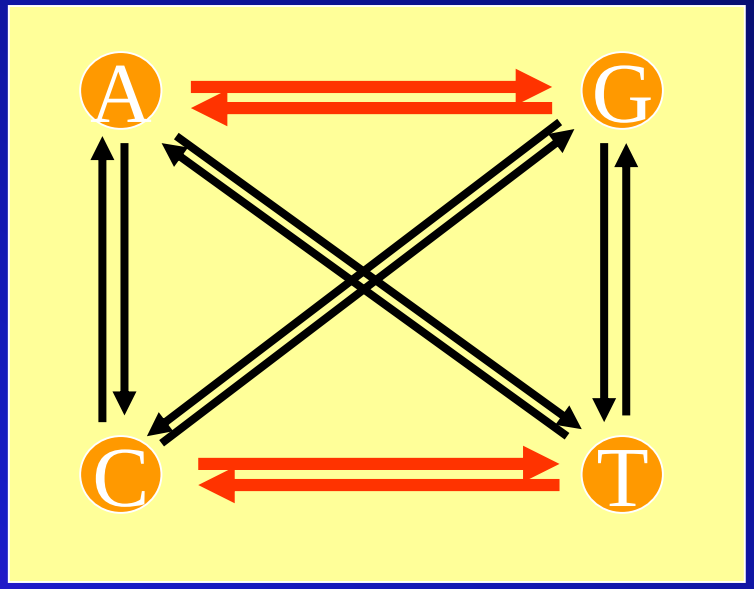
Три филогенетические программы используют Баесов принцип - MrBayes, BEST и BEAST.

MrBayes: Программа для Байесовой оценки филогении (Redelings & Suchard, 2005).
http://sourceforge.net/projects/mrbayes/

BEAST: (**B**ayesian **E**volutionary **A**nalysis **S**ampling **T**rees) - Байесов эволюционный анализ базы филогенетических деревьев
http://evolve.zoo.ox.ac.uk/beast/primateTutorial.html

BEST (Bayesian **E**stimation of **S**pecies **T**rees**)** Байесовые вероятности филогенетических деревьев
http://www.stat.osu.edu/~dkp/BEST/

Модель эволюции ДНК Modeltest



Transitions

Transversions

Transitions

56 вероятных моделей эволюции ДНК

Модель эволюции ДНК определяется с помощью программ **Modeltest, Modeltest2.**

Качество филогенетических деревьев

- Bootstrap
- Jack Knife
- Decay Index
- Bayesian probability

Network construction

Chloroplast haplotype genealogy (network analysis)

AACGTAACAGCGTTCC

AACGTAACAGCGTTTC

AACATAAGATCGTTTC

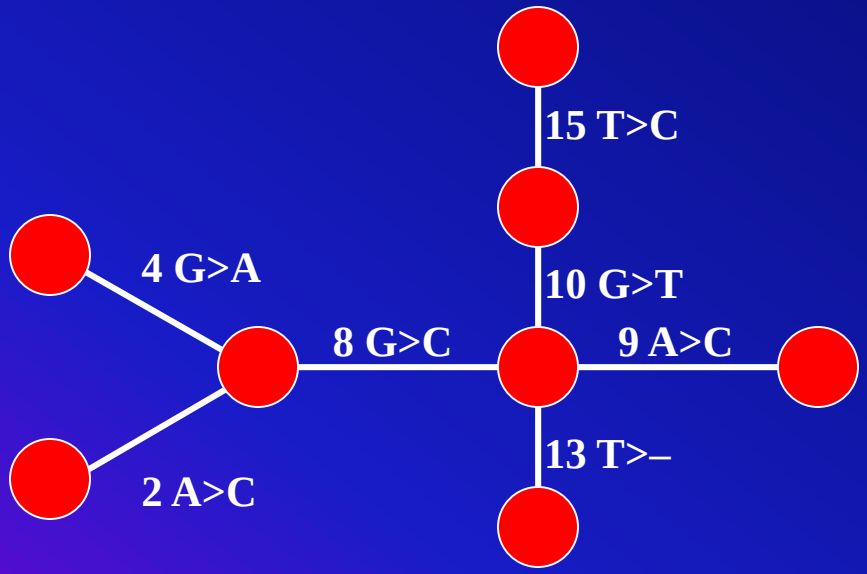
AACGTAAGATCGTTTC

AACGTAACATCGTTTC

AACGTAACCTCGTTTC

ACCGTAAGATCGTTTC

AACGTAACATCG–TTC



Chloroplast haplotype genealogy (network analysis)

AACGTAACAGCGTTCC

AACATAAGATCGTTTC

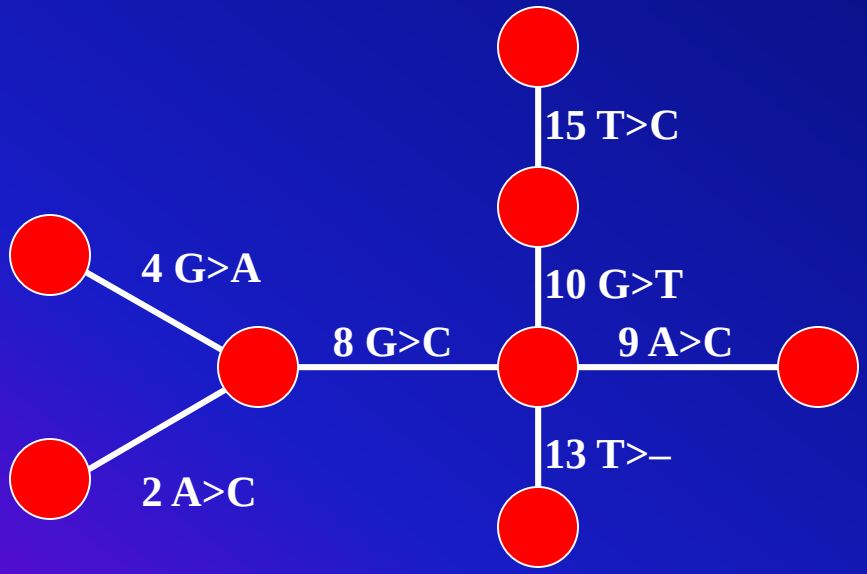
AACGTAAGATCGTTTC

AACGTAACATCGTTTC

AACGTAACCTCGTTTC

ACCGTAAGATCGTTTC

AACGTAACATCG–TTC



Chloroplast haplotype genealogy (network analysis)

AACGTAACAGCGTTCC

AACATAAGATCGTTTC

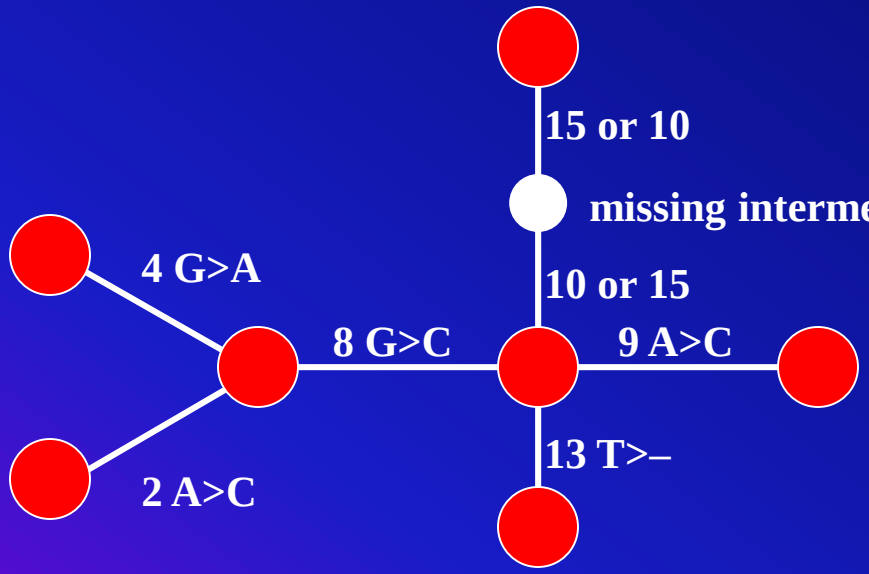
AACGTAAGATCGTTTC

AACGTAACATCGTTTC

AACGTAACCTCGTTTC

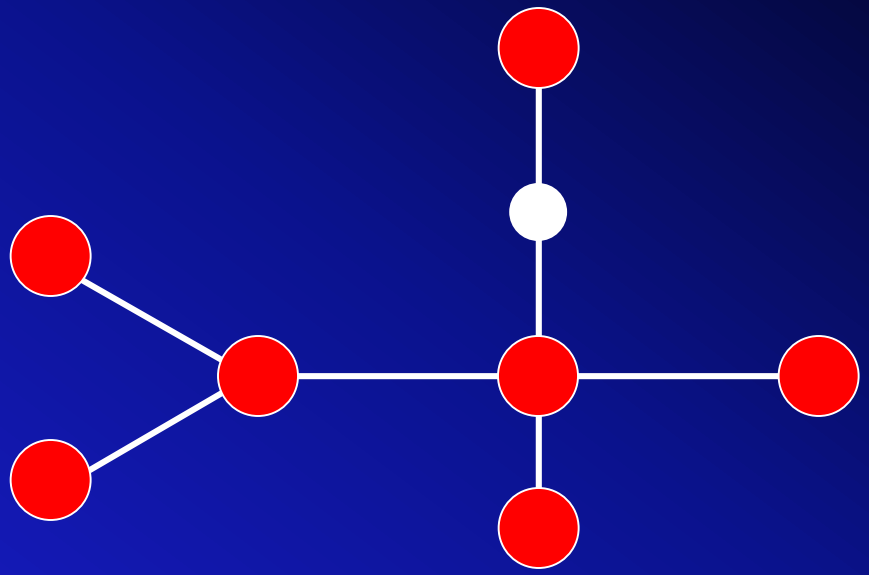
ACCGTAAGATCGTTTC

AACGTAACATCG–TTC

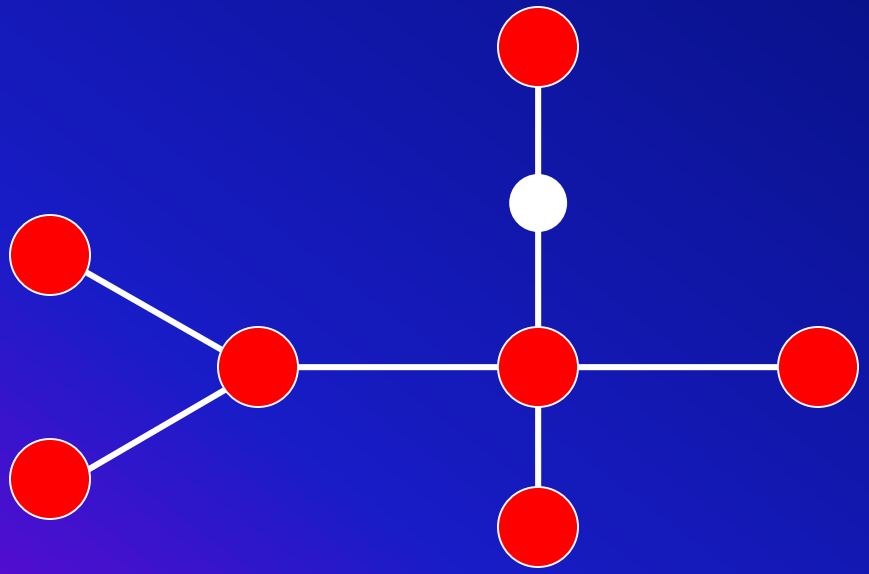


Chloroplast haplotype genealogy (network analysis)

Phylogenetic tree

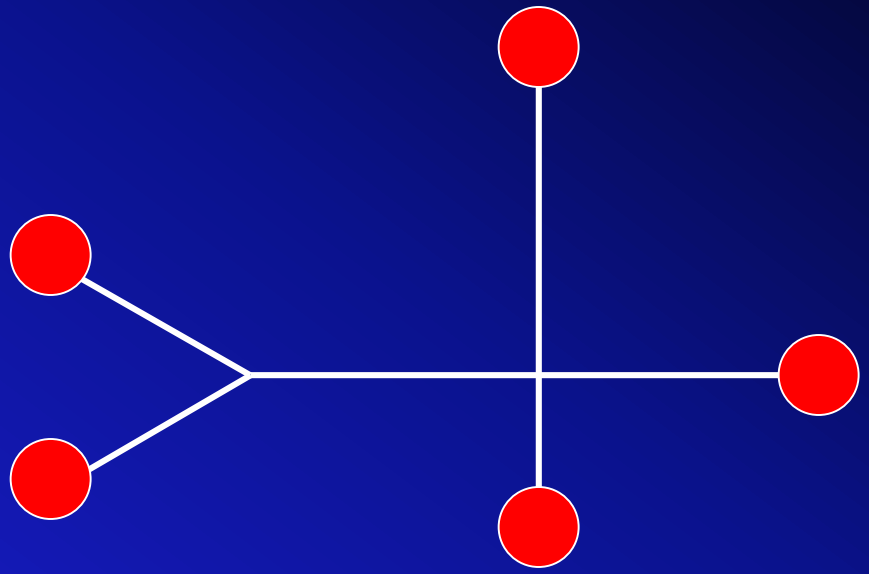


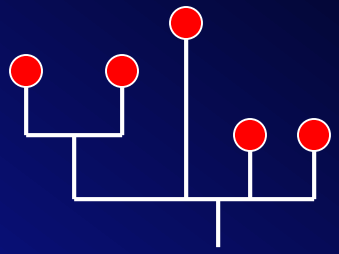
Genealogy/network



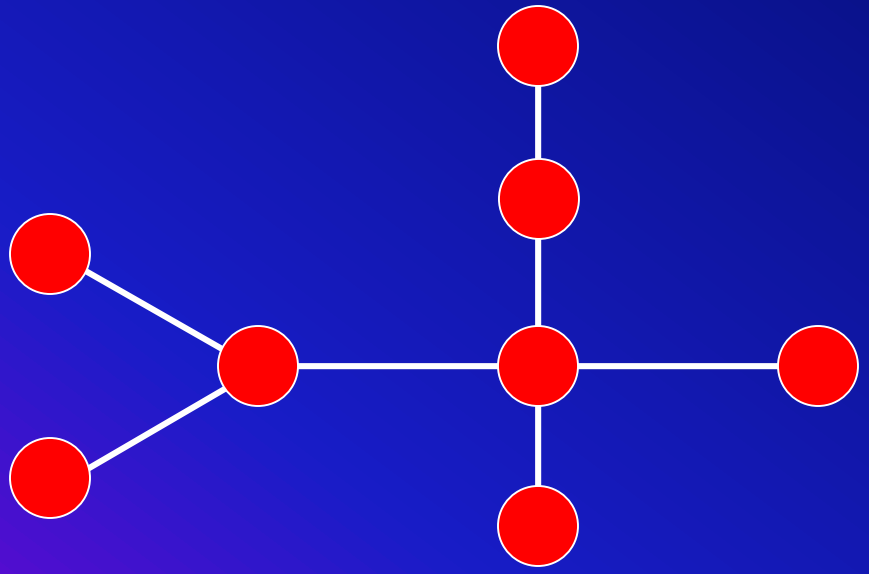
Chloroplast haplotype genealogy (network analysis)

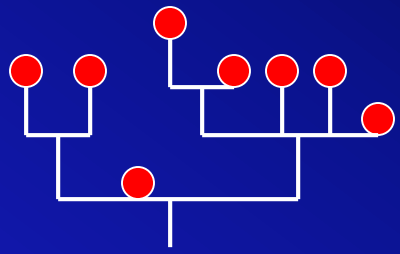
Phylogenetic tree





Genealogy/network





Chloroplast haplotype genealogy (network analysis) of genus Hordeum

(Jacobs and Blatner, 2011)