Вычислительная филогенетика

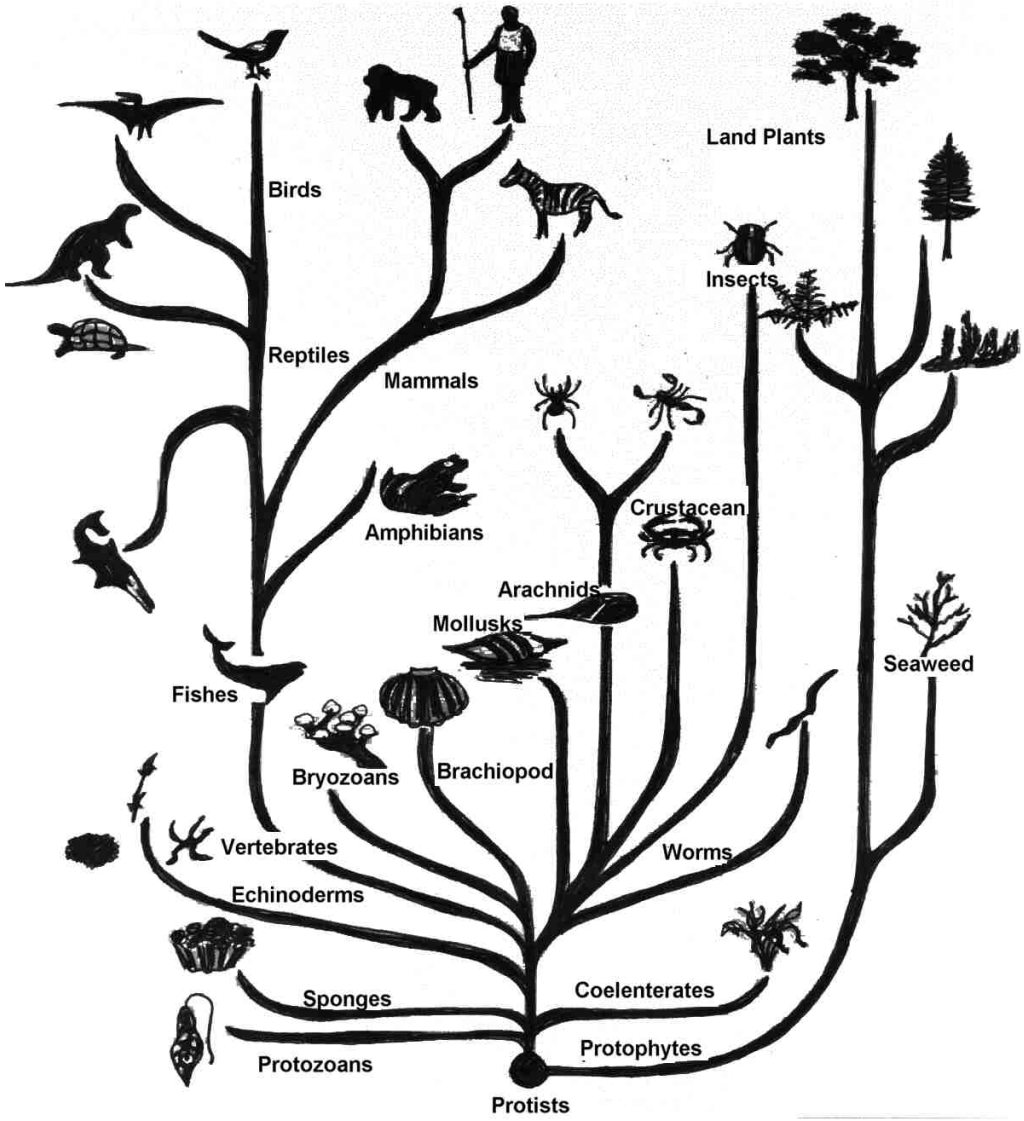
Филогенетические деревья

С.А.Спирин

sspirin@hse.ru

11 сентября 2019

Древо жизни



В наше время выяснять детали происхождения видов помогают последовательности ДНК и белков

Гены и белки

Геном
$3 \cdot 10^9$ букв у человека,
~ 10^6 букв у бактерий

содержит →

Кодирующие участки
<2% генома у человека,
~ 90% у бактерий

кодируют →

Белки
~ 25 000 у человека,
600 – 6000 у бактерий

Генетический код

	T(U)	C	A	G
T(U)	TTT Phe	TCT Ser	TAT Tyr	TGT Cys
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys
	TTA Leu	TCA Ser	TAA Stop	TGA Stop
	TTG Leu	TCG Ser	TAG Stop	TGG Trp
C	CTT Leu	CCT Pro	CAT His	CGT Arg
	CTC Leu	CCC Pro	CAC His	CGC Arg
	CTA Leu	CCA Pro	CAA Gln	CGA Arg
	CTG Leu	CCG Pro	CAG Gln	CGG Arg
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser
	ATC Ile	ACC Thr	AAC Asn	AGC Ser
	ATA Ile	ACA Thr	AAA Lys	AGA Arg
	ATG Met	ACG Thr	AAG Lys	AGG Arg
G	GTT Val	GCT Ala	GAT Asp	GGT Gly
	GTC Val	GCC Ala	GAC Asp	GGC Gly
	GTA Val	GCA Ala	GAA Glu	GGA Gly
	GTG Val	GCG Ala	GAG Glu	GGG Gly

Аминокислоты

A Ala Alanine Аланин
R Arg Arginine Аргинин
N Asn Asparagine Аспарагин
D Asp Aspartic Acid Аспарагиновая кислота
C Cys Cysteine Цистеин
Q Gln Glutamine Глютамин
E Glu Glutamic Acid Глутаминовая кислота
G Gly Glycine Глицин
H His Histidine Гистидин
I Ile Isoleucine Изолейцин
L Leu Leucine Лейцин
K Lys Lysine Лизин
M Met Methionine Метионин
F Phe Phenylalanine Фенилаланин
P Pro Proline Пролин
S Ser Serine Серин
T Thr Threonine Треонин
W Trp Thryptophan Триптофан
Y Tyr Tyrosine Тирозин
V Val Valine Валин
"Stop" в таблице кода означает стоп-кодон — сигнал окончания трансляции.

Мутации

gatcaacactacttgacttcaa**g**acttaccataaagaaaac

gatcaacactacttgacttcaa**a**acttaccataaagaaaac

точечная замена

---

gatcaacactacttgacttcaa**ga**cttaccataaagaaaac

gatcaacactacttgacttcaacttaccataaagaaaac

делеция

---

gatcaacactacttgacttcaagacttaccataaagaaaac

gatcaacactacttgacttcaaga**ta**cttaccataaagaaaac

инсерция
(вставка)

Мутации (точечные замены) в гене

… AATCCGTCAAGTCTA…

… **Asn Pro Ser Ser Leu** …

1) “молчащая”(синонимическая)мутация

… AATCCGTCGAGTCTA…

… **Asn Pro Ser Ser Leu** …

2) замена остатка на близкий по свойствам

… AATCCGACAAGTCTA…

… **Asn Pro Thr Ser Leu** …

3) замена остатка на остаток с иными свойствами

… AATCCGTCAAGACTA…

… **Asn Pro Ser Arg Leu** …

Эволюция белков

Мутации возникают случайно.

Конкретная мутация может быть:

- летальной;
- вредной;
- слабовредной;
- нейтральной;
- полезной.

Мутация порождает **полиморфизм** данного белка в популяции.

Доля каждого варианта подвержена случайным изменением (модель: «случайное блуждание с поглощением»).

За исторически короткое время один из вариантов (старый или новый) исчезает. Во втором случае говорят, что мутация **закрепилась.**

Мы видим лишь закрепившиеся мутации

А шанс закрепиться есть лишь у безвредных мутаций…

```
CYB5_CHICK          1 MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHP     50
                        .:|...||  .:||.|||:||||:|:|||:|:||::||:||||:|||
CYB5_HUMAN          1 ---MAEQSDEA--VKYYTLEEIQKHNHSKSTWLILHHKVYDLTKFLEEHP     45

CYB5_CHICK         51 GGEEVLREQAGGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKP    100
                      |||||||||||||||||||||||||.:|:|||||||||||||||.||
CYB5_HUMAN         46 GGEEVLREQAGGDATENFEDVGHSTDAREMSKTFIIGELHPDDRPKLNKP     95

CYB5_CHICK        101 AETLITTVQSNSSSWSNWVIPAIAAIIVALMYRSYMSE-    138
                      .||||||:.|:||.|:||||||:|:.||||||.||:|
CYB5_HUMAN         96 PETLITTIDSSSSWWTNWVIPAISAVAVALMYRLYMAED    134
```

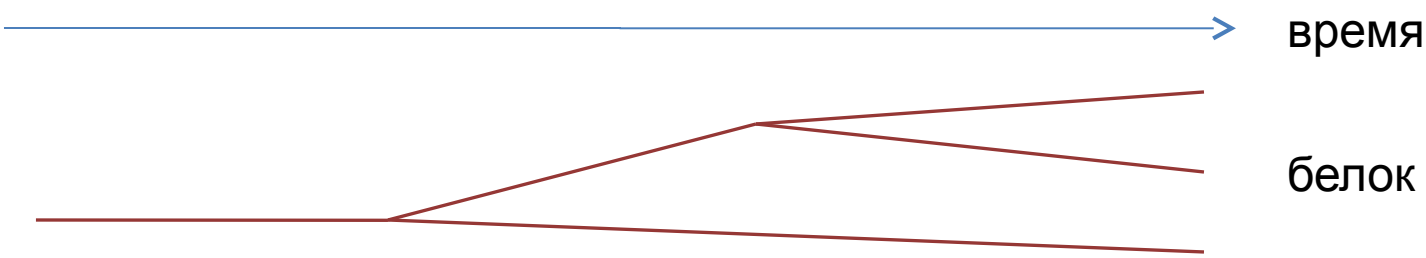
История белка

Приближённая картина: один белок – это конкретный белок в конкретный момент времени у конкретного вида живых организмов.

Можно (теоретически) проследить историю данного белка во времени. С течением времени последовательность белка меняется. Это и называется **эволюцией** белка.

При разделении вида на два все белки этих видов начинают эволюционировать независимо

Кроме того, нередко случается дупликация гена в геноме; после дупликации соответствующие белки также эволюционируют независимо



Эволюция видов и эволюция белков

Когда виды разделяются, то разделяются пути эволюции всех их белков…

В результате большинству белков одного вида соответствует **ортолог** в другом виде.

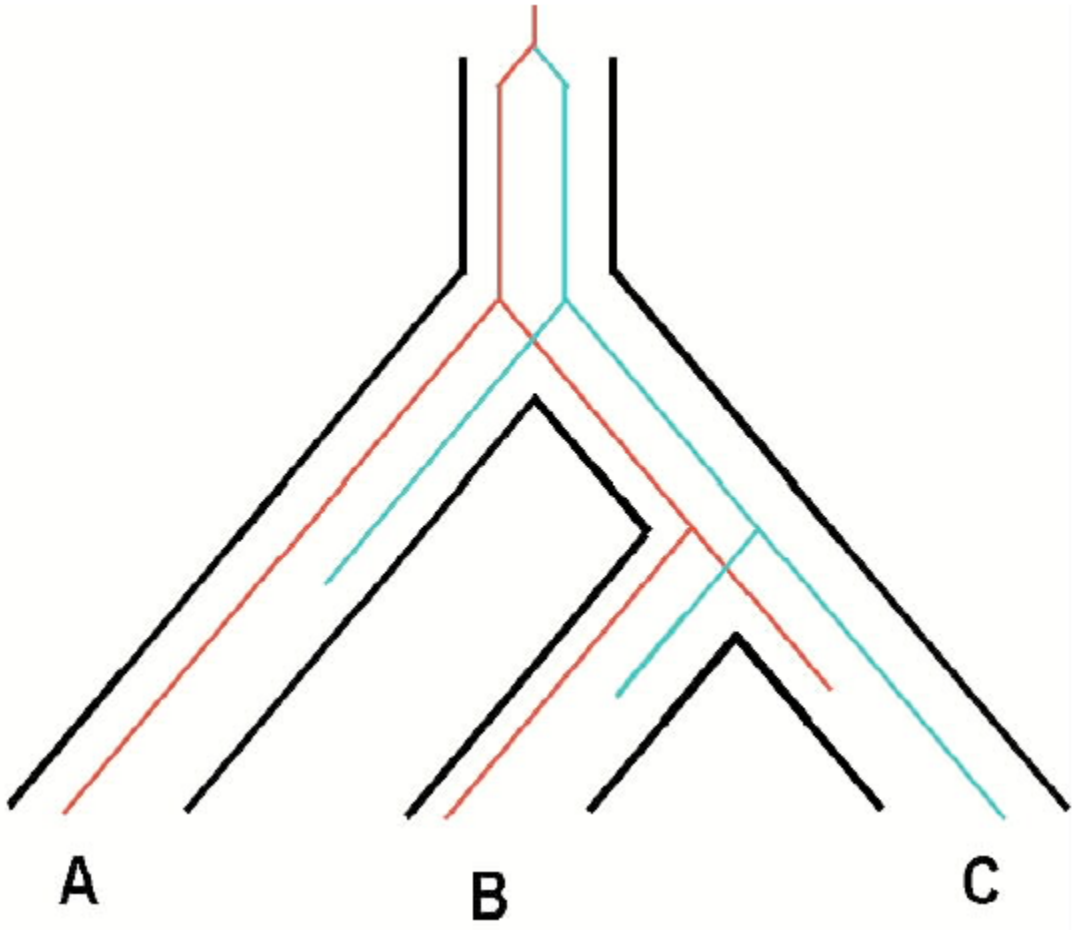
Но:

1) Бывают дупликации белков без разделения видов: два родственных белка существуют в одном геноме и эволюционируют (почти) независимо – такие белки называются **паралогами**
2) Бывают потери генов.

   Если в двух видах потерялись по одному белку из пары паралогов, то получается, что общий предок белков, которые выглядят как ортологи, «жил» существенно раньше, чем общий предок видов.
3) Бывают горизонтальные переносы генов.
4) Бывает, что два белка объединяются в один многодоменный, и наоборот.

   Поэтому правильнее говорить об эволюции белковых доменов.

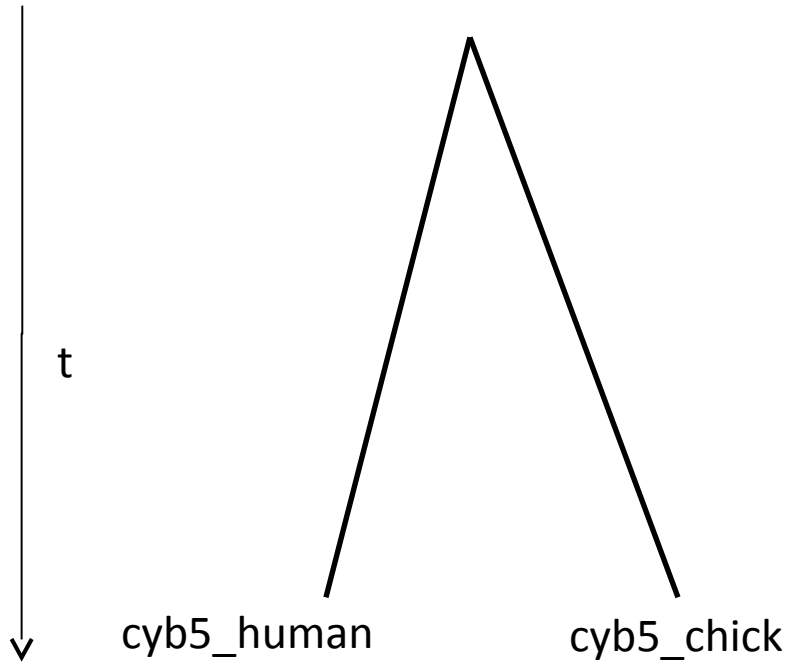
Дерево видов и дерево белков



Путь эволюции

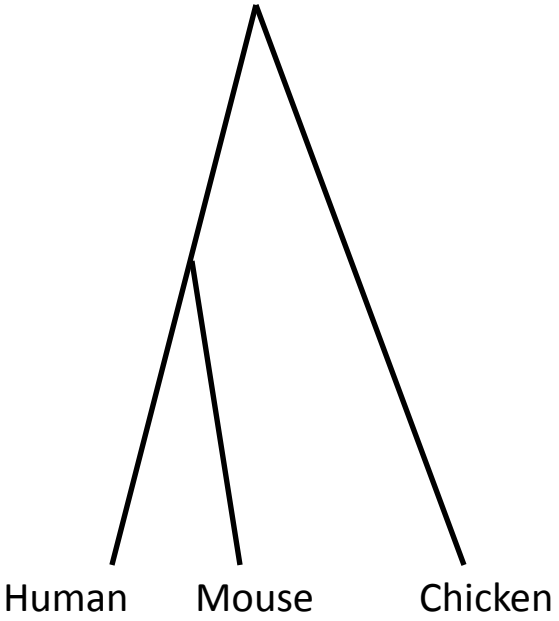
cyb5_human ——— cyb5_chick

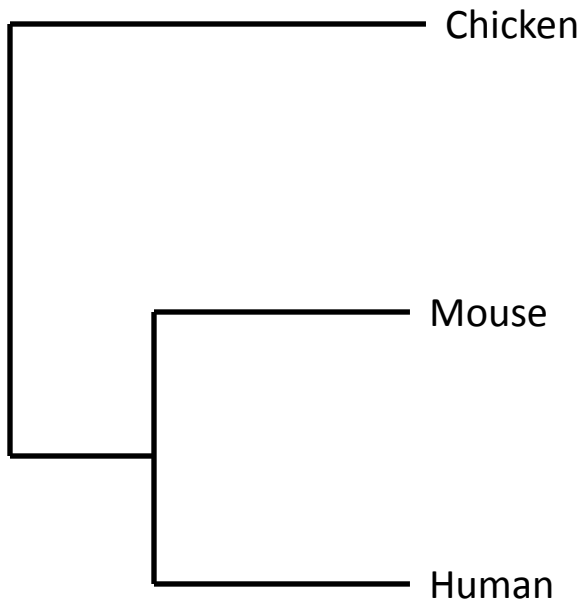
или



Каждая точка такой фигуры изображает существовавшую когда-то последовательность

Филогенетическое дерево

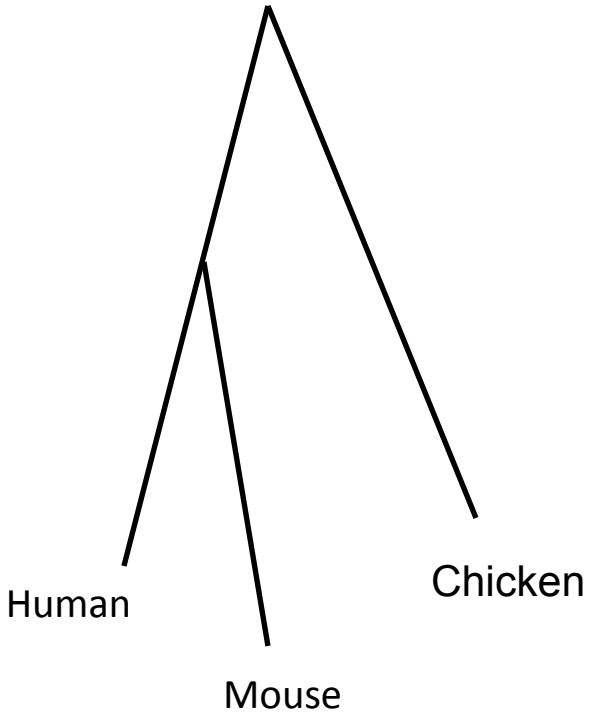


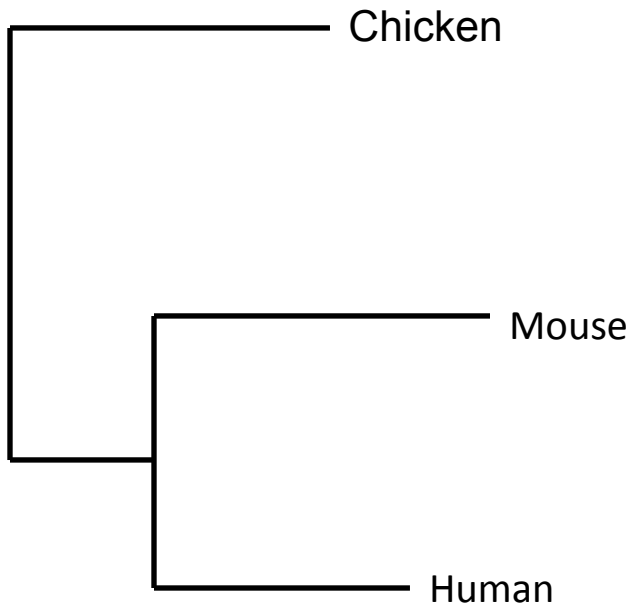


Слева — угловой вид, любая точка — последовательность, линии расходятся там, где разошлись пути эволюции.

Справа — прямоугольный вид, моментам расхождения отвечают не точки, а целые вертикальные линии.

«Молекулярные часы»: всегда идут, но иногда неточно

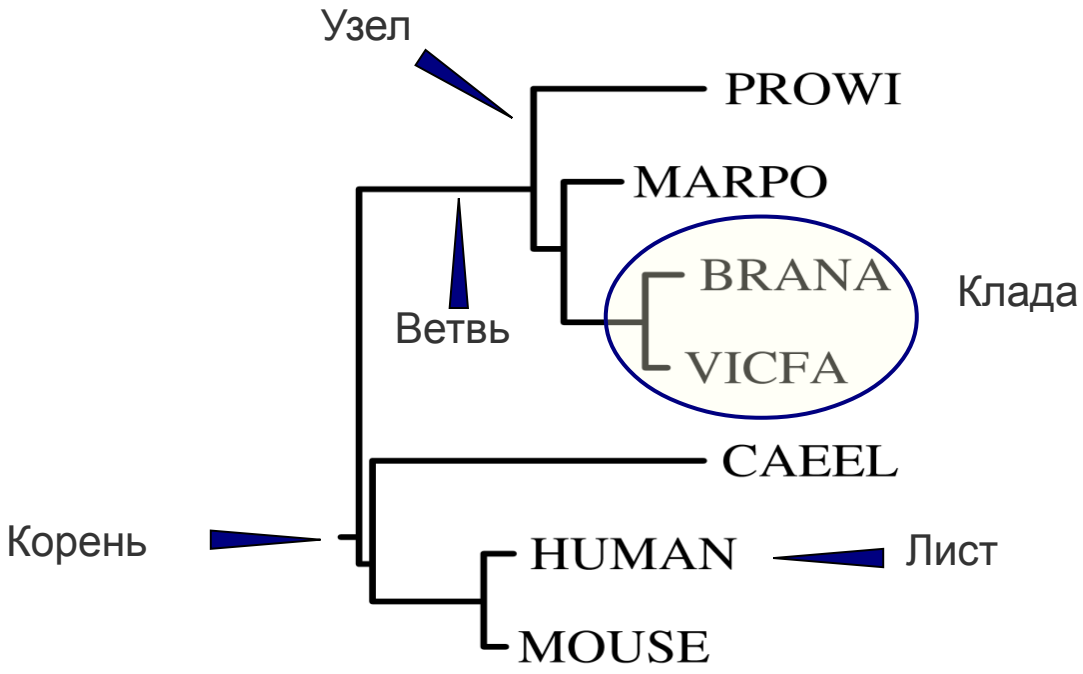




Когда хотят отразить разное число мутаций, произошедших на пути от общего предка, рисуют что-нибудь подобное.

Филогенетическое дерево (терминология)

- **Узел (node)** — точка разделения предковой последовательности. Соответствует внутренней вершине графа, изображающего эволюцию.
- **Лист (leaf)** — реальный (современный) объект; внешняя вершина графа.
- **Ветвь (branch)** — связь между узлами или между узлом и листом; ребро графа.
- **Корень (root)** — гипотетический общий предок всех рассматриваемых объектов.
- **Кла́да (clade)** — группа всех потомков некоторого ранее существовавшего объекта.



Длины ветвей дерева

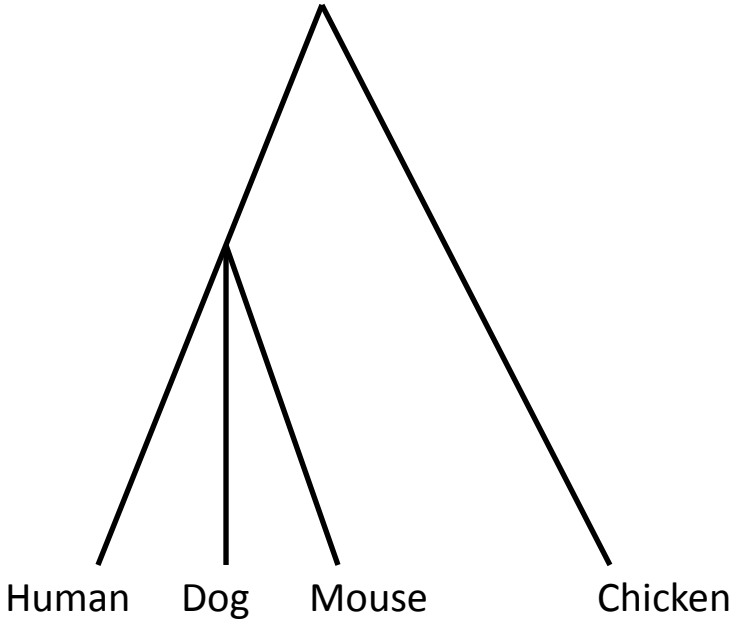
Каждая точка дерева – некоторая последовательность, существовавшая в некоторый момент времени (в прошлом, если эта точка – не лист).

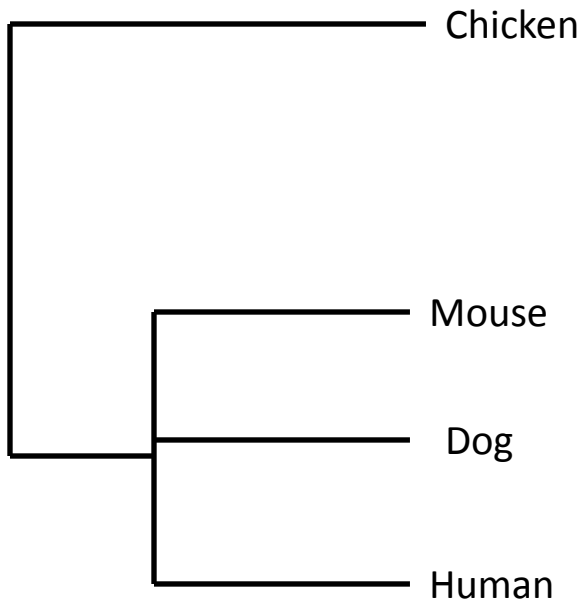
Длины ветвей могут иметь двоякий смысл:

1) интервал времени между моментами существования двух последовательностей;

2) число мутаций, случившихся на пути от одной последовательности до другой.

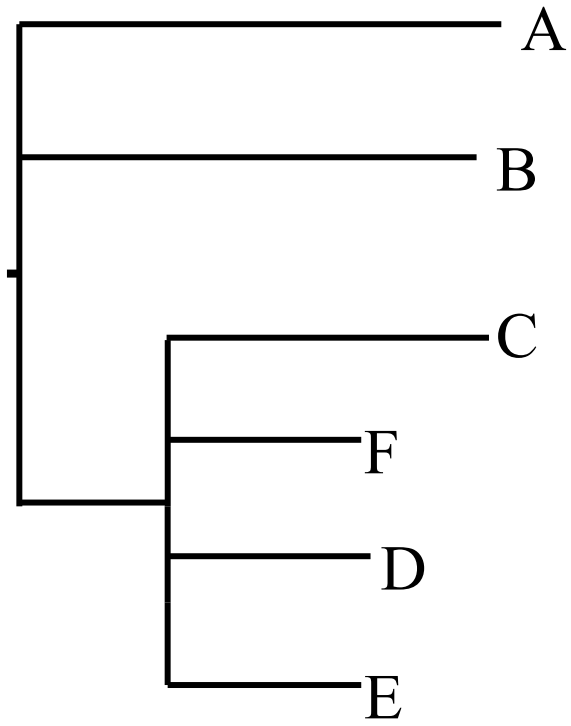
Небинарное дерево



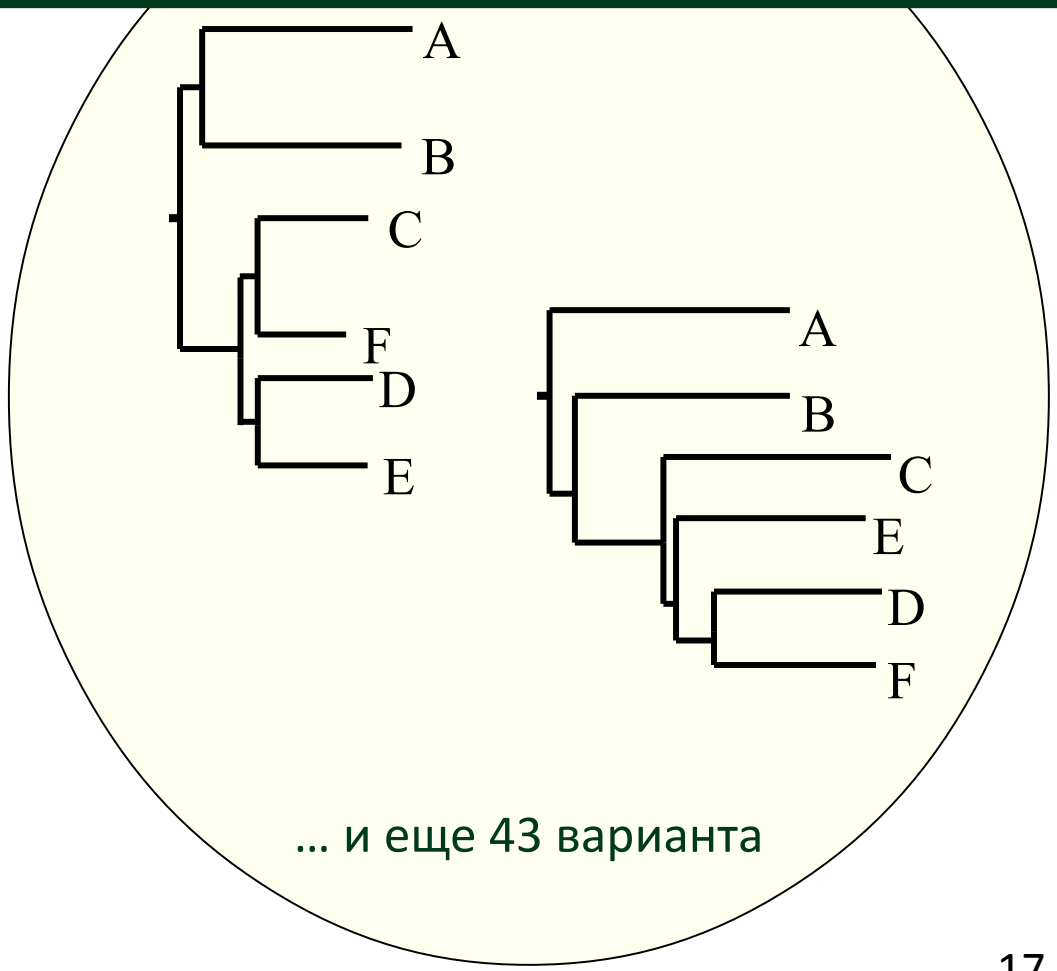


Часто вместо «небинарное» (non-binary) говорят «неразрешённое» (not resolved) дерево: некоторые узлы не разрешены до бинарных узлов.

Небинарное дерево следует понимать как множество возможных «разрешений»

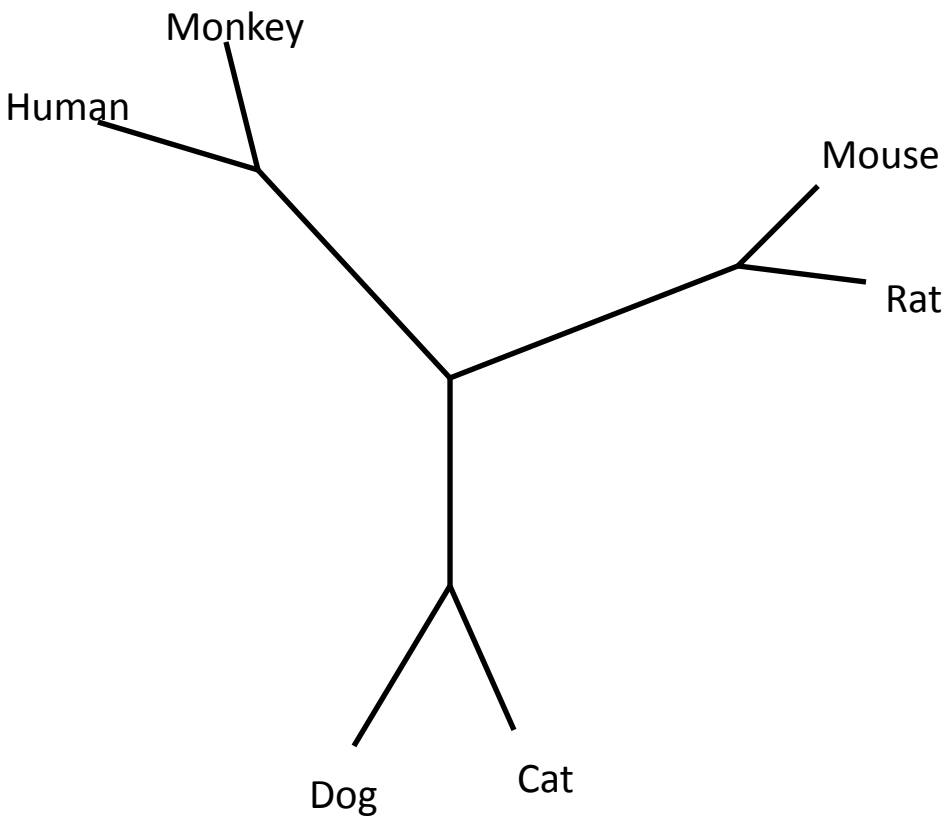


=

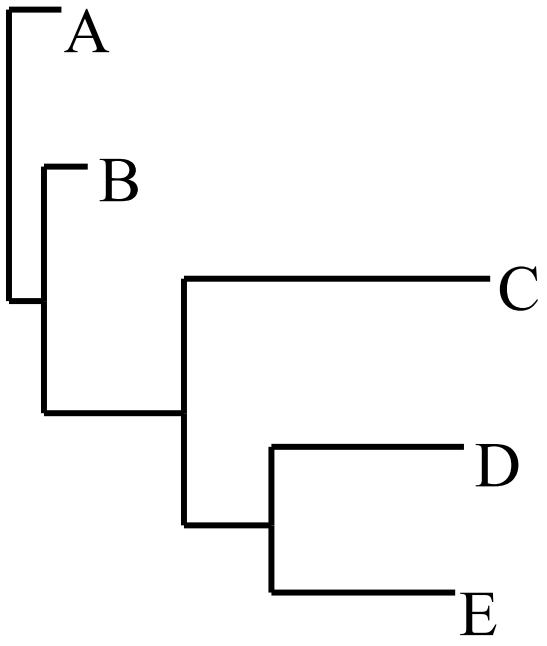


Термины «небинарное» и «неразрешённое» будем считать синонимами.

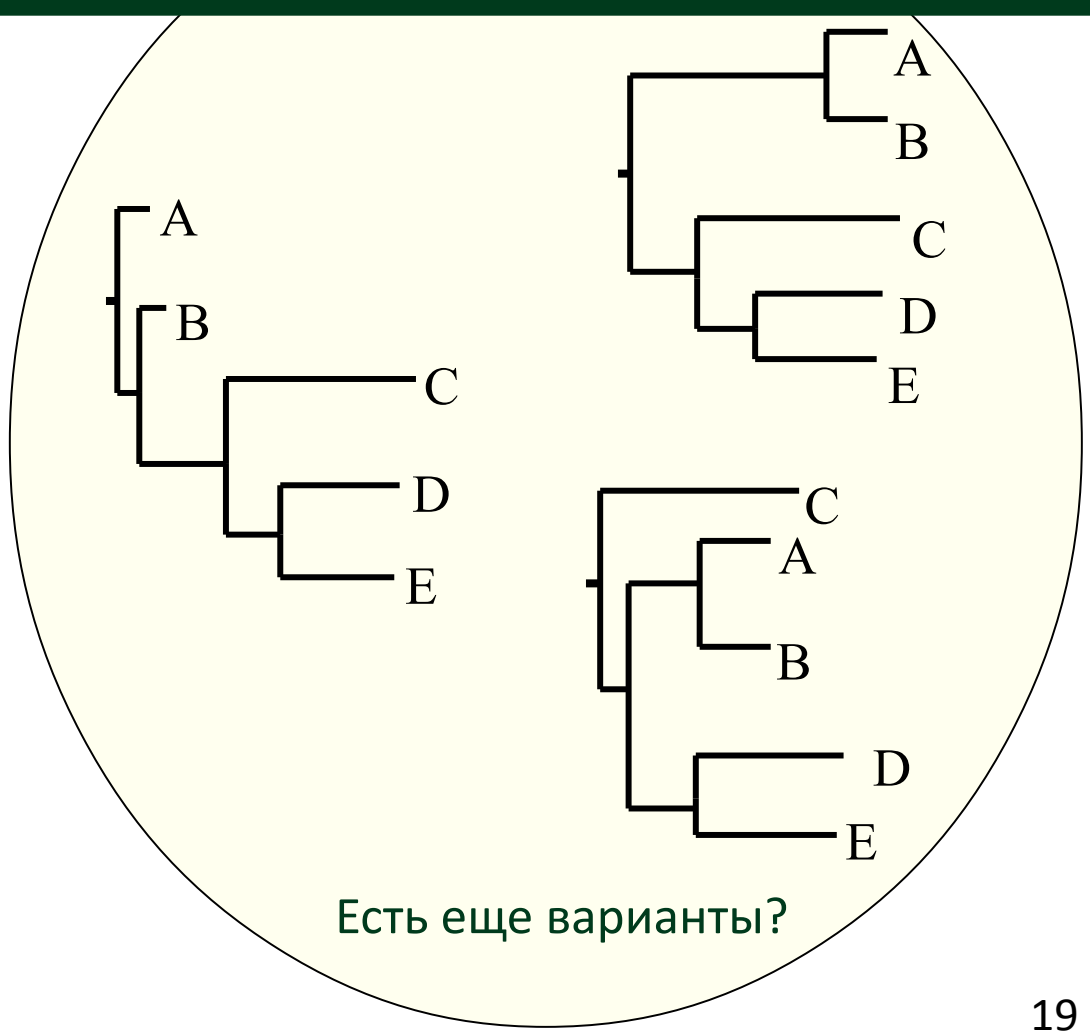
Неукоренённое дерево



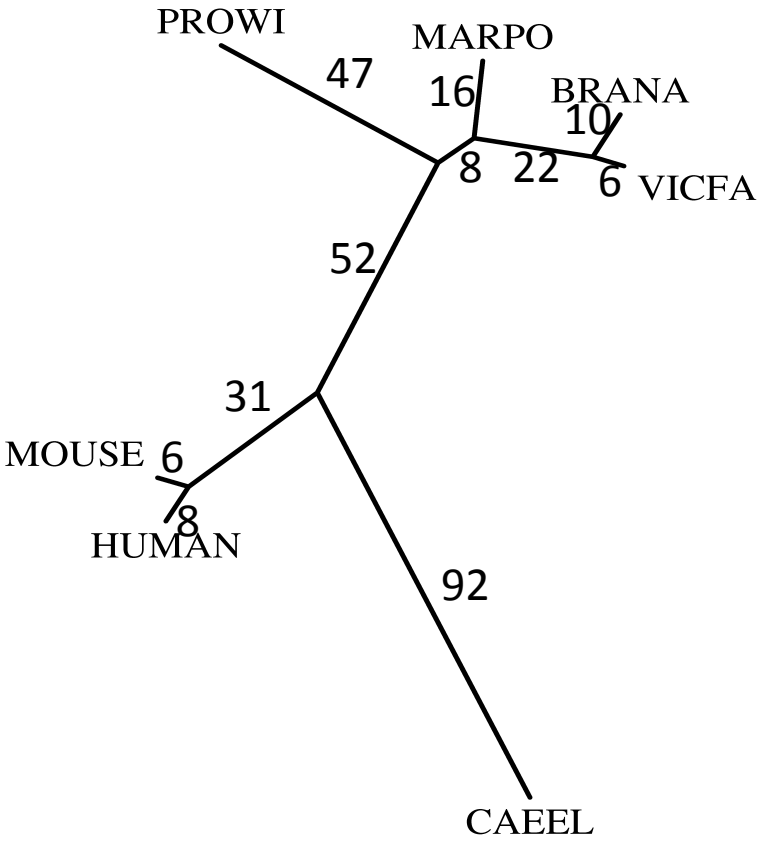
Неукоренённое дерево следует понимать как множество возможных укоренений



=

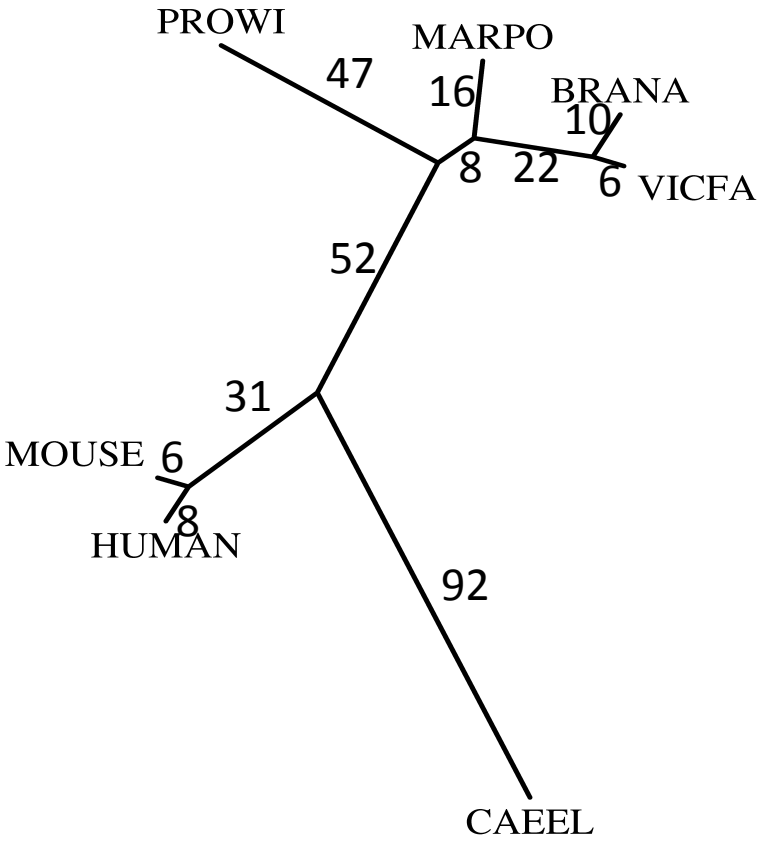


Длины ветвей и расстояния по дереву между листьями



D(MOUSE, CAEEL) = 6+31+92 = 129

Длины ветвей и расстояния по дереву между листьями



D(MOUSE, CAEEL) = 6+31+92 = 129

Дерево с заданными длинами ветвей порождает метрическое пространство, элементами которого являются листья

Ультраметрические деревья

Дерево называется ультраметрическим, если на нём есть точка, расстояния от которой до всех листьев одинаковы.

В этом случае множество листьев является ультраметрическим пространством: для любых трёх листьев a,b,c верно $d(a,b) \leq \max\big(d(a,c), d(b,c)\big)$.

Если все листья представляют **современные** последовательности, а длины ветвей имеют смысл **времени**, то дерево ультраметрическое.

Молекулярные часы

Гипотеза молекулярных часов: за одинаковое время происходит в среднем одинаковое число мутаций

Если гипотеза верна, то можно оценивать **эволюционное время** между современными последовательностями и на основании этих оценок строить **укоренённое ультраметрическое дерево.**

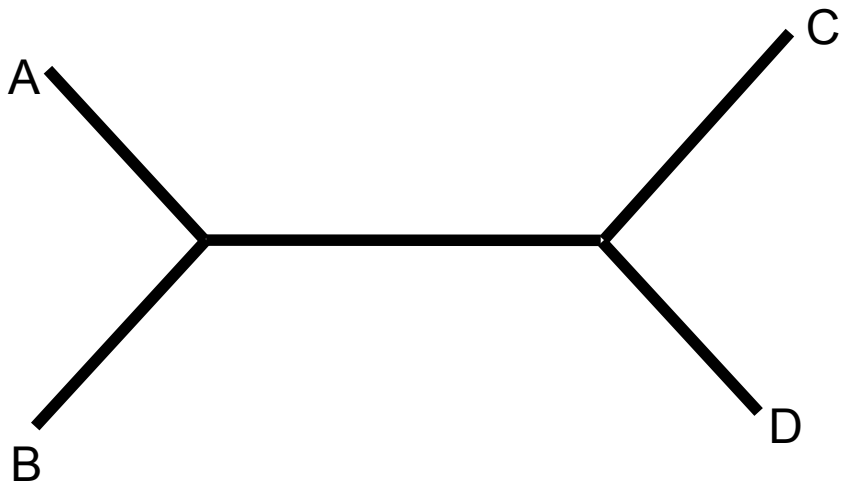
Но гипотеза МЧ часто не выполняется.

Расстояние как число мутаций

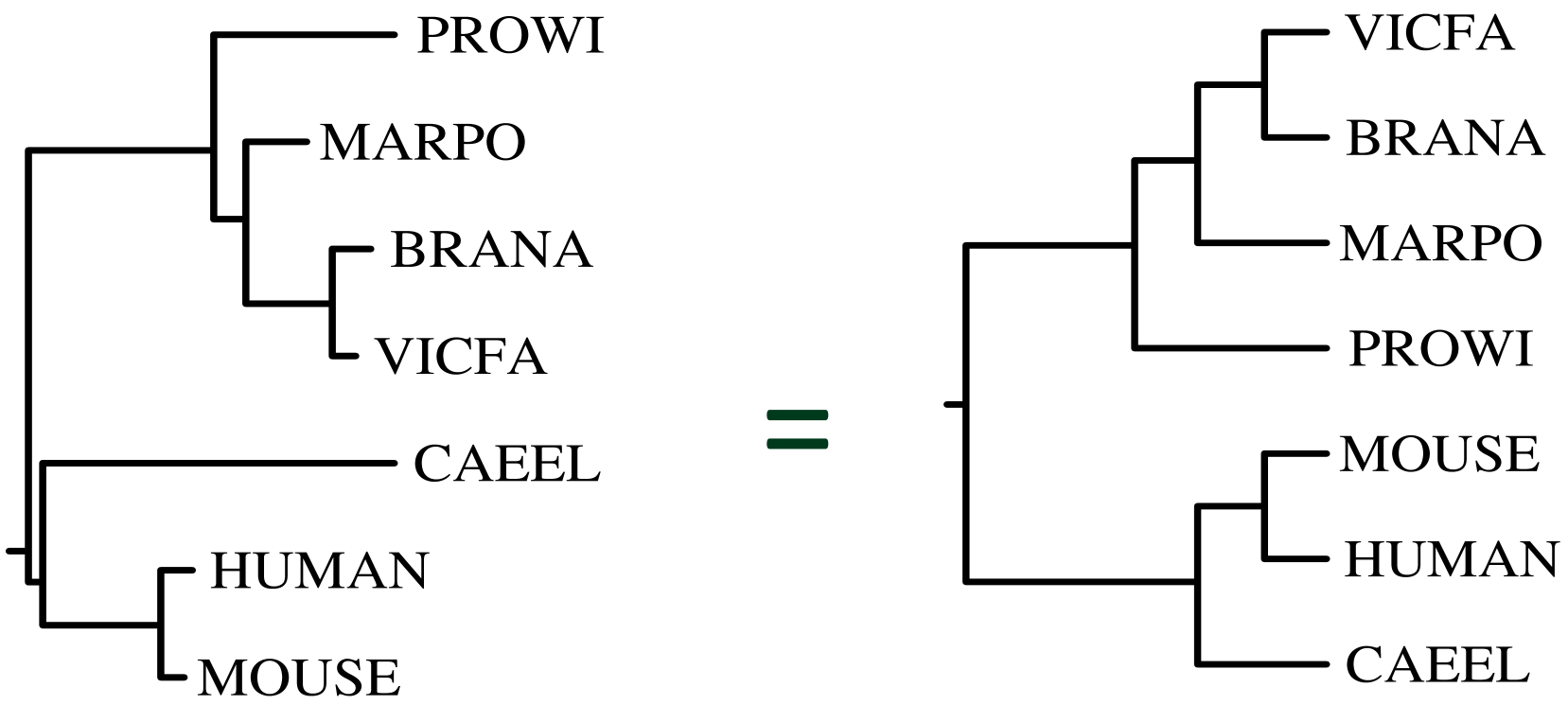
Расстояние между последовательностями ультраметрично, если его понимать как эволюционное время...

Но если неверно предположение о «молекулярных часах», то удобнее понимать расстояние как числа произошедших мутаций. **Такое расстояние не обязательно ультраметрично.**

Для расстояний по дереву выполняется свойство, названное **«аддитивность»:** для любых четырёх листьев A,B,C,D из трёх сумм
1) $d(A,B) + d(C,D)$ 2) $d(A,C) + d(B,D)$ 3) $d(A,D) + d(B,C)$
две равны между собой и больше третьей.



Топология дерева



Топология дерева

Каждая ветвь разбивает множество листьев на два.

В каждом дереве есть **тривиальные** ветви (отделяющие один лист от всех остальных), они не зависят от топологии.

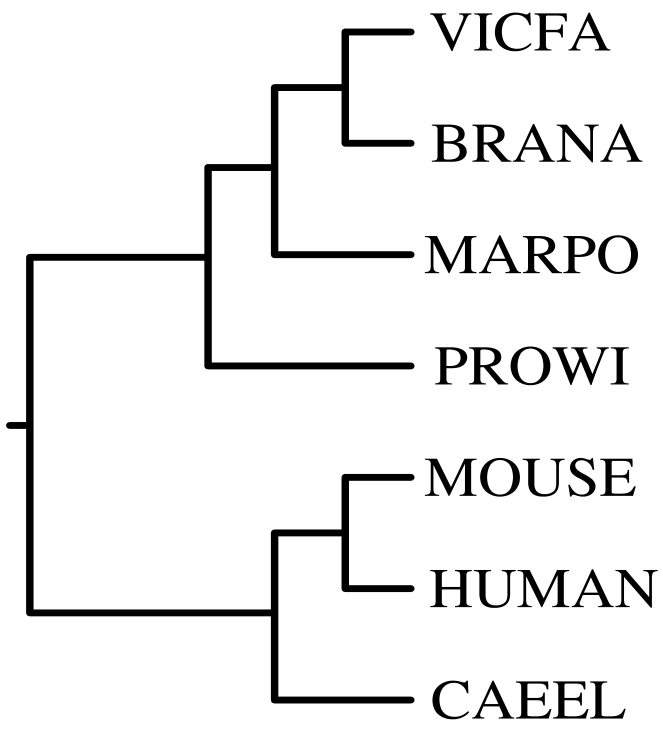
Топологию (неукоренённого) дерева можно однозначно записать набором нетривиальных разбиений. Например:

{HUMAN, MOUSE} vs {CAEEL,PROWI,MARPO,BRANA,VICFA}

{HUMAN, MOUSE,CAEEL} vs {PROWI,MARPO,BRANA,VICFA}

{HUMAN,MOUSE, CAEEL,PROWI} vs {MARPO,BRANA,VICFA}

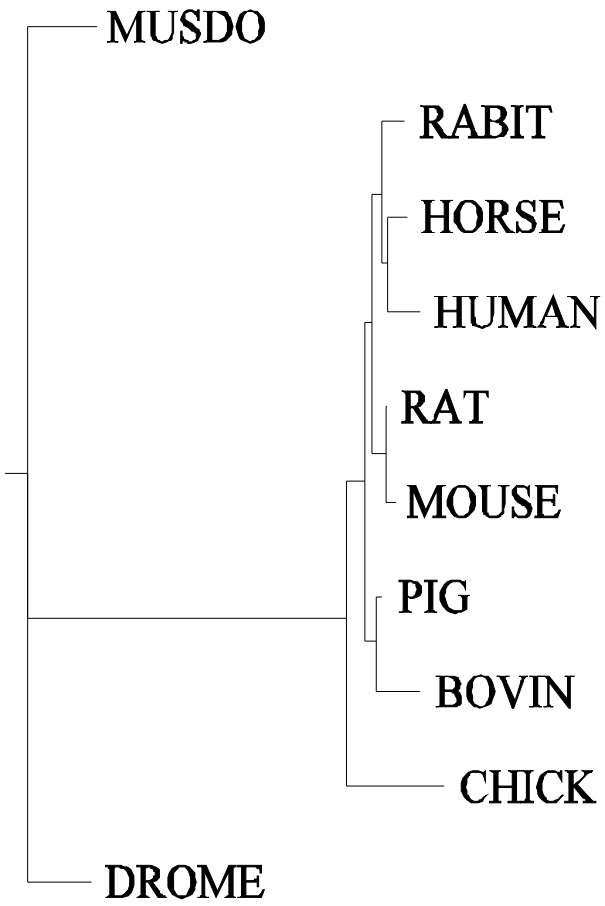
{HUMAN,MOUSE,CAEEL,PROWI,MARPO} vs {BRANA,VICFA}

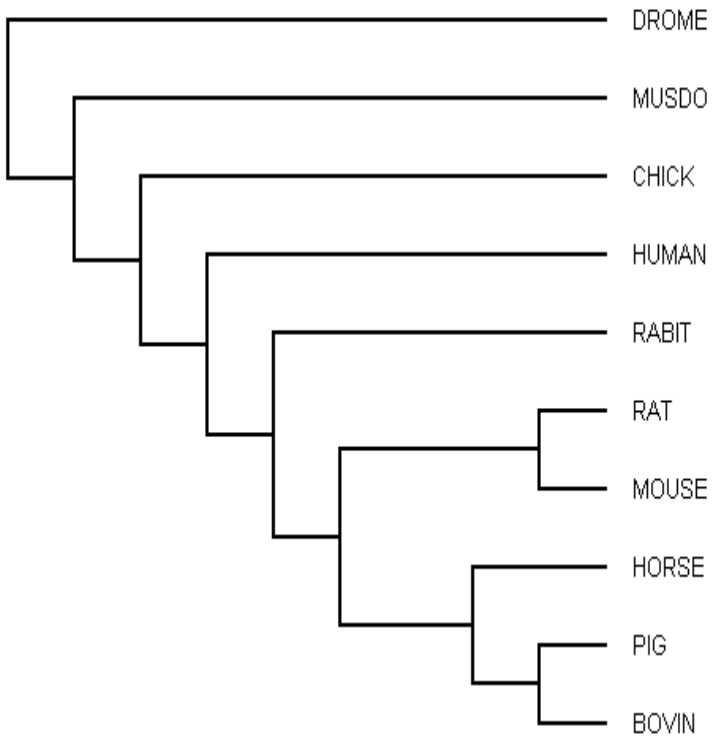


HUMAN	MOUSE	CAEEL	VICFA	BRANA	MARPO	PROWI
+	+	-	-	-	-	-
+	+	+	-	-	-	-
+	+	+	-	-	-	+
+	+	+	-	-	+	+

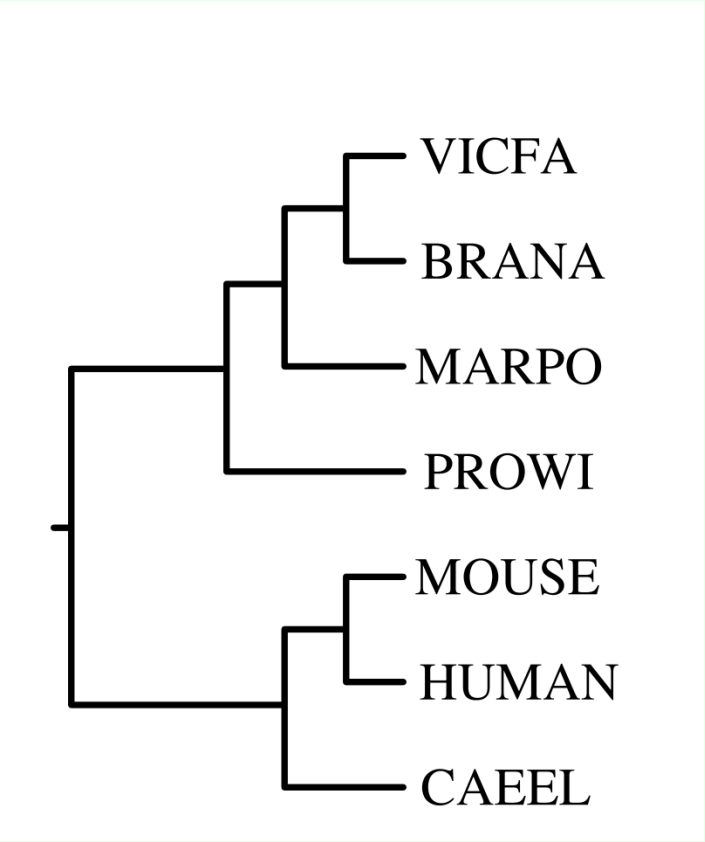
Представление топологии дерева разбиениями позволяет отождествлять ветви разных деревьев с одним и тем же множеством листьев.

В частности, если имеются две реконструкции эволюции по одним и тем же данным, то можно сказать, в каких ветвях они согласуются, а в каких – расходятся.





Скобочная формула (формат Newick)



Newick Standard:

((((VICFA:3, BRANA:3):3, MARPO:6):2, PROWI:8):7, ((MOUSE:3, HUMAN:3):3, CAEEL:6):15);

«The reason for the name is that the second and final session of the committee met at Newick's restaurant in Dover, and we enjoyed the meal of lobsters.»
Joseph Felsenstein, http://evolution.genetics.washington.edu/phylip/newicktree.html

Программы работы с деревьями

- MEGA

http://www.megasoftware.com/
Визуализация и построение деревьев (оконный интерфейс)

- Пакет PHYLIP

http://evolution.genetics.washington.edu/phylip.html
Построение и визуализация деревьев (интерактивный интерфейс с запуском из командной строки).

- FigTree

http://tree.bio.ed.ac.uk/software/figtree/
Визуализация деревьев

- iTol

https://itol.embl.de/
Визуализация деревьев online