Филогенетический анализ и молекулярная эволюция

Лекция 5

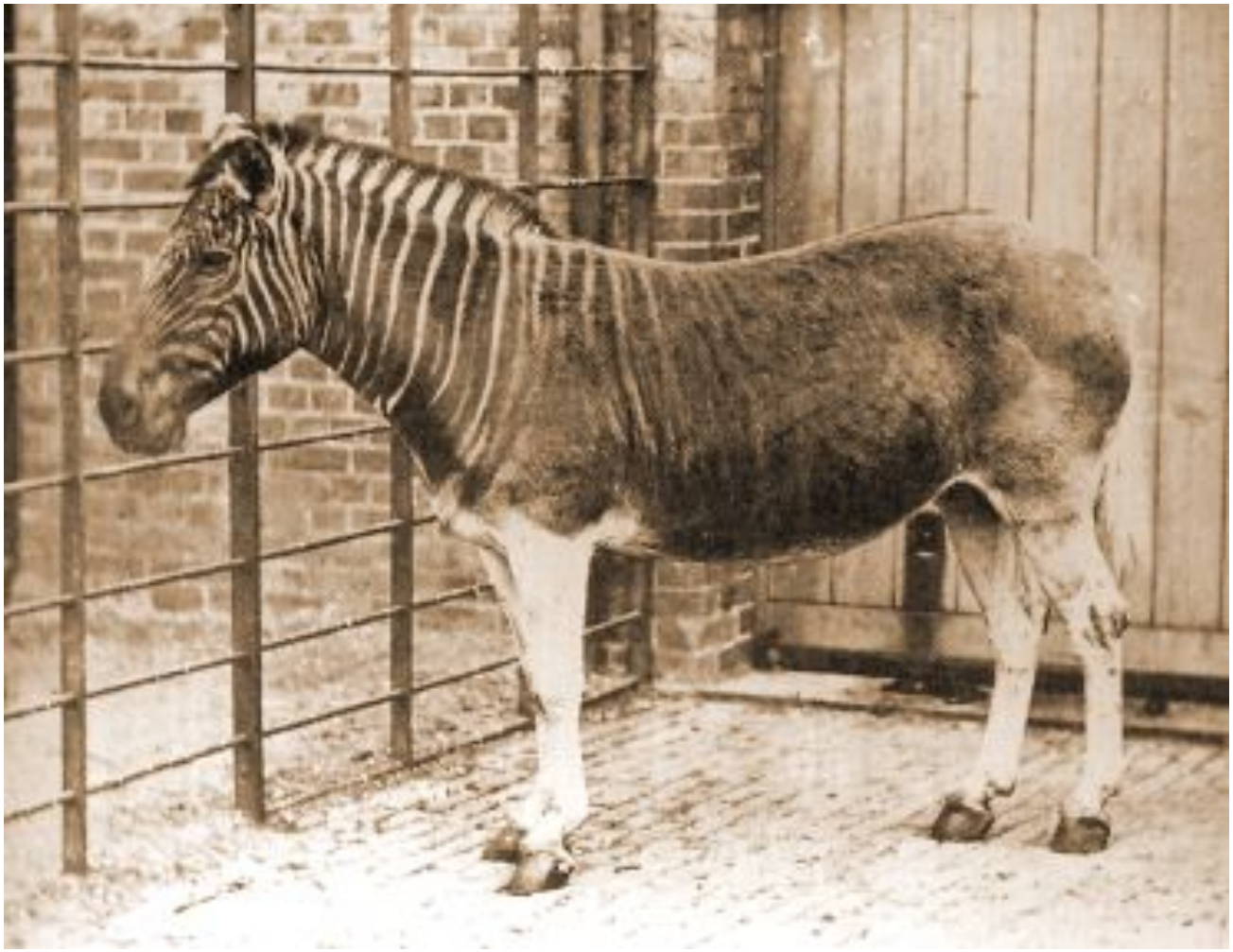
Задачи молекулярной эволюции

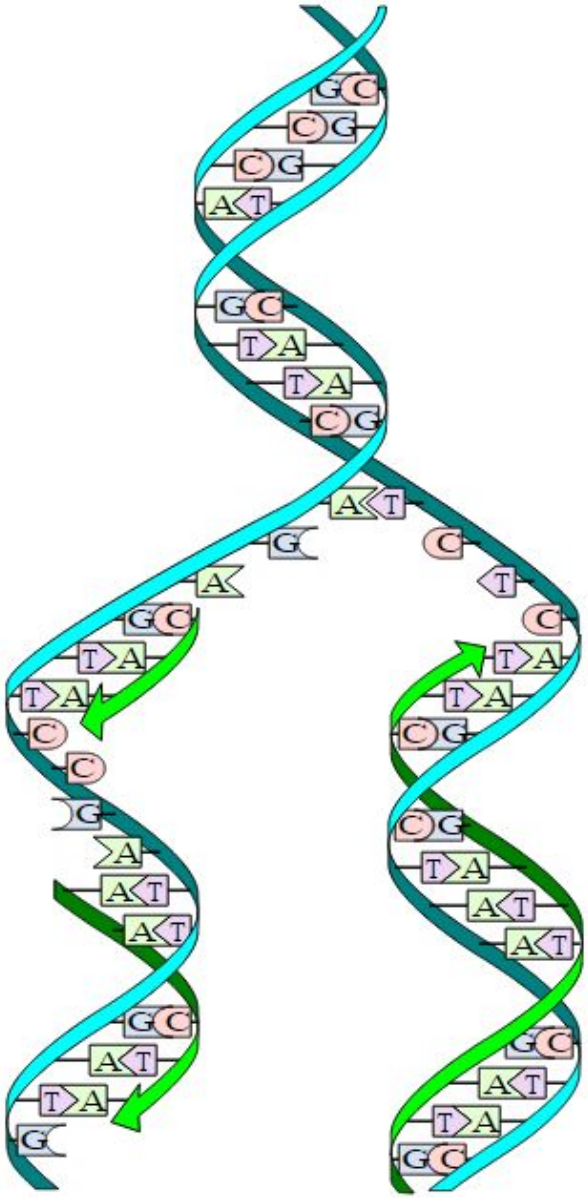
- Изучение законов изменения наследственной информации в живых системах, включая неклеточные и клеточные формы жизни
- Изучение истории развития жизни на Земле, установление родственных отношений между формами жизни (филогения)

Филогенетический анализ и молекулярная эволюция

- На молекулярном уровне эволюция является процессом мутации и селекции
- Молекулярная эволюция изучает изменения генов и белков на протяжении разных ветвей древа жизни.
- Филогенез является представлением эволюционных отношений. Традиционно, филогения опиралась на сопоставлении морфологических признаков между организмами. В настоящее время данные о молекулярных последовательностях также используются для филогенетического анализа.

Квагга (вымерла) больше похожа на зебру или лошадь?





Мутации

- Ошибки, происходящие при репликации генома
- Могут происходить как в половых, так и в соматических клетках
- Соматические мутации не наследуются в поколениях, исключены из эволюционного процесса

Классификации мутаций

- По числу затронутых нуклеотидов (по длине мутации)

  Точечные мутации (point mutations)

  Изменения в нескольких соседних нуклеотидах

Классификации мутаций

- По типам событий, происходящих при мутации

  Замена одного нуклеотида на другой (nucleotide substitution)

  Вставка одного или более нуклеотидов (insertion)

  Частный случай – дупликация (duplication)

  Удаление одного или нескольких соседних нуклеотидов (deletion)

  Поворот участка длиной минимум 2 нуклеотида на 180 градусов – инверсия (inversion)

- В кодирующих участках вставки и делеции могут приводить к сдвигу рамки считывания => изменение всей аминокислотной последовательности полипептидной цепи => потеря функции белка => нежизнеспособность потомства!

Нуклеотидные замены Транзиции

- Замена пурина на другой пурин:
  - A □ G или G □ A

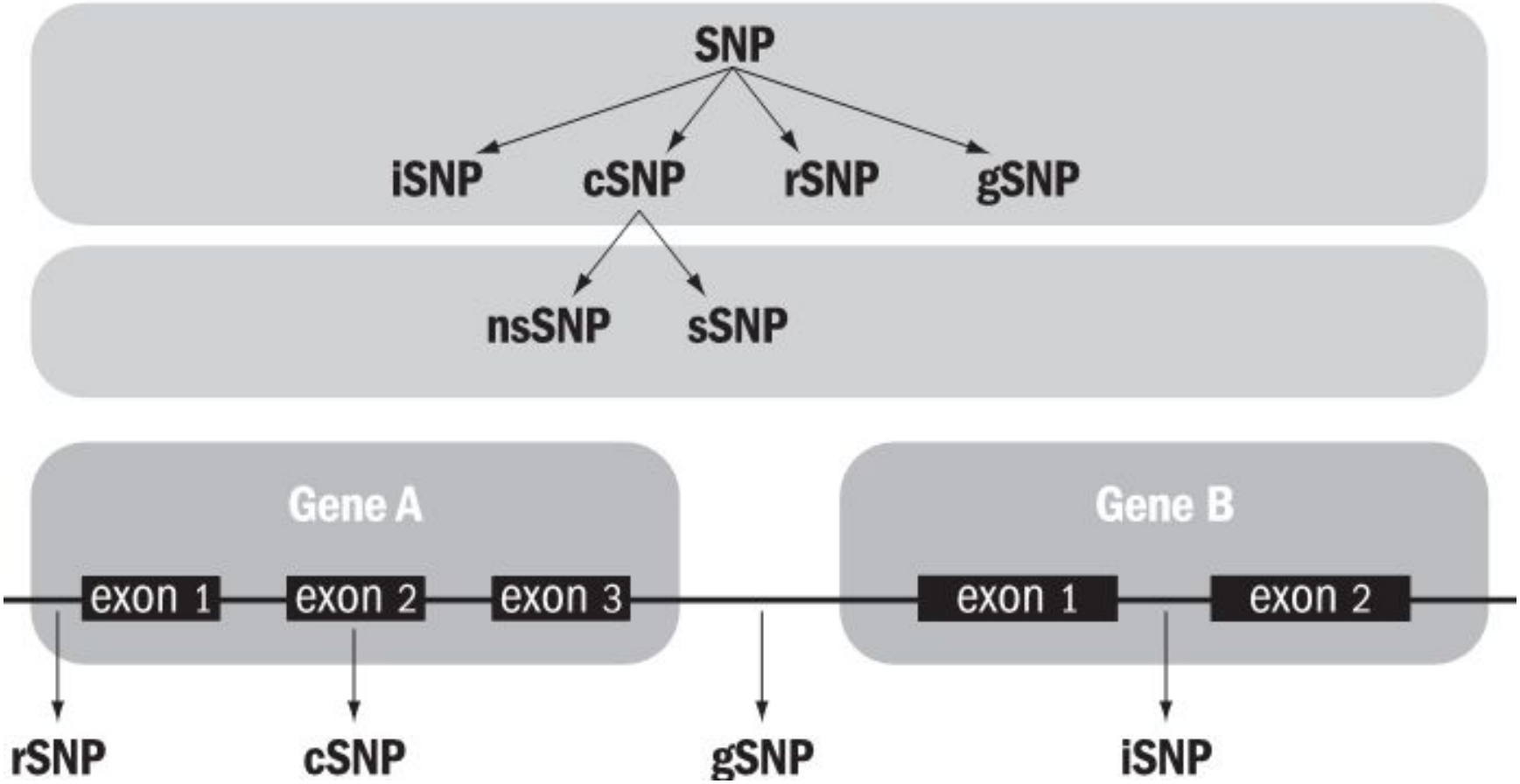
- Замена пиримидина на другой пиримидин:
  - C □ T или T □ C

Нуклеотидные замены Трансверсии

- Замена между пуринами и пиримидинами:
  - A □ T
  - A □ C
  - G □ T
  - G □ C
  - T □ A
  - T □ G
  - C □ A
  - C □ G

SNP

- Однонуклеотидный полиморфизм (англ. Single nucleotide polymorphism, SNP) — любая единичная замена основания (A, T, G или C) в геноме (или в другой сравниваемой последовательности) представителей одного вида или между гомологичными участками гомологичных хромосом индивида.

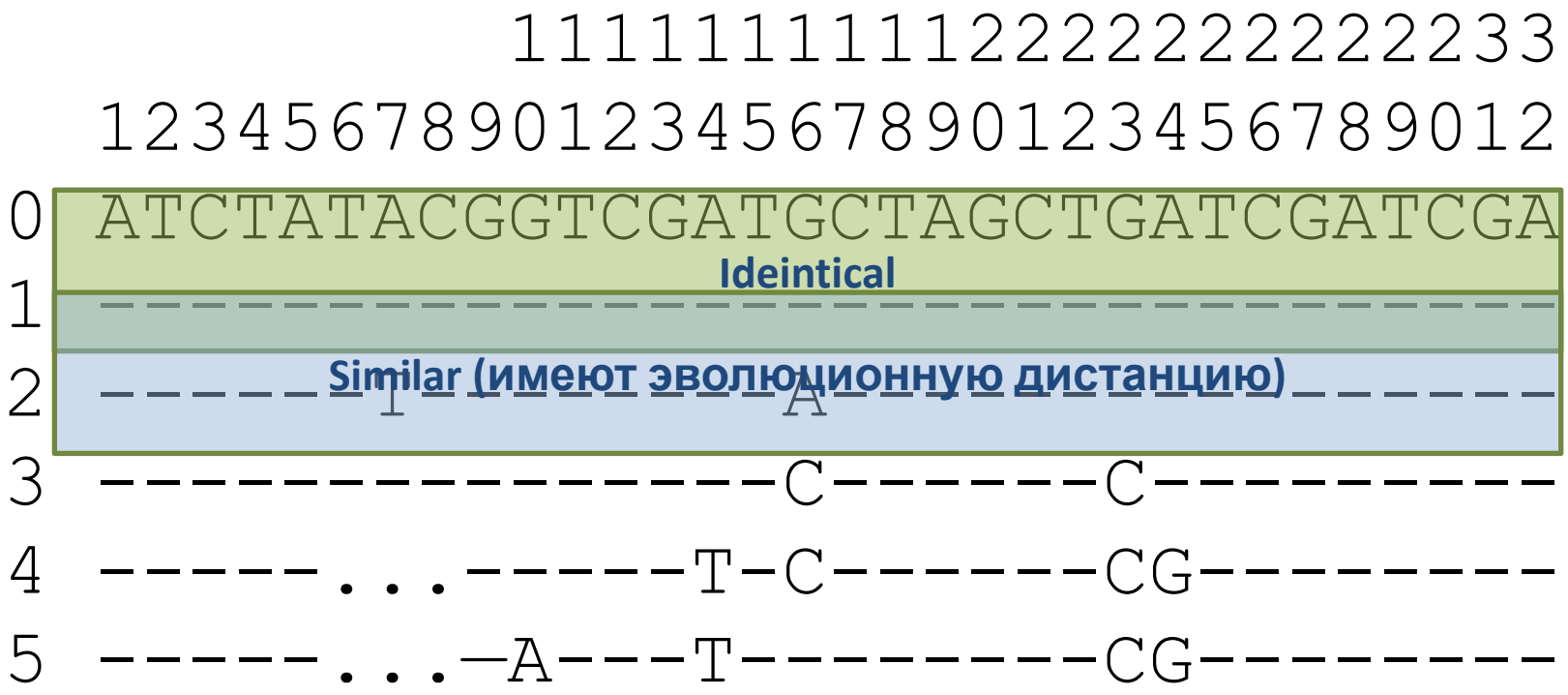

SNP
iSNP
cSNP
rSNP
gSNP
nsSNP
sSNP
Gene A
exon 1
exon 2
exon 3
Gene B
exon 1
exon 2
rSNP
cSNP
gSNP
iSNP

Нонсенс и миссенс мутации

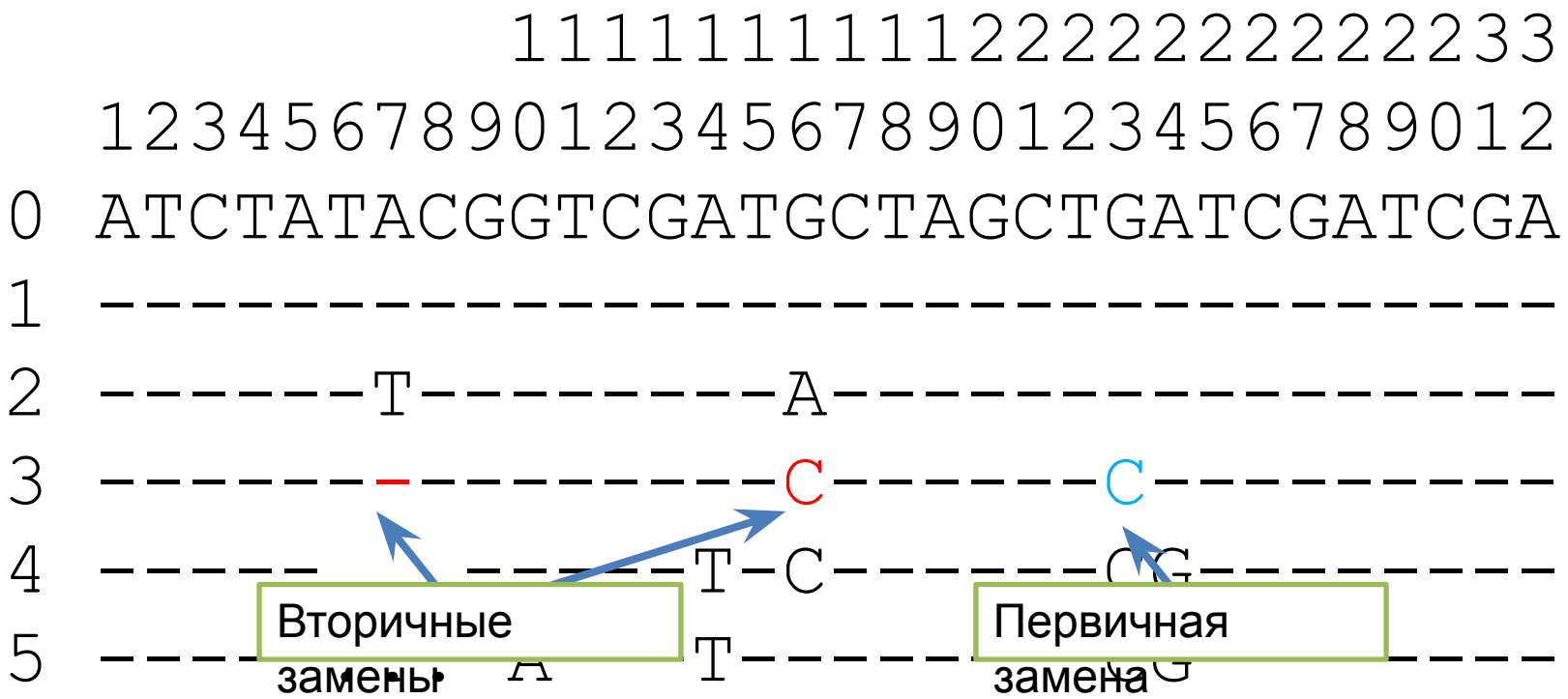
- Замена кодирующего триплета на стоп-кодон – **нонсенс мутация** (nonsence mutation)

- Замена кодирующего триплета на другой кодирующий (кроме стоп-кодона) – **миссенс мутация** (missence mutation)

Эволюция нуклеотидной последовательности



Эволюция нуклеотидной последовательности



Эволюция нуклеотидной последовательности

Идентичные

Похожие

Находятся на определенной эволюционной дистанции

Первичные и вторичные замены

Обратная замена или реверсия

Изменчивые (вариабельные) позиции

Консервативные позиции (инвариантные и относительно консервативные)

Эволюция нуклеотидной последовательности

Дивергенция – разделение на независимые родственные эволюционные линии

Параллельные мутации – мутации, произошедшие в родственных эволюционных линиях независимо друг от друга

Диверсификация – процесс, ведущий к образованию гетерогенной группы

Эволюция нуклеотидной последовательности

Гомологичные последовательности – последовательности, имеющие общее эволюционное происхождение

Группа гомологичных последовательностей, произошедших от общего предка называется **монофилетической группой**.

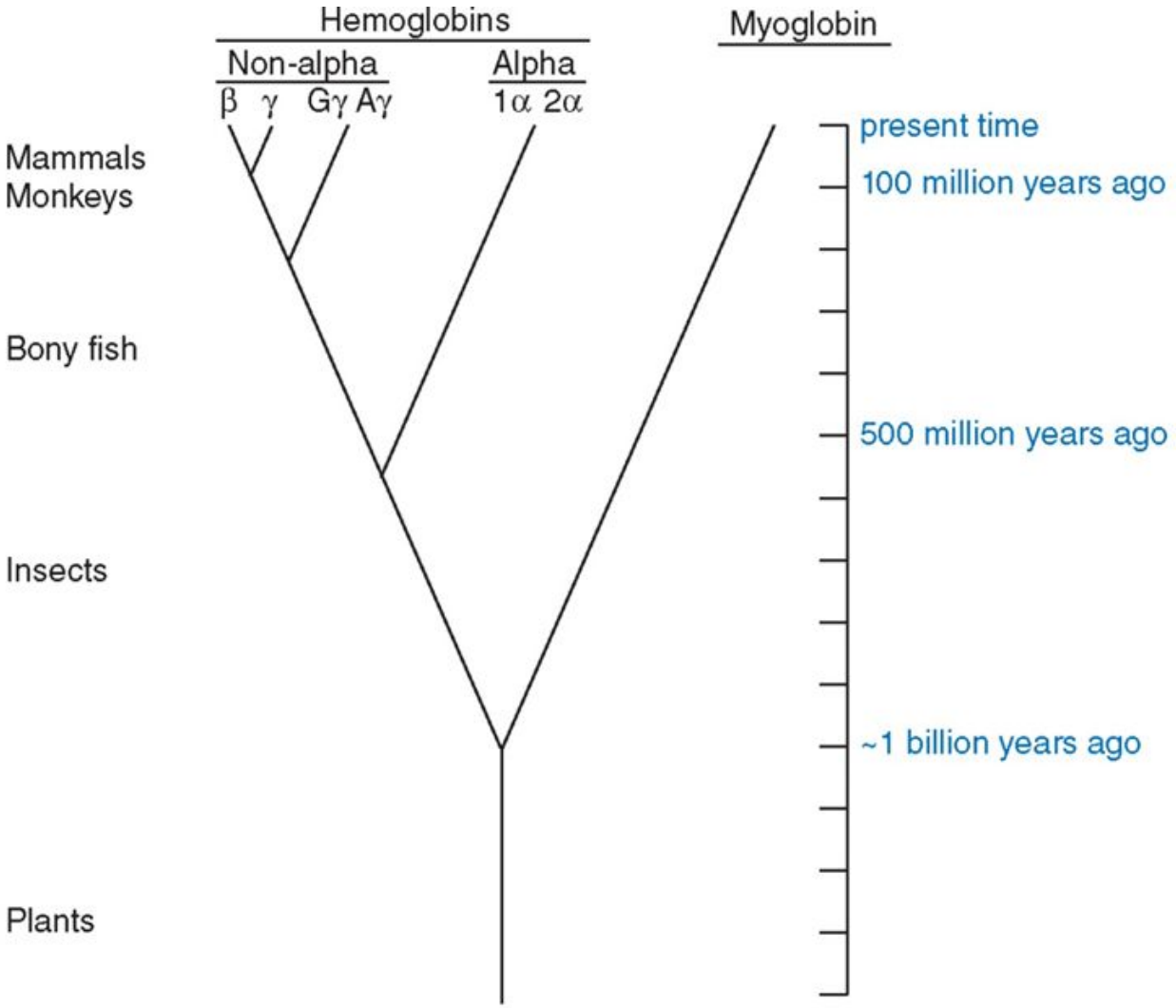


FIGURE 7.2 Dayhoff et al. (1972) summarized the relationship of the globin subfamilies in the context of evolutionary time. The dates of speciation events were inferred from fossil-based studies.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
Companion Website: www.wiley.com/go/pevsnerbioinformatics

Эволюция нуклеотидной последовательности

- **MRCA** – most recent common ancestor (наиболее недавний общий предок)

- **LUCA** – last universal common ancestor (Последний общий предок всех форм жизни на Земле)

Бритва Оккама:

«без необходимости не стоит утверждать многого»

То, что можно объяснить посредством меньшего, не следует выражать посредством большего

Консенсусные последовательности

- Искусственная последовательность, содержащая в каждой позиции нуклеотид, встречаемый наиболее часто у анализируемых последовательностей
- Обычно, присутствующий минимум в 50% анализируемых последовательностей (строгость пороговых критериев!)

Консенсусные последовательности

- В консенсусной последовательности можно отображать разнообразие нуклеотидов в конкретной позиции (гетерогенность позиции)
- Если в некоторой позиции у одних последовательностей обнаружена A, а у других – T, то A/T-гетерогенность можно отобразить как W (IUPAC)

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

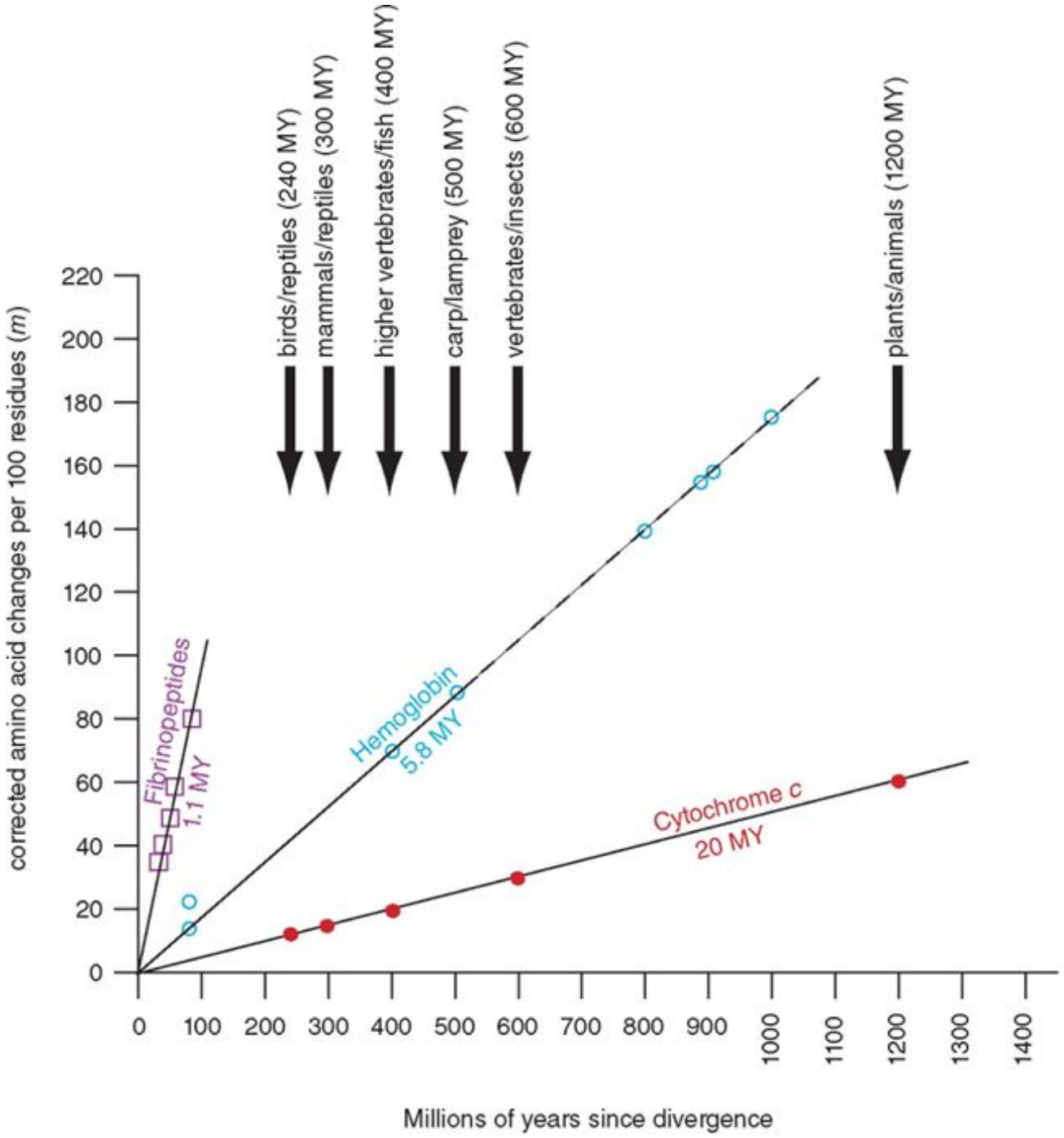
IUPAC amino acid code	Three letter code	Amino acid
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

Концепция молекулярных часов

- Закрепление мутаций в популяции занимает определённое время и постоянна
- Цукеркандль и Поллинг сформулировали концепцию молекулярных часов в эволюции живых систем:
  - Для конкретной генетической последовательности скорость эволюции постоянна во времени и одинакова у всех дочерних последовательностей

Концепция молекулярных часов

- Если известно, что дивергенция двух, различающихся между собой на один нуклеотид последовательностей произошла 10 лет тому назад, то дивергенция между этими двумя последовательностями и третьей, имеющей 2 отличия от них, произошла 20 лет назад (с учетом доверительного интервала)


birds/reptiles (240 MY)
mammals/reptiles (300 MY)
higher vertebrates/fish (400 MY)
carp/lamprey (500 MY)
vertebrates/insects (600 MY)
plants/animals (1200 MY)
corrected amino acid changes per 100 residues (m)
220
200
180
160
140
120
100
80
60
40
20
0
Fibrinopeptides
1.1 MY
Hemoglobin
5.8 MY
Cytochrome c
20 MY
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
Millions of years since divergence

Влияние отбора на мутации

- Мутации, улучшающие приспособленность организма, подвергаются действию положительного естественного отбора — эволюционным силам, факторам, направленным на отбор такой мутации, ее преимущественное закрепление в популяции.
- Мутации, ухудшающие приспособленность организма, подвергаются действию отрицательного естественного отбора — эволюционным силам, факторам, направленным на преимущественное удаление такой мутации из популяции.
- Мутации, не изменяющие приспособленности организма к окружающей среде, называют *селективно нейтральными*.
- Процесс изменения частоты мутации в популяции под действием стохастических процессов называют *случайным генетическим дрейфом* (random genetic drift).
- Для кодирующих нуклеотидных последовательностей основными факторами естественного отбора являются эволюционные факторы, действующие на уровне белка. Соответственно действию этих факторов подвергаются несинонимичные, изменяющие белок, а не синонимичные замены.
- В целом синонимичные замены принято считать селективно нейтральными, или близкими к нейтральным.

Измерение отбора путем анализа последовательностей белок-кодирующих генов (Hurst, 2002; Li, 1997)

Отношение *Ka*/*Ks* (где *Ka* – частота несинонимичных замен, *Ks* – синонимичных; обе вычисляются с коррекцией на множественные замены) является количественной мерой отбора, действующего на уровне белковых последовательностей.

http://www.rumvi.com/products/ebook/логика-случая-о-природе-и-происхождении-биологической-эволюции/c4e78da6-26bb-491e-b262-aa73b9043637/preview/preview.html

- $Ka/Ks = 1$ – нейтральная эволюция белковой последовательности (кодируемый белок не подвергается отбору).
- Для большинства белок-кодирующих генов $Ka/Ks \ll 1$ – отсекающий отбор.
- Для прокариот типично $Ka/Ks < 0{,}1$.
- Для эукариот типично $Ka/Ks \approx 0{,}1–0{,}2$.
- $Ka/Ks > 1$ – положительный отбор; достаточно редко встречается для белок-кодирующих генов, но для некоторых категорий генов, несомненно, присутствует, например, для генов, участвующих в антипаразитической защите или в сперматогенезе, а также в вирусных белках, таких как гемагглютинин вируса гриппа.
- Для измерения Ka/Ks для индивидуальных сайтов используют методы максимального правдоподобия; большинство белок-кодирующих генов содержат несколько сайтов, подверженных положительному отбору.

- Использование *Ka*/*Ks* для измерения уровня отбора предполагает нейтральность синонимичных сайтов.
- Однако *Ka* и *Ks* положительно коррелируют между собой – таким образом, отбор затрагивает и синонимичные сайты.
- Некодирующие сайты, такие как интронные последовательности, могут использоваться как фон нейтральной эволюции при измерении отбора на синонимичных сайтах (*Ks*/*Ki*, где *Ki* – частота замен для интронных сайтов).

Критерий Макдональда – Крейтмана (Aquadro, 1997; McDonald and Kreitman, 1991) широко используется для измерения отбора. Он сравнивает внутривидовые вариации (частота полиморфизма, *P*) с межвидовыми вариациями (дивергенция, *D*).

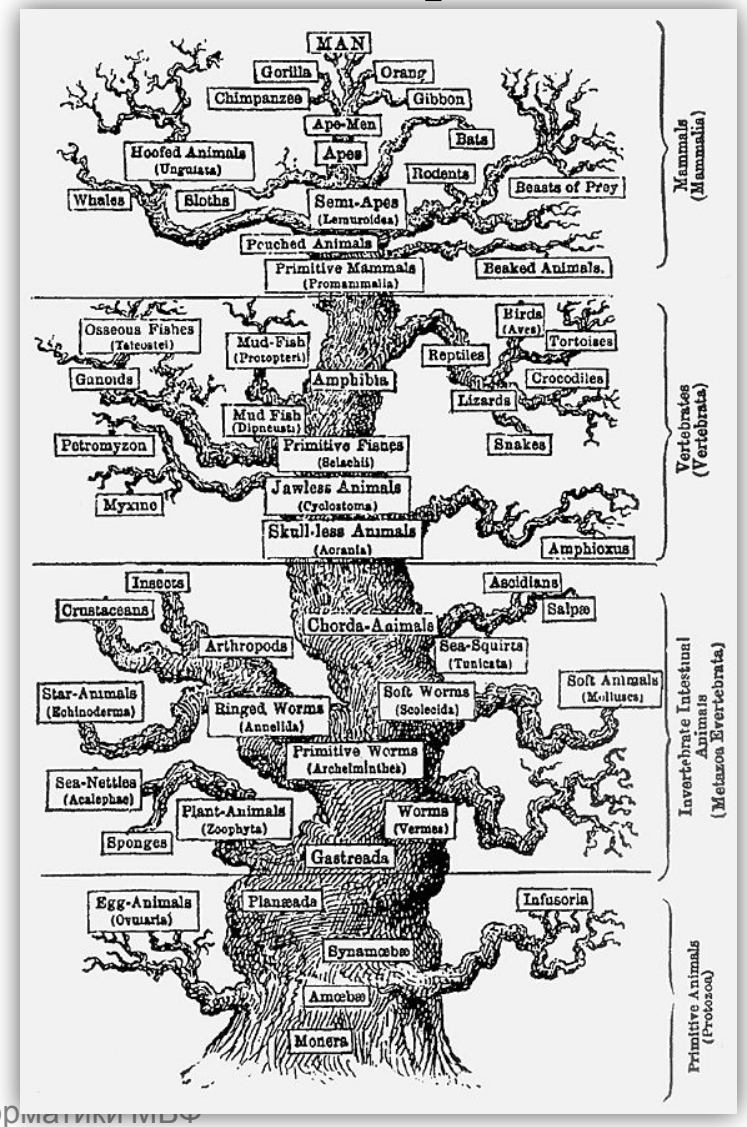
- *Dn*/*Ds* = *Pn*/*Ps* – нейтральная эволюция белковой последовательности.
- *Dn*/*Ds* < *Pn*/*Ps* – отсекающий отбор.
- *Dn*/*Ds* > *Pn*/*Ps* – положительный отбор.

Dn - скорость эволюции по несинонимичным сайтам; *Ds* - по синонимичным

Филогенетические деревья

The time will come, I believe, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of Nature.

Charles Darwin



Причина подобия – общее происхождение!

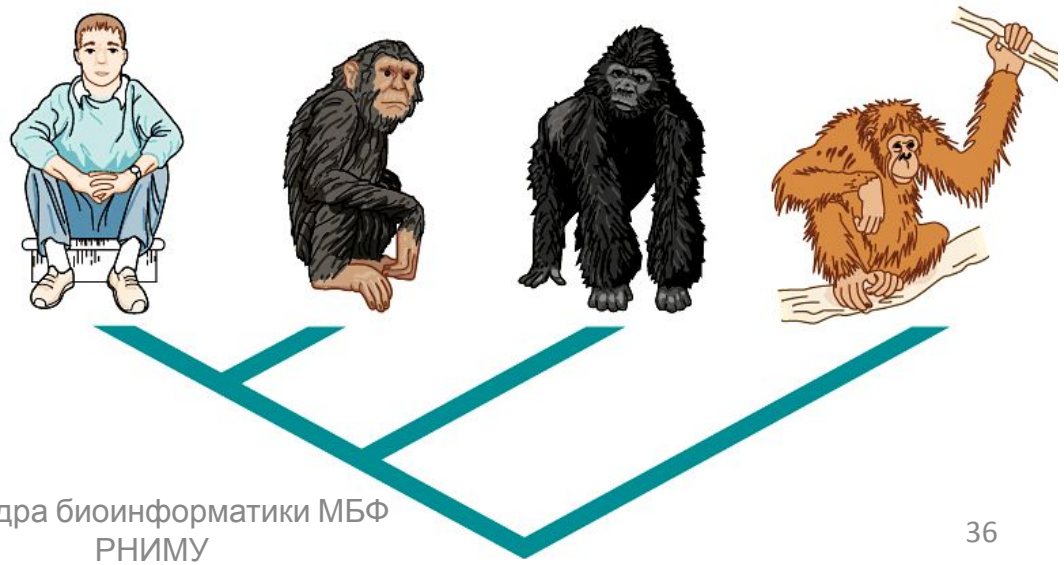
- **Гомология** – происхождение от общего предка
- **Подобие** – наблюдаемые данные, собранные сейчас не подразумевающие каких-либо исторических гипотез
- **Кластеризация** – объединение вместе сходных по определенным признакам объектов (**Иерархическая классификация**)
- **Филогения** – описание биологических взаимосвязей, обычно в виде филогенетического дерева

Зачем нужны филогенетические деревья?

<u>Биологические задачи</u>:

сравнение 3-х и более объектов
(кто на кого более похож ....)

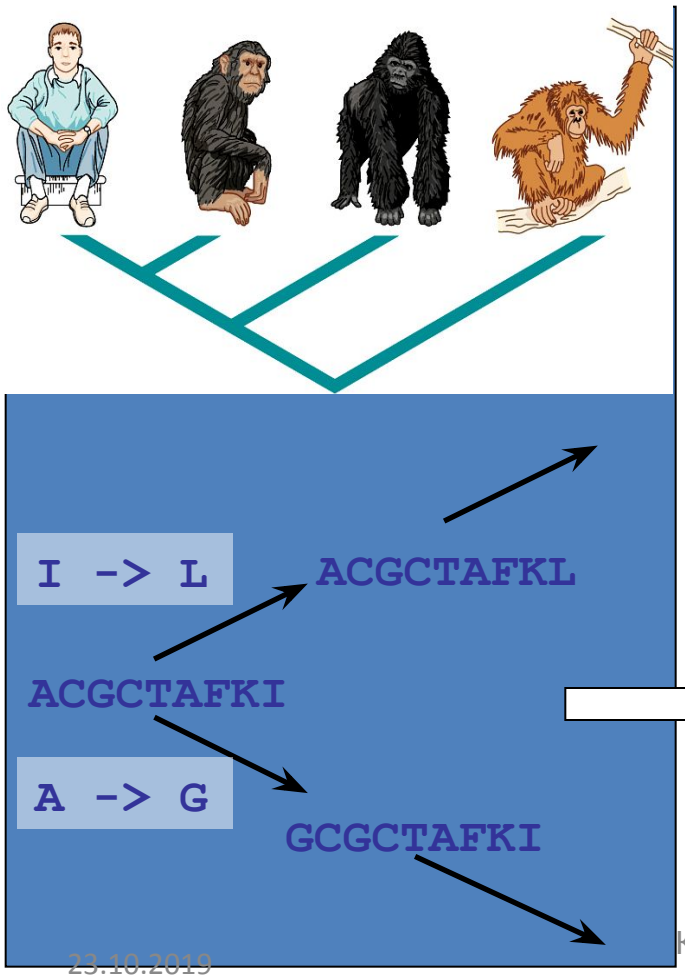
реконструкция эволюции
(кто от кого, как и когда произошел…)



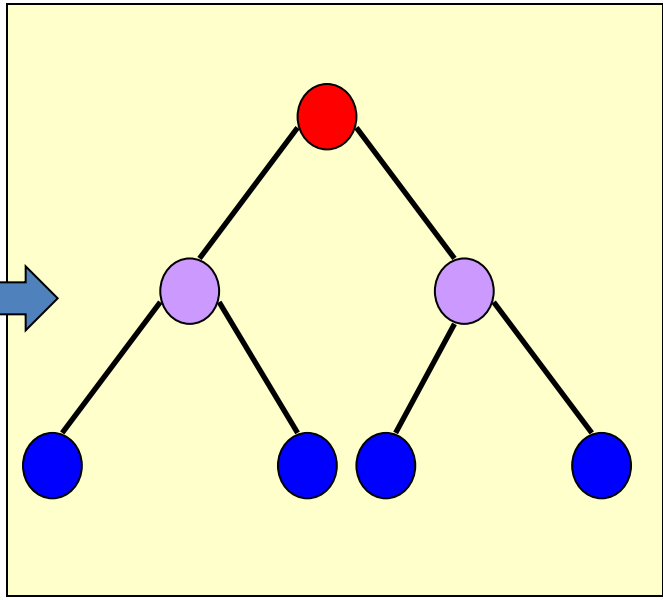
Реальные события : эволюция в природе или в лаборатории, компьютерная симуляция

Данные: например, а.к. последовательности или количество щетинок

Построенное дерево древовидный граф, ***вычисленный на основе данных***, может отражать или не отражать реальные события




```
>Seq1
ASGCTAFKL
. . .
>Seq3
GCGCTLFKI
>Seq4
GCGCTGFKI
. . . . .
```



Узел (node) — точка разделения предковой последовательности (вида, популяции) на две независимо эволюционирующие. Соответствует внутренней вершине графа, изображающего эволюцию.

Лист (**leaf, OTU** – оперативная таксономическая единица) — реальный (современный) объект; внешняя вершина графа.

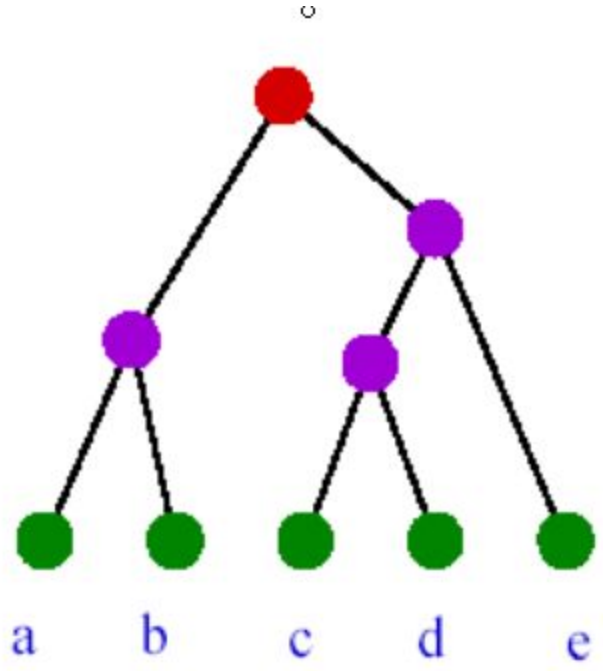
Ветвь (branch) — связь между узлами или между узлом и листом; ребро графа.

Корень (root) — гипотетический общий предок.

Клада (clade) - группа двух или более таксонов или последовательностей ДНК, которая включает как своего общего предка, так и всех его потомков.

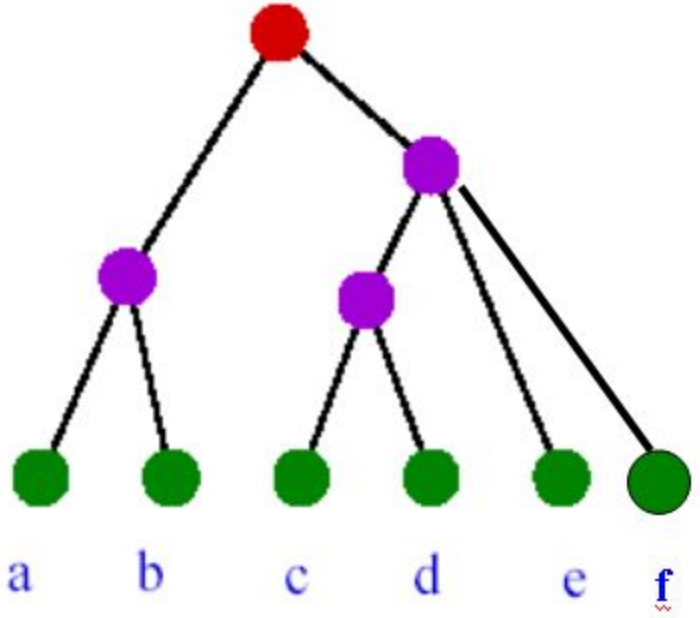
Бинарное (разрешённое)

(в один момент времени может произойти только одно событие)



Небинарное (неразрешённое)

(может ли в один момент времени произойти два события?)

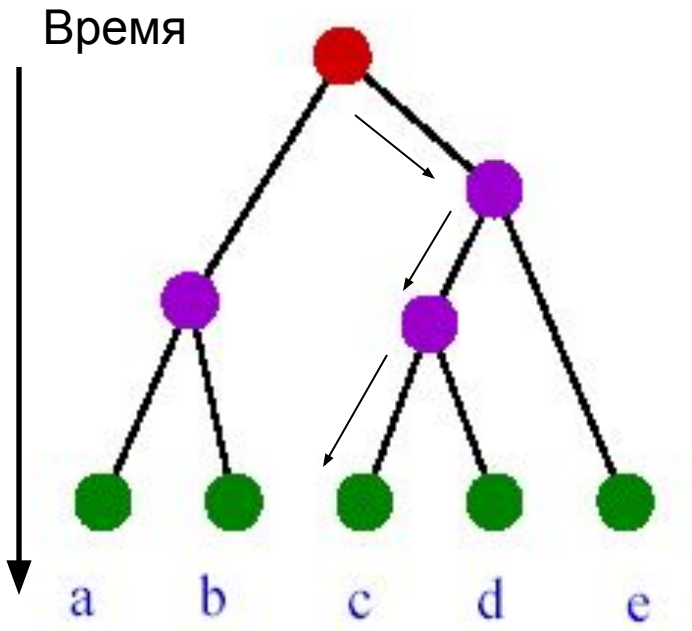


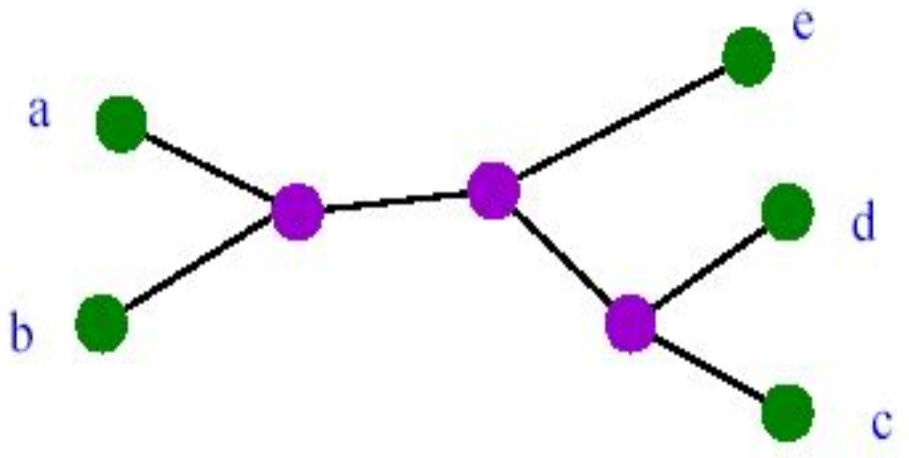
Время

Какие бывают деревья?

Укорененное дерево (rooted tree)
отражает направление эволюции

Неукорененное (бескорневое) дерево (unrooted tree) показывает только связи между узлами





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

Существует (2n-5)!! разных бескорневых деревьев с n листьями

Рутинная процедура, или как строят деревья?

Составление выборки последовательностей

Множественное выравнивание

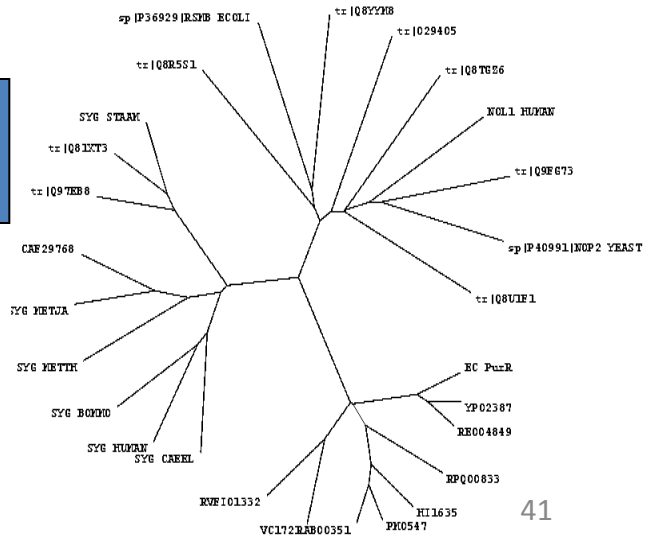
```
f53969          ALLKTPAEFDAYELNSSIKGAGTDEACLIEILSSRSNAEIKEINRIYKQEYKKTLEDAIK 313
con101          AMLKTPSQYDAYELKEAIKGAGTDEACLIEILASRSNAEIREINQVFKAENKKSLEDAIS 347
hum4            GMMTPTVLYDVQELRRAMKGAGTDEGCLIEILASRTPEEIRRISQTYQQQYGRSLEDDIR 140
ANX4_MOUSE      GLMTPTVLYDVQELRRAMKGAGTDEGCLIEILASRTPEEIRRINQTYQQQYGRSLEKDIC 140
max1            GLLMPAPVYDAYELKAAMKGAGTEEACLIDILASRSNSEMNAINEVYKKEYGKTLEDAVC 143
```

Построение дерева

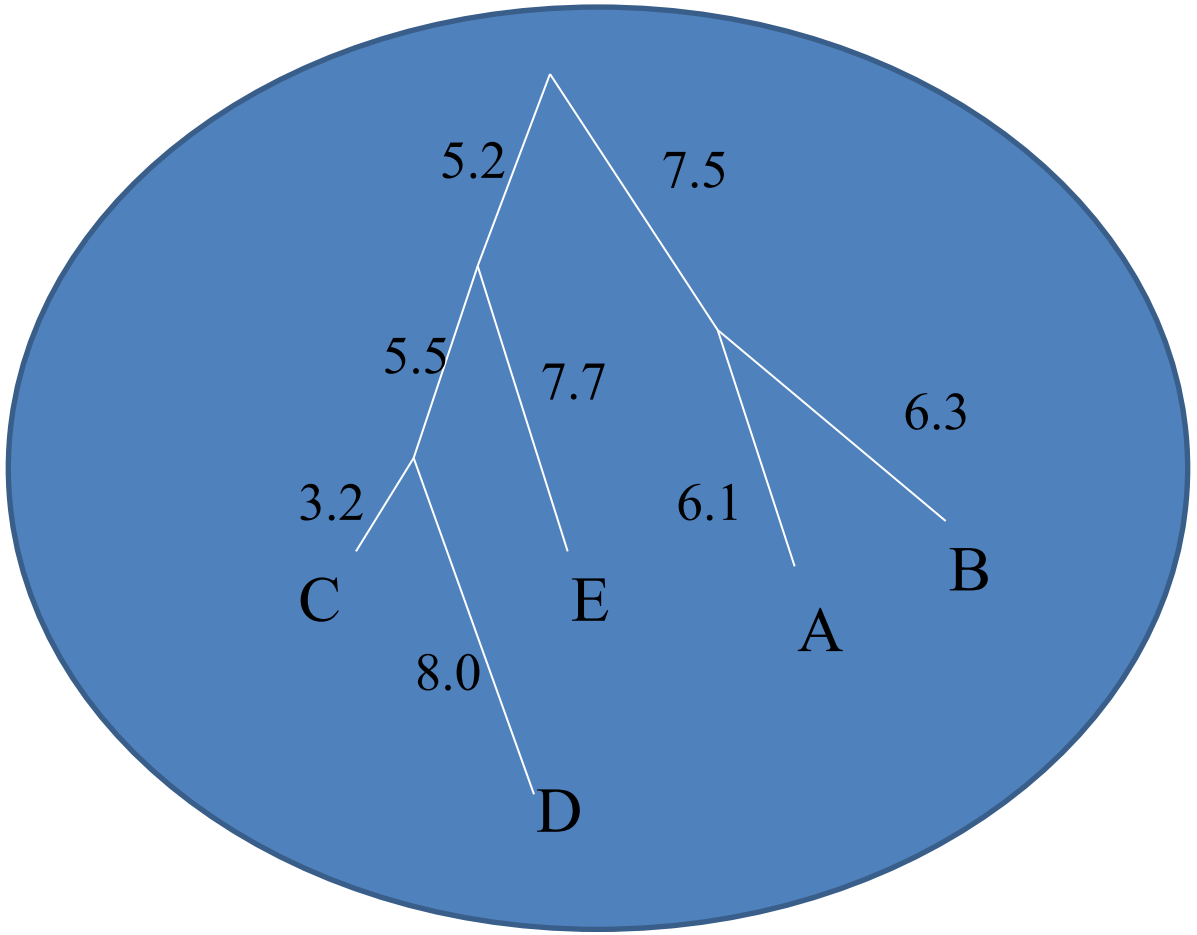
фрагмент записи в виде скобочной формулы:

(((((con101:38.51018,(f53969:28.26973,((f67220:8.39851, max4:27.50591):4.92893,con92:30.19677):13.62315):9.53075):25.83145,

Визуализация и редактура дерева



Скобочная формула (Newick format)



`(((C,D),E)),(A,B));` только топология

`(((C:3.2,D:8.0):5.5,E:7.7):5.2,(A:6.1,B:6.3):7.5);` длины ветвей

Как выбирать последовательности для дерева?

- ✔ Кроме случаев очень близких последовательностей, проще работать с белками (а не с ДНК)
- ✔ Придерживайтесь небольшой выборки (< 50 последовательностей)
- ✔ Избегайте:
  - фрагментов;
  - ксенологов;
  - рекомбинантных последовательностей;
  - многодоменных белков и повторов
- ✔ Используйте outgroup (последовательность, ответвившаяся от общего предка заведомо (но минимально!) раньше разделения интересующих групп-клад)

Самое главное – хорошее выравнивание!

- ✔ Максимальный вклад в финальное дерево: нельзя построить хорошее дерево по плохому выравниванию
- ✔ Блоки, содержащие много гэпов, плохо выровненные N- и C- концы можно просто вырезать.

Основные алгоритмы построения филогенетических деревьев

Методы, основанные на оценке расстояний (матричные методы):

Вычисляются эволюционные расстояния между всеми листьями (OTUs) и строится дерево, в котором расстояния между вершинами наилучшим образом соответствуют матрице попарных расстояний.

- **UPGMA**
- **Neighbor-joining**
- Минимальная эволюция
- Квартеты («топологический»)
- ...

Максимального правдоподобия, Maximal likelihood, ML

Используется модель эволюции и строится дерево, которое наиболее правдоподобно при данной модели

Максимальной экономии (бережливости), maximal parsimony, MP

Выбирается дерево с минимальным количеством мутаций, необходимых для объяснения данных

Пример матрицы расстояний

```
1       2       3       4       5       6       7       8
00  10.53    9.77  12.78  12.03  16.54  13.53  25.00     HUMAN 1
  0.00    9.02  12.03   9.77  15.79   9.02  27.27     HORSE 2
      0.00   9.77   9.02  16.54  12.03  24.24     RABIT 3
          0.00   2.26  17.29  10.53  25.76     MOUSE 4
              0.00  15.79   8.27  25.76     RAT 5
                  0.00  10.53  29.55     BOVIN 6
                      0.00  25.00     PIG 7
                          0.00     CHICK 8
```

Расстояние (уровень дивергенции) между соответствующими последовательностями из геномов мыши и свиньи

Как понимать расстояние между объектами?

- Как время, в течение которого они эволюционировали
- Как число «эволюционных событий» (мутаций)

В первом случае объекты образуют
ультраметрическое пространство
(если все объекты наблюдаются в одно время, что, как правило, верно)

Но время непосредственно измерить невозможно

Гипотеза «молекулярных часов»

(E.Zuckerkandl, L.Pauling, 1962)

За равное время во всех ветвях эволюции накапливается равное число мутаций

Если гипотеза молекулярных часов принимается, число различий между выровненными последовательностями можно считать примерно пропорциональным времени. Отклонения от ультраметричности можно считать случайными. Эволюция реконструируется в виде **ультраметрического** дерева.

Укоренённое дерево называется ультраметрическим, если расстояние от корня до любого из листьев одинаково.

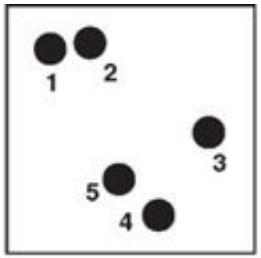
UPGMA

Unweighted Pair Group Method with Arithmetic Mean

разновидность кластерного метода

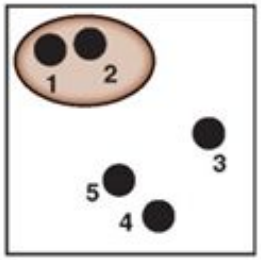
Расстояние между кластерами вычисляется как среднее арифметическое всевозможных расстояний между последовательностями из кластеров

(a)

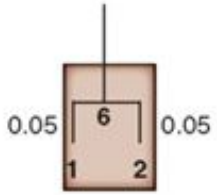


	1	2	3	4	5
1	—				
2	0.1	—			
3	0.8	0.8	—		
4	0.8	1	0.3	—	
5	0.9	0.9	0.3	0.2	—

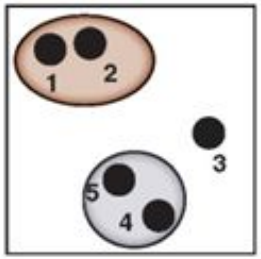
(b)



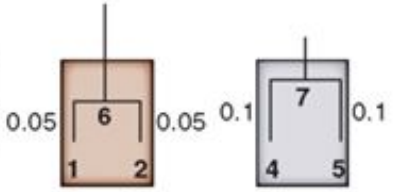
	(1,2)	3	4	5
(1,2)	—			
3	0.8	—		
4	0.9	0.3	—	
5	0.9	0.3	0.2	—



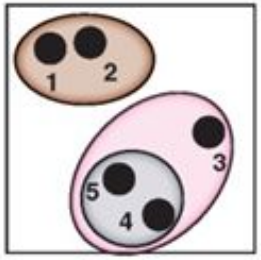
(c)



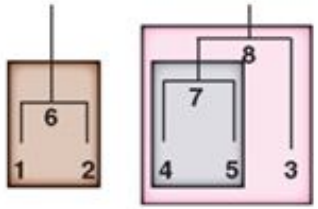
	(1,2)	3	(4,5)
(1,2)	—		
3	0.8	—	
(4,5)	0.9	0.3	—



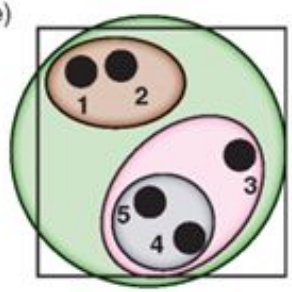
(d)

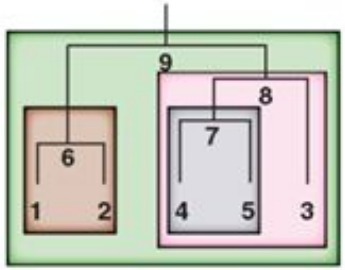


	(1,2)	[3,(4,5)]
(1,2)	—	
[3,(4,5)]	0.85	—

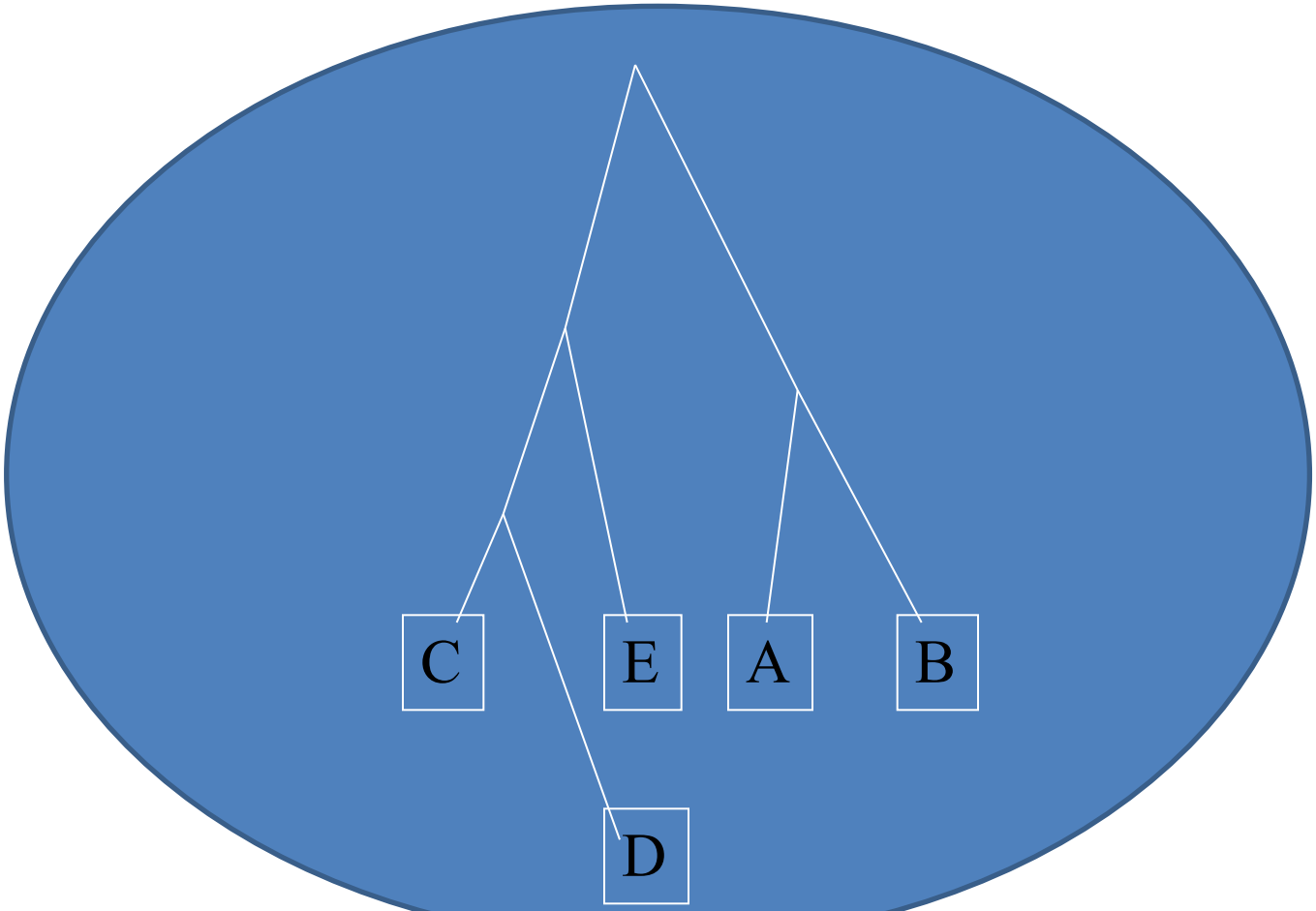


(e)





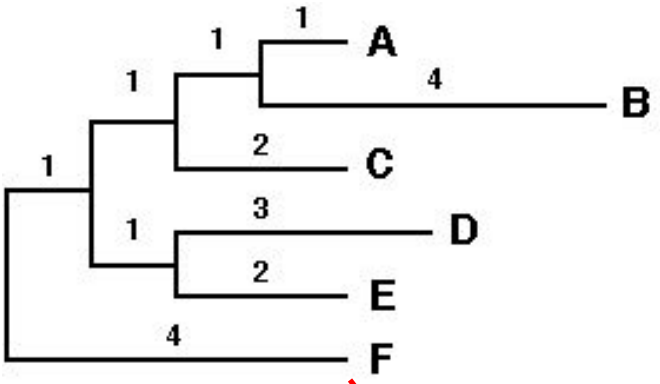
Гипотеза молекулярных часов не всегда справедлива



(длина ветвей пропорциональна числу мутаций)

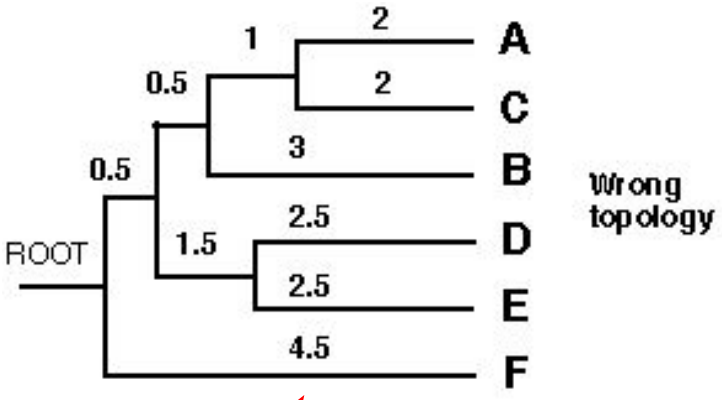
Недостатки UPGMA

Алгоритм строит ультраметрическое дерево, а это означает, что скорость эволюции предполагается одинаковой для всех ветвей дерева. Использовать этот алгоритм имеет смысл только в случае ультраметрических данных (справедливости «молекулярных часов»).



Реальное дерево

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

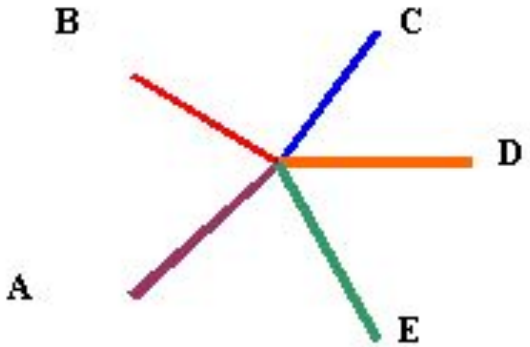


UPGMA

Метод ближайших соседей (Neighbor-joining, NJ)

- ✔ Строит неукоренённое дерево
- ✔ Может работать с большим количеством данных
- ✔ Достаточно быстрый
- ✔ Хорошо зарекомендовал себя на практике: если есть недвусмысленное с точки зрения эксперта дерево, то оно будет построено.
- ✔ Могут появиться ветви с длиной <0

Метод Neighbor-joining



Рисуем «звездное» дерево и будем «отщипывать» от него по паре листьев

Пусть $u_i = \Sigma_k \ M_{ik}/(n-2)$ — среднее расстояние от листа *i* до других листьев

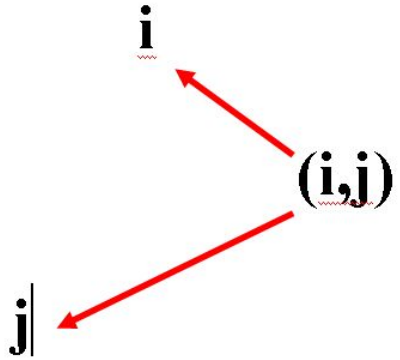
1. Рассмотрим все возможные пары листьев. Выберем 2 листа *i* и *j* с минимальным значением величины

$$M_{ij} - u_i - u_j$$

т.е. выбираем 2 узла, которые близки друг к другу, но далеки ото всех остальных.

Метод ближайших соседей (Neighbor-joining, NJ)

2. Кластер (i, j) – новый узел дерева



Расстояние от i или от j до узла (i,j):

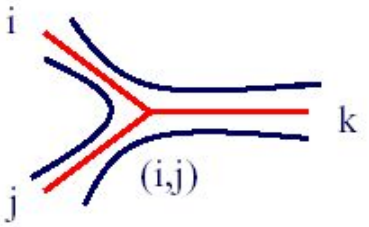
$D(i, (i,j)) = 0{,}5 \cdot (Mij + ui - uj)$

$D(j, (i,j)) = 0{,}5 \cdot (Mij + uj - ui)$

т.е. длина ветви зависит от среднего расстояния до других вершин

3. Вычисляем расстояние от нового кластера до всех других

$$M(ij)k = \frac{Mik + Mjk - Mij}{2}$$



5. В матрице M убираем i и j и добавляем (i, j).

Повторяем, пока не останутся 3 узла ...

Стандартная ситуация

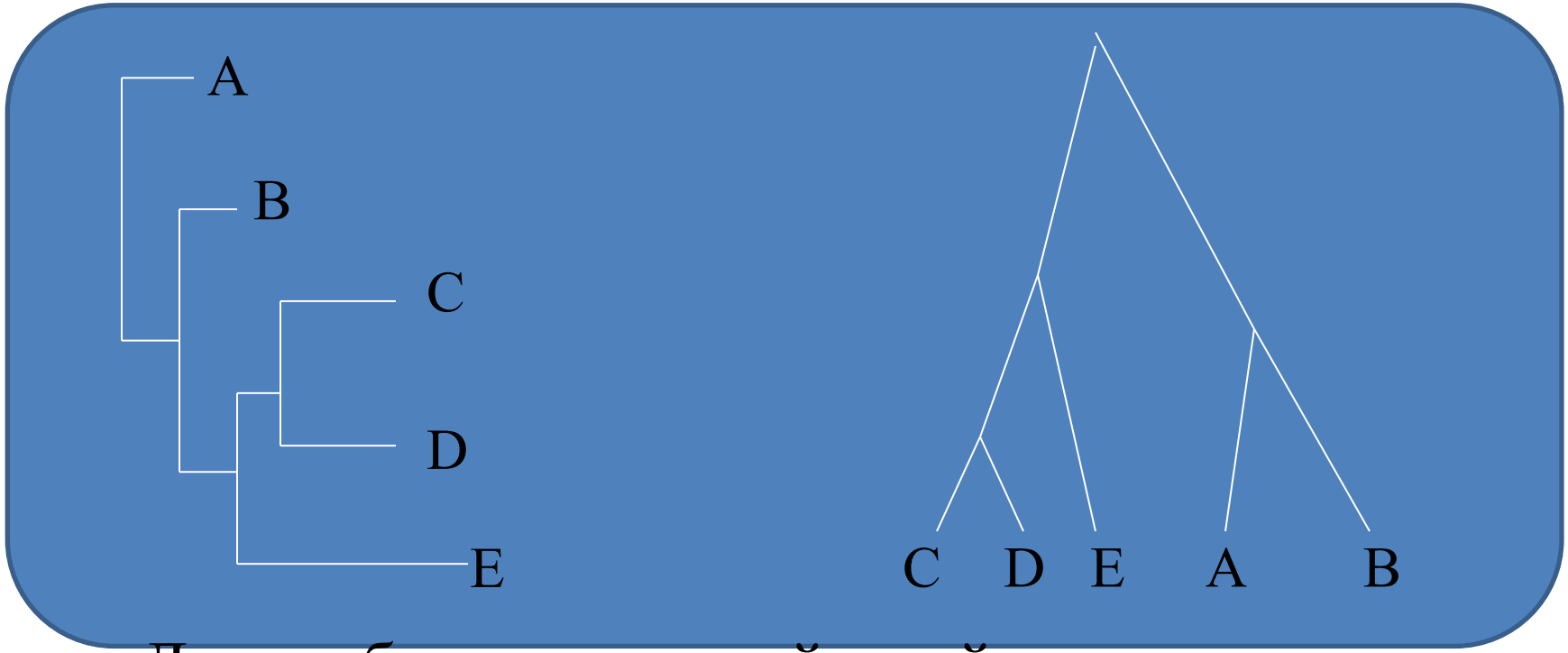
- ✔ Понимаем расстояние как число мутаций
- ✔ Реальное (неизвестное нам) дерево — укоренённое, но не ультраметрическое
- ✔ Мы реконструируем неукоренённое дерево (топологию и длины ветвей). Его надо понимать как **множество** всех возможных укоренений.

Если данные таковы, что гипотеза молекулярных часов не проходит, то реконструкция укорененного дерева намного менее надёжна, чем реконструкция неукоренённого

Как изобразить дерево? Топология дерева

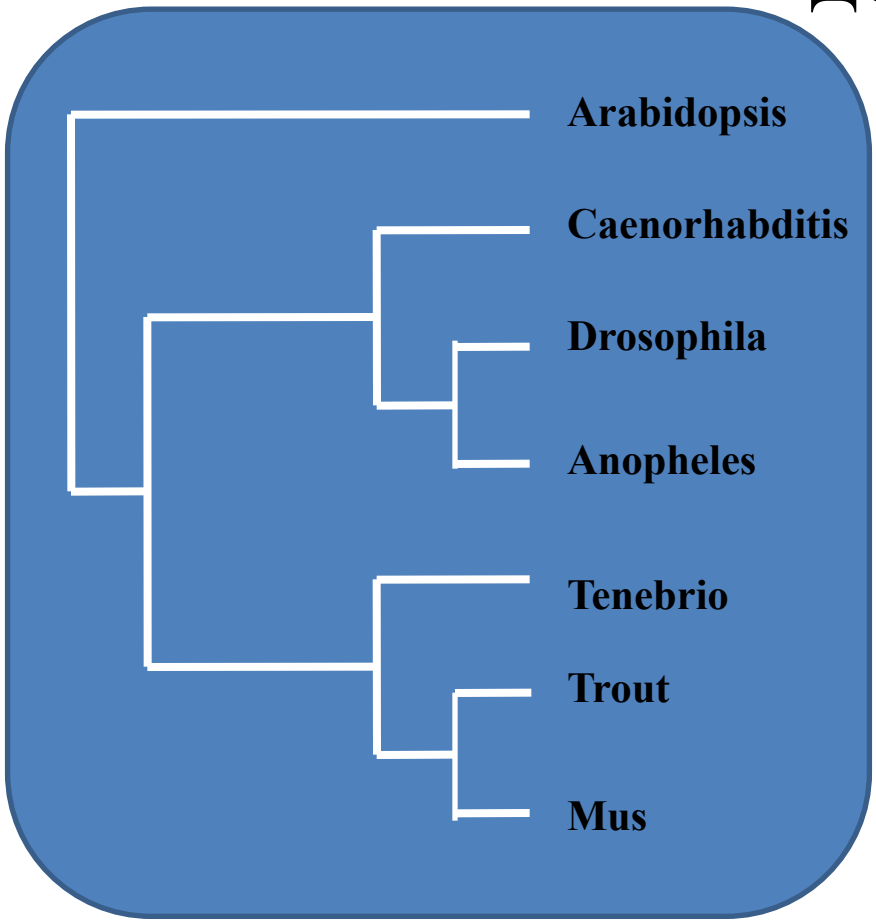
Топология дерева — только листья, узлы, (корень) и связывающие их ветви

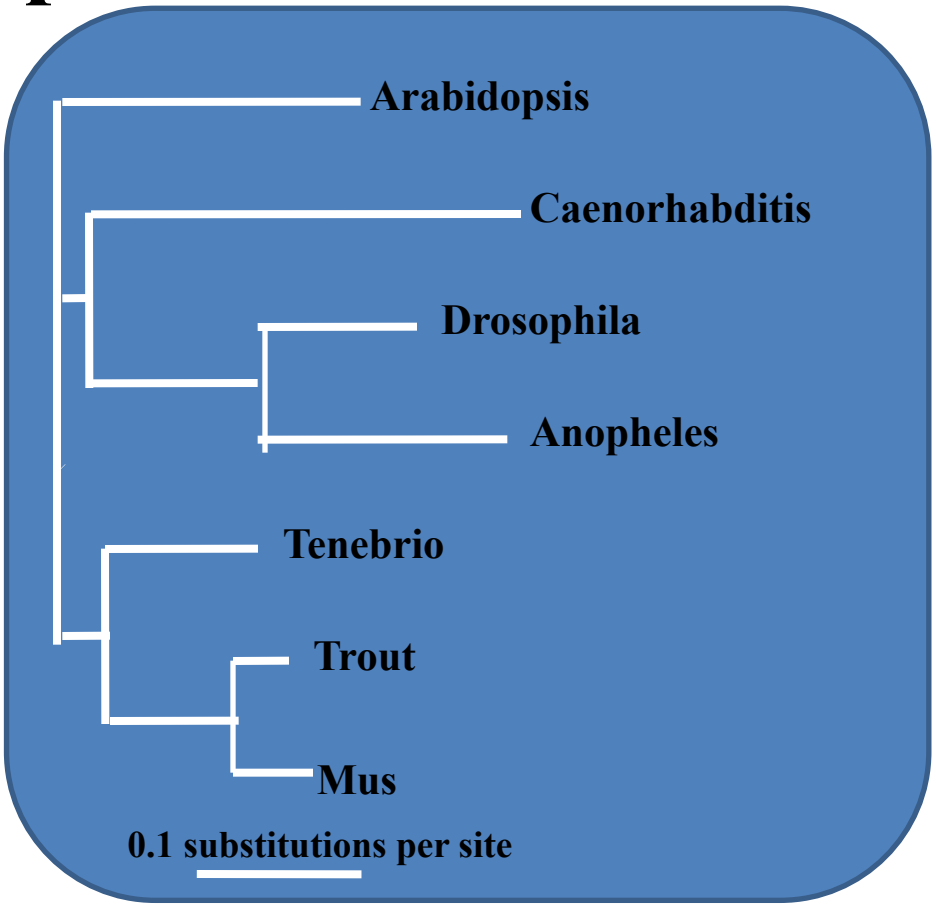
(топология не зависит от способа изображения дерева)



Два изображения одной и той же топологии

Как можно нарисовать построенное дерево?



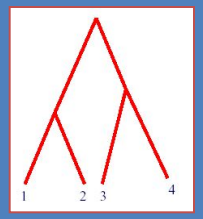


Кладограмма:
представлена только топология, длина ребер игнорируется.

Филограмма:
Длина ребер пропорциональна эволюционному расстоянию между узлами.

Достоверность топологии. Bootstraps

Есть множественное выравнивание и построенное по нему дерево.
Верим ли мы в топологию дерева?

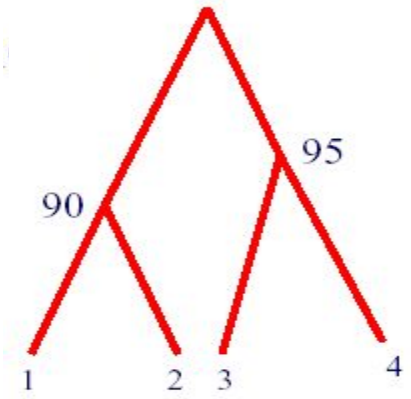


- Создадим псевдоданные:

  N множественных выравниваний той же длины, что и исходное, каждое из псевдовыравниваний - случайный набор столбцов из исходного (выборка с возвращением!)

- Построим N деревьев:

  на каждой внутренней ветви отметим долю случаев из N, в которых появлялся этот узел.



Обычно верят в топологию, если метки ветвей на бутстрепном дереве больше 70-80% . Если меньше 50%, то не верим. В иных случаях – думаем…

Какие on-line программы строят деревья?

- ✔ ClustalW. “Tree type” – nj, phylip: строит только методом NJ, но результат – в разных форматах, no bootstraps
- ✔ Phylip (Felsenstein, 1993) – пакет программ для построения филогенетических деревьев (stand-alone)

  On-line (partly): например,

  http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html
- ✔ PAUP (Phylogenetic Analysis Using Parsimony)

Phylogeny : Phylip programs

Documentation.
FAQ (Frequently Asked Questions).

- **Programs for molecular sequence data [sequence.doc]**
  - DNA

    dnadist [advanced form] [dnadist.doc]
    Distances from DNA sequences.

    dnapars [advanced form] [dnapars.doc]
    Parsimony method for DNA.

    dnaml
    (Maximum likelihood method) has been removed ; please use rather fastDNAml, which is much faster and equivalent.
  - Proteins

    protdist [advanced form] [protdist.doc]
    Distances from protein sequences.

    protpars [advanced form] [protpars.doc]
    Parsimony method for protein sequences.

- **Programs for distance matrix data [distance.doc]**

neighbor [advanced form] [neighbor.doc]
Neighbor-joining and UPGMA methods

fitch [advanced form] [fitch.doc]
Fitch-Margoliash and least-squares methods

kitsch [advanced form] [kitsch.doc]
Fitch-Margoliash and least-squares methods with molecular clock

- **Programs for trees**

drawtree [advanced form] [drawtree.doc]
Draw a tree (as computed *before* by phylogenetic analysis methods).

drawgram [advanced form] [drawgram.doc]
Draw a phenogram (see comment for drawtree).

consense [advanced form] [consense.doc]
Consensus tree program.

unroot NEW

Пакет Phylip

- ✔ protdist — оценка эволюционных расстояний между белковыми последовательностями (вход — множественное выравнивание, выход — матрица попарных расстояний)
- ✔ dnadist — то же для нуклеотидных последовательностей
- ✔ protpars – оценка числа нуклеотидных мутаций для наблюдаемой частоты белковых замен (близкие последовательности)
- ✔ neighbor — реконструкция филогении по матрице расстояний методами NJ и UPGMA
- ✔ drawtree — рисование неукоренённого дерева
- ✔ drawgram — рисование кладограмм и филограмм

Bootstrapping with Phylip

- ✔ Надо выбрать Bootstrap options в protdist, выставить не менее 100 итераций, нечетное число в “Random number of seed”
- ✔ Затем, при запуске “Neighbor” снова выбрать “Bootstrap options” и выставить указанное в пред. пункте количество наборов данных и отметить “Compute a consensus tree”

Общий план действий с пакетом Phylip

- ✔ Множественное выравнивание -> protdist
- ✔ Bootstrap options - ?
- ✔ Результат – или сразу, или URL по e-mail (предлагают продолжить с программой построения дерева)
- ✔ Выбрать Neighbor, Neighbor-Joining, Boostrap…?, outgroup – позиция outgroup в выравнивании
- ✔ Выход: outfile.consense – текстовый рисунок
- ✔ + outtree.consense – в Newick формате
- ✔ Представление дерева в графическом режиме одной из программ – Drawtree или Drawgram (без bootstraps) - или другими программами

MEGA: филогенетический анализ последовательностей