

Вычислительная филогенетика

Часть 3

Парсимония, сравнение деревьев, укоренение, бутстреп, паралоги

С.А.Спирин

25 сентября 2019

Maximum parsimony

- Минимальное число мутаций, необходимых для получения данных последовательностей по данному дереву.
- Можно посчитать это число для каждой позиции выравнивания, а затем сложить по всем позициям.
- Не зависит от укоренения.
- Не включает вычисление длин ветвей.

Как посчитать минимальное число мутаций в одной позиции

- Введём операцию «парсимония» на множествах :
 - $A * B = A \cap B$, если $A \cap B \neq \emptyset$
 - $A * B = A \cup B$, если $A \cap B = \emptyset$
- Для каждой позиции выравнивания заполним все узлы дерева множествами по правилу: множество в узле равно парсимонии множеств в двух дочерних узлах
- В листьях поставим одноэлементные множества, содержащие буквы из выравнивания.
- Дойдём до корня и посчитаем, сколько раз пришлось брать объединение (то есть множества в потомках данного узла не пересекались). **Это число и будет минимальным числом мутаций.**

СРАВНЕНИЕ ДЕРЕВЬЕВ

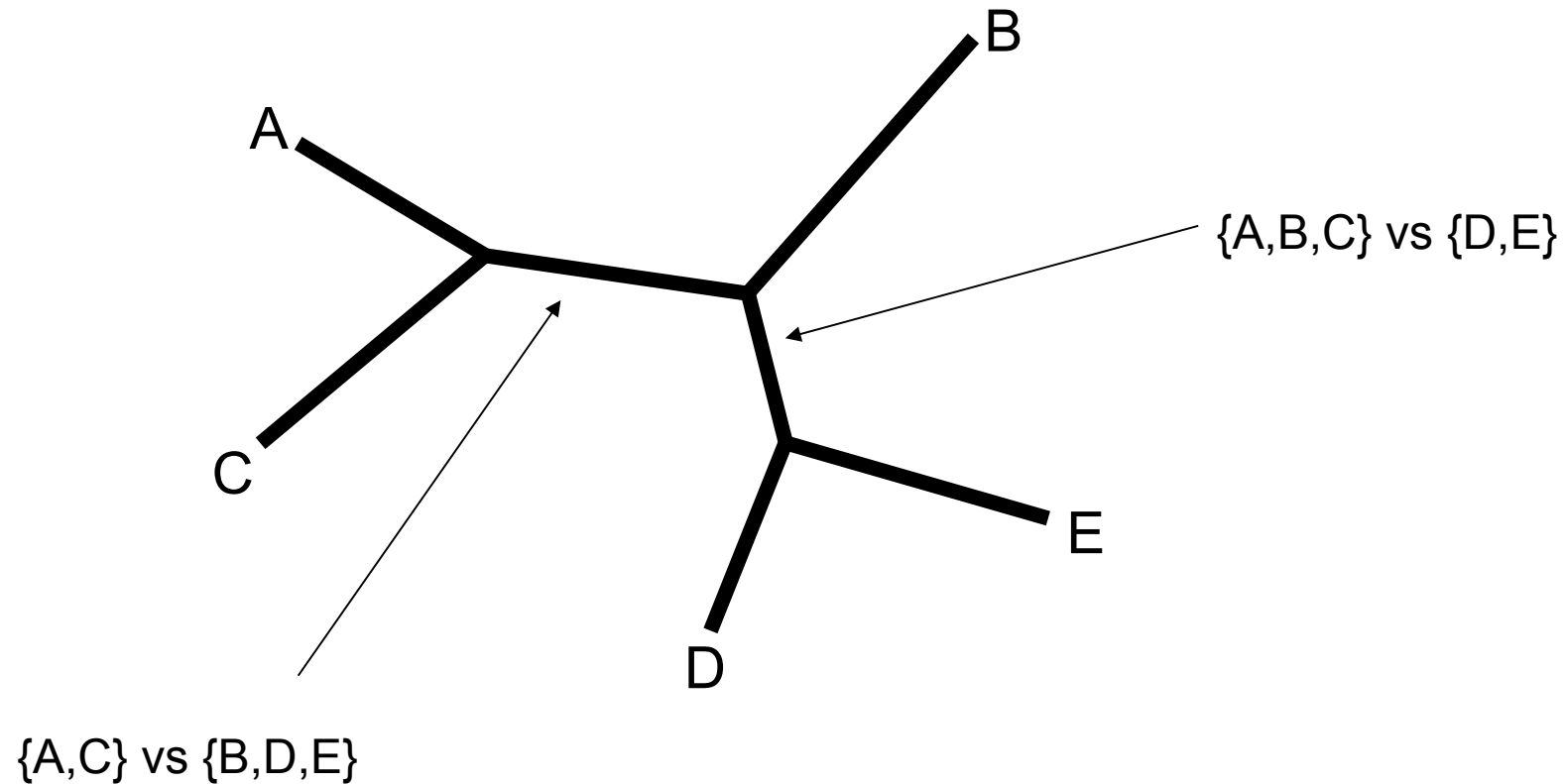
Сравнение деревьев

Программы реконструкции филогении так же ненадёжны, как и любые другие компьютерные программы предсказания биологических фактов.

Поэтому (в частности) возможны различные варианты реконструкции по одним и тем же данным.

Встаёт задача сравнения различных деревьев с одним и тем же множеством листьев.

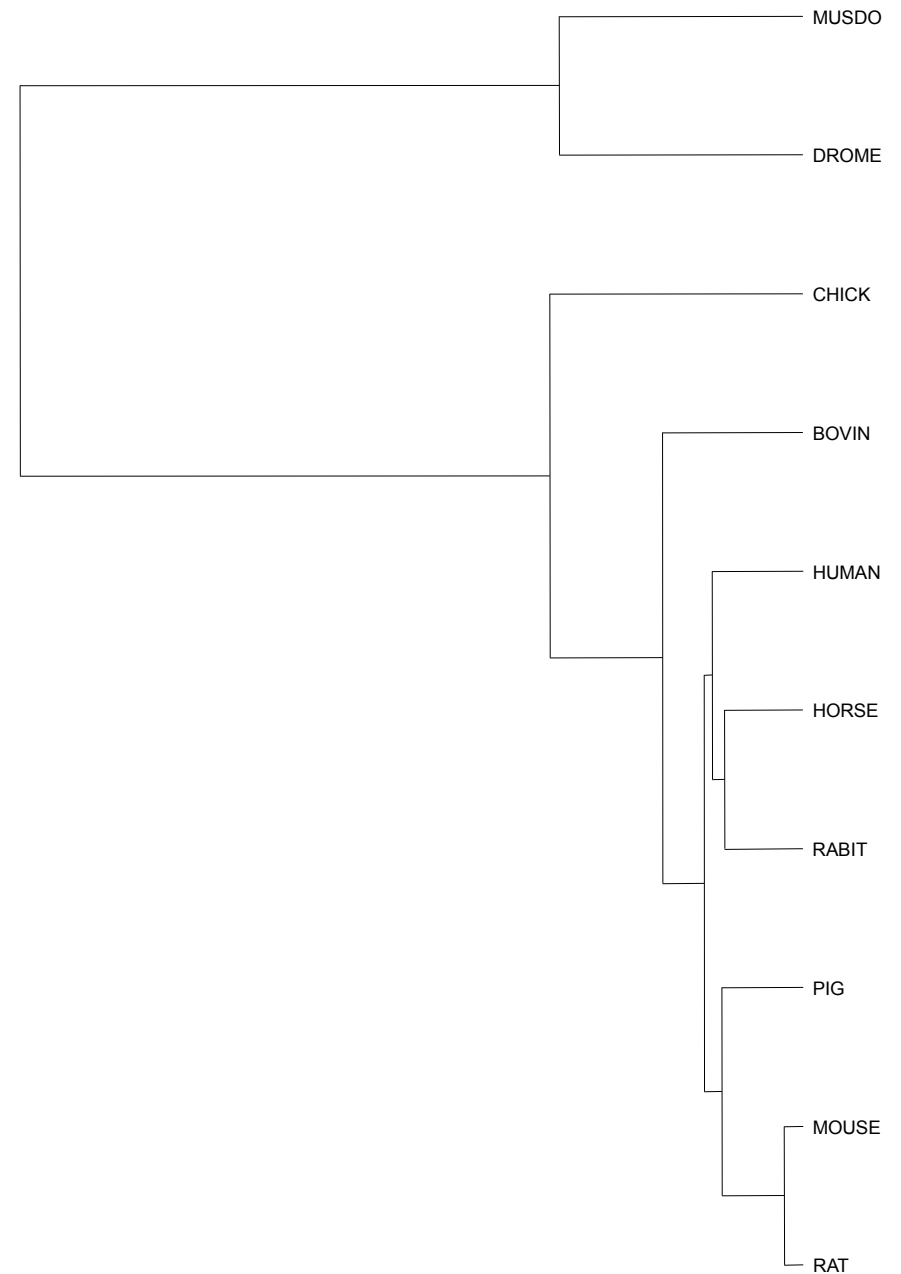
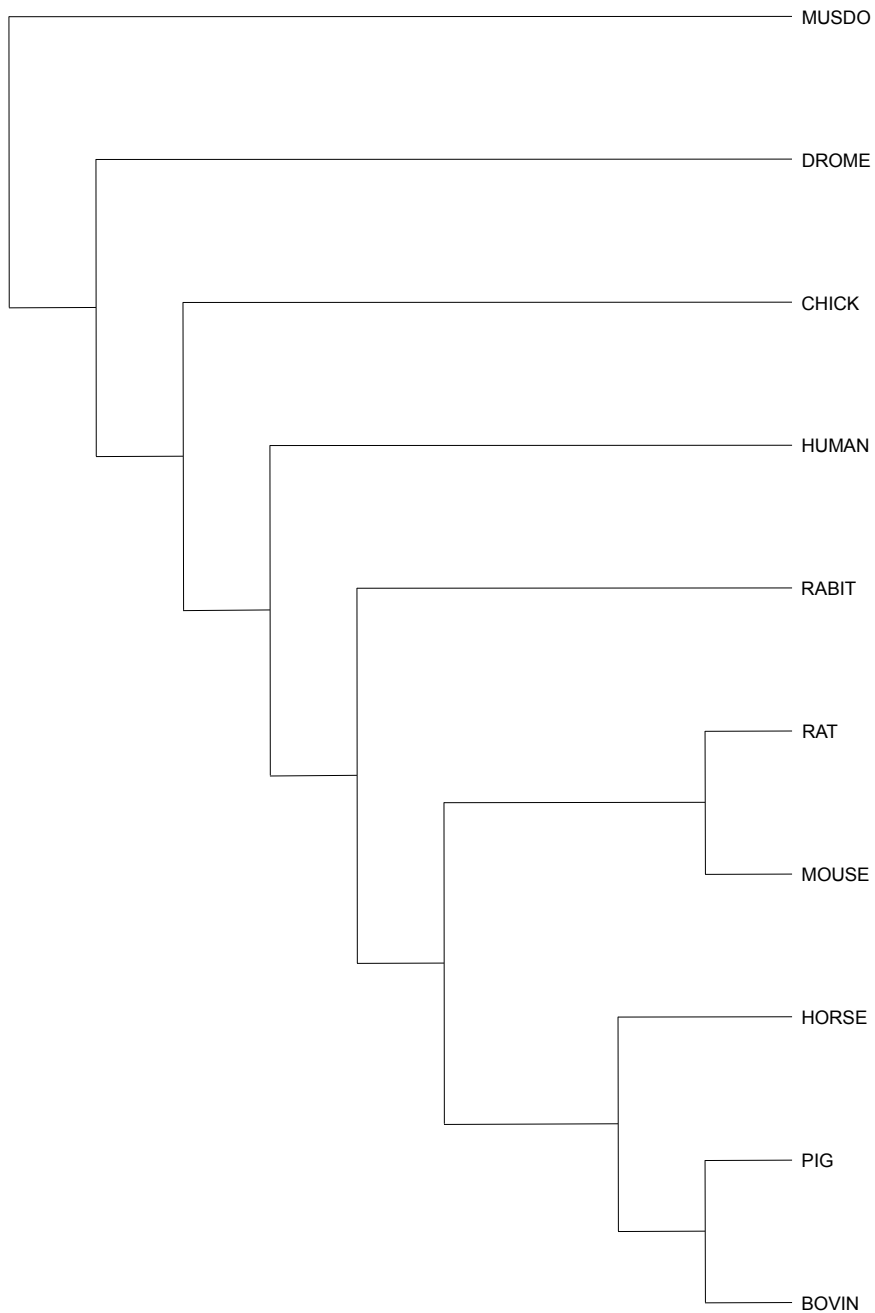
Напоминание: ветвь дерева как разбиение множества листьев

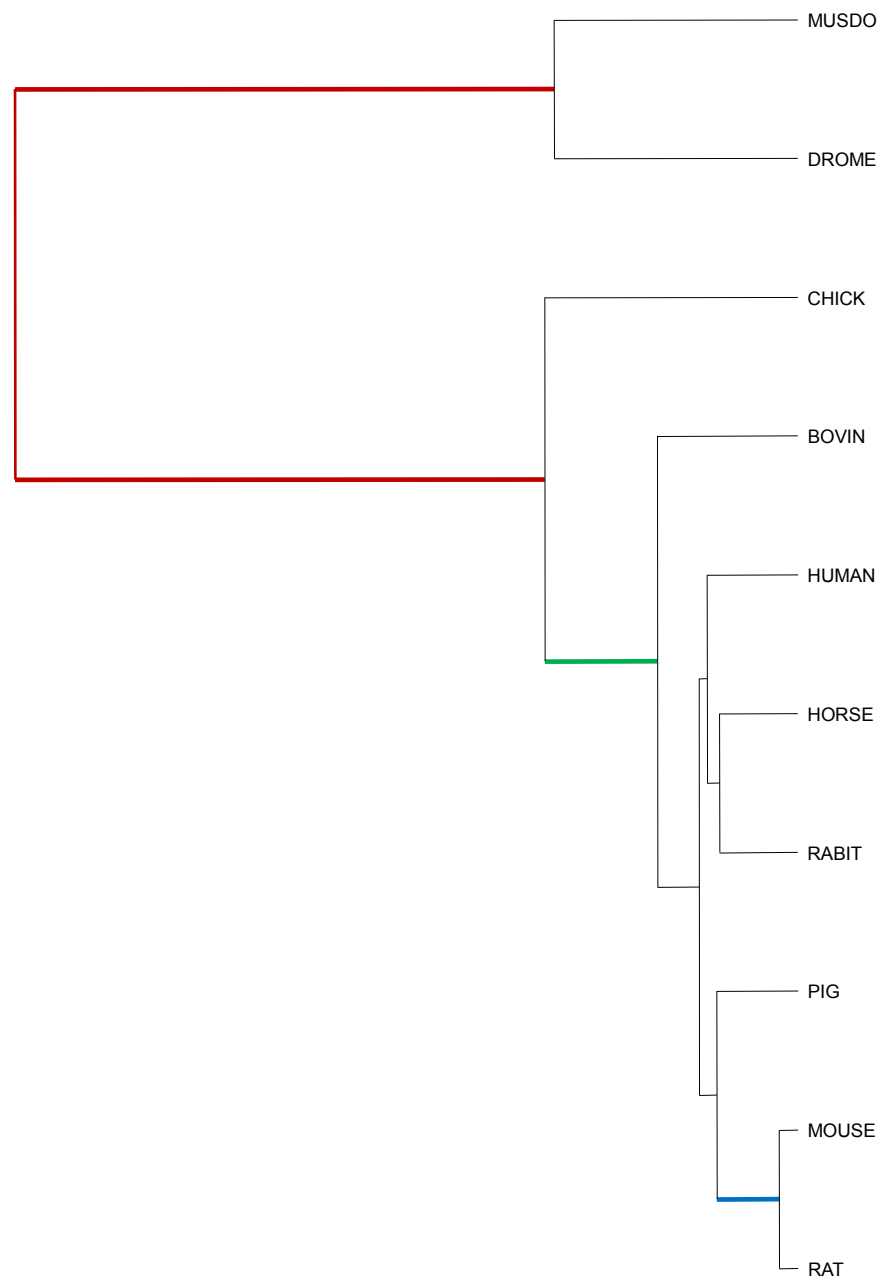
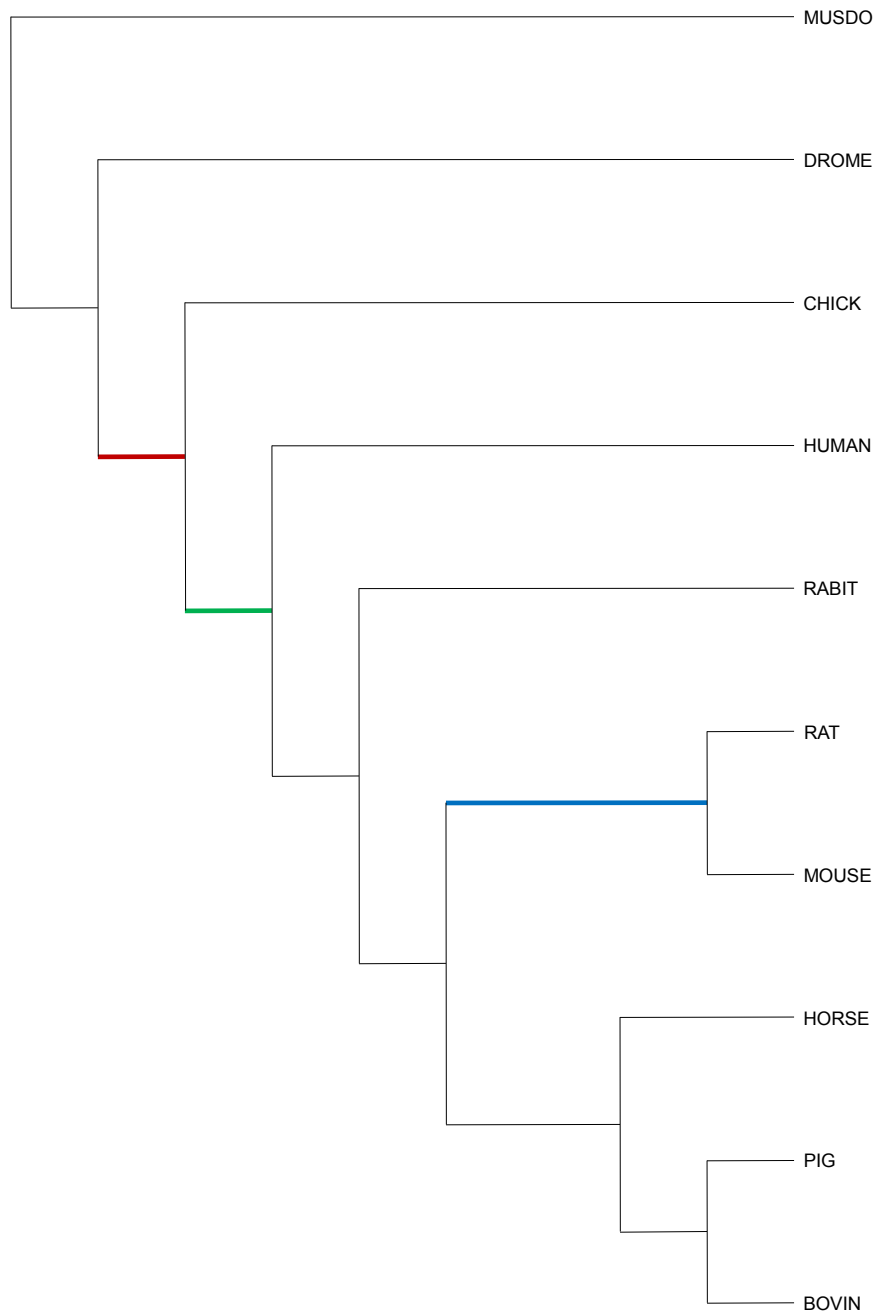


Расстояние Робинсона – Фоулдса

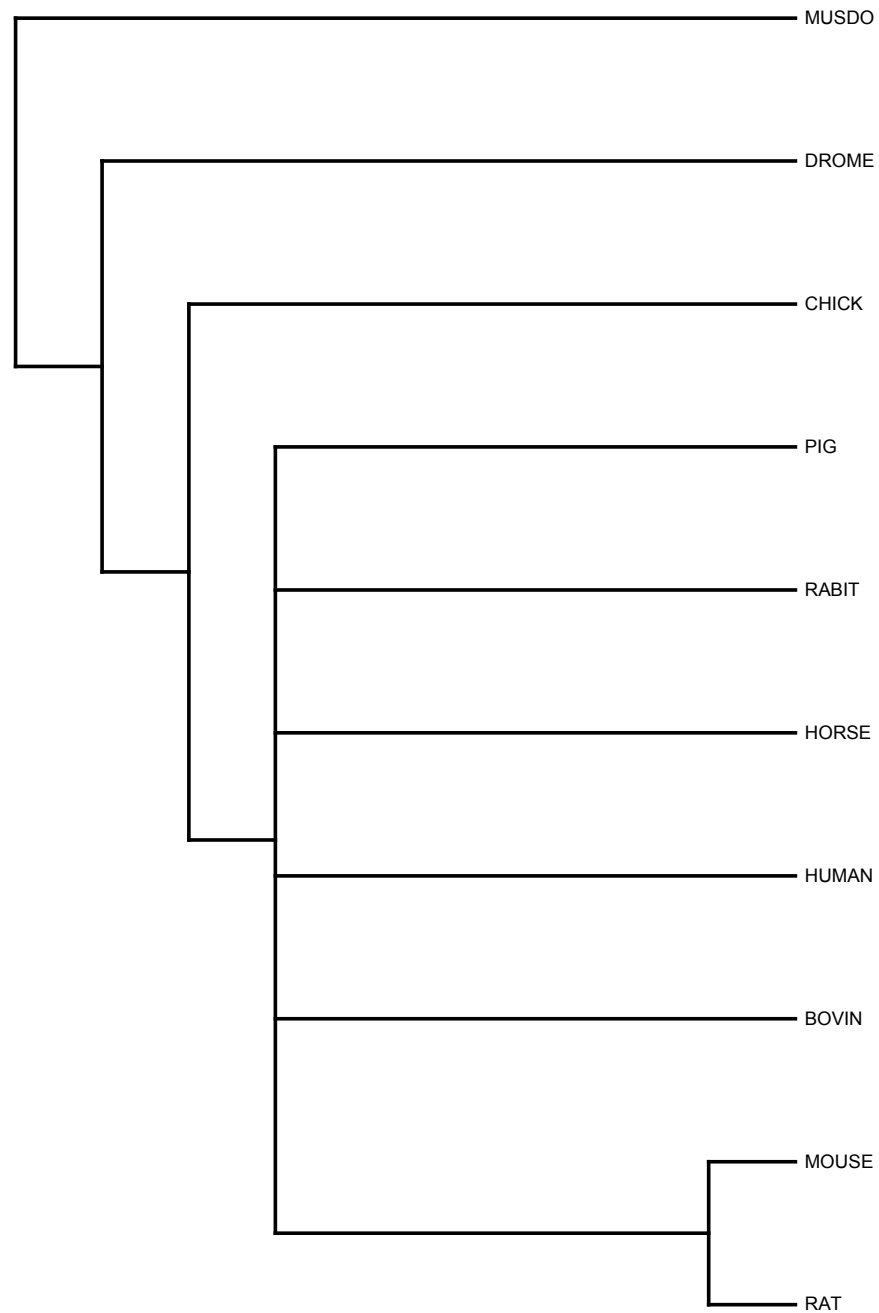
- Расстояние между двумя деревьями с одинаковым множеством листьев равно числу ветвей (т.е. разбиений), встретившихся только в одном из этих деревьев.
- Предложены разные другие меры расстояний (об этом позже)

Что общего у этих двух деревьев?





Консенсусное дерево

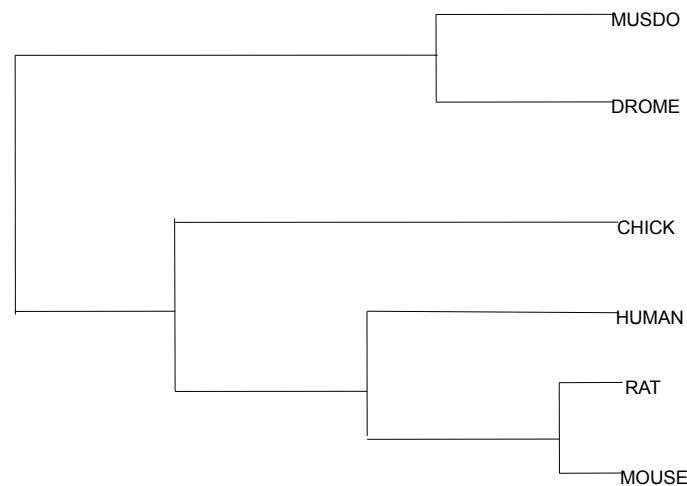
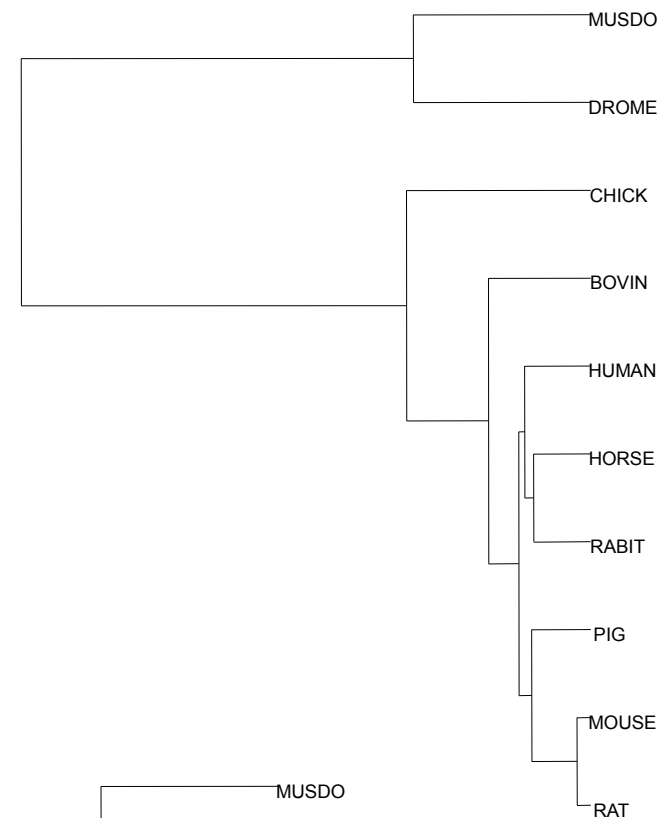
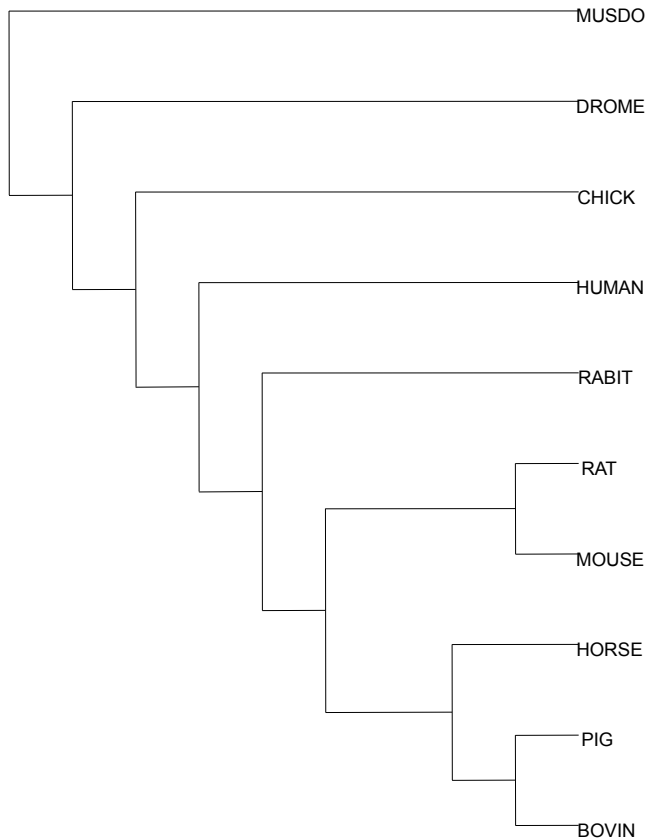


Совмещение данных от многих деревьев

По набору деревьев **с одинаковым множеством** листьев:

- 1** Консенсус (consensus)
Включает только те ветви, которые встретились **во всех** деревьях исходного набора
- 2** Дерево большинства (majority-rule tree)
Включает только те ветви, которые встретились **в большинстве** деревьев исходного набора
- 3** Дерево расширенного большинства (extended majority-rule tree)
К дереву большинства добавляются ветви, не противоречащие уже имеющимся, начиная с наиболее «поддержанных»

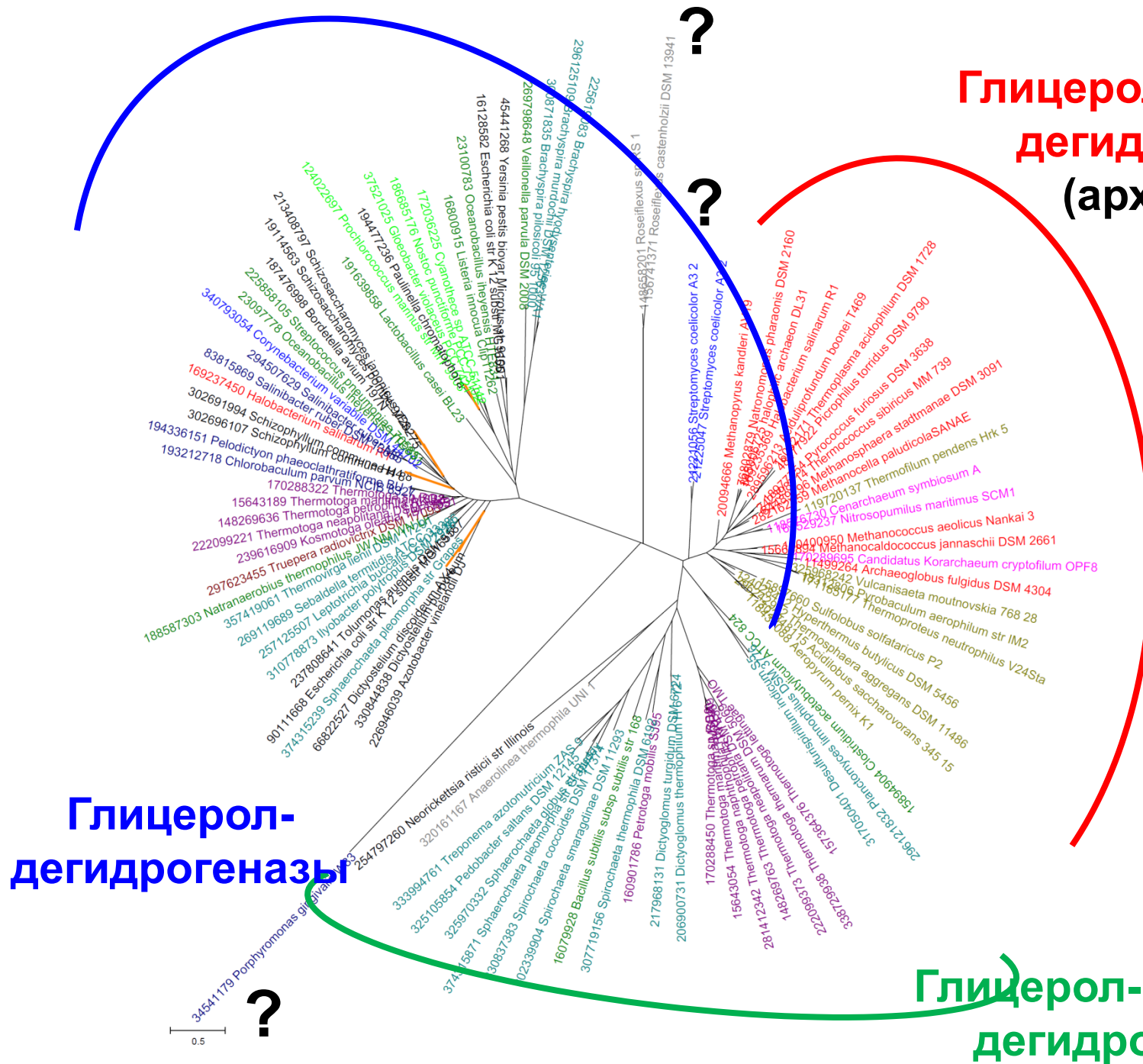
Наибольшее общее поддерево



Н.О.П. определено не однозначно!

УКОРЕНЕНИЕ

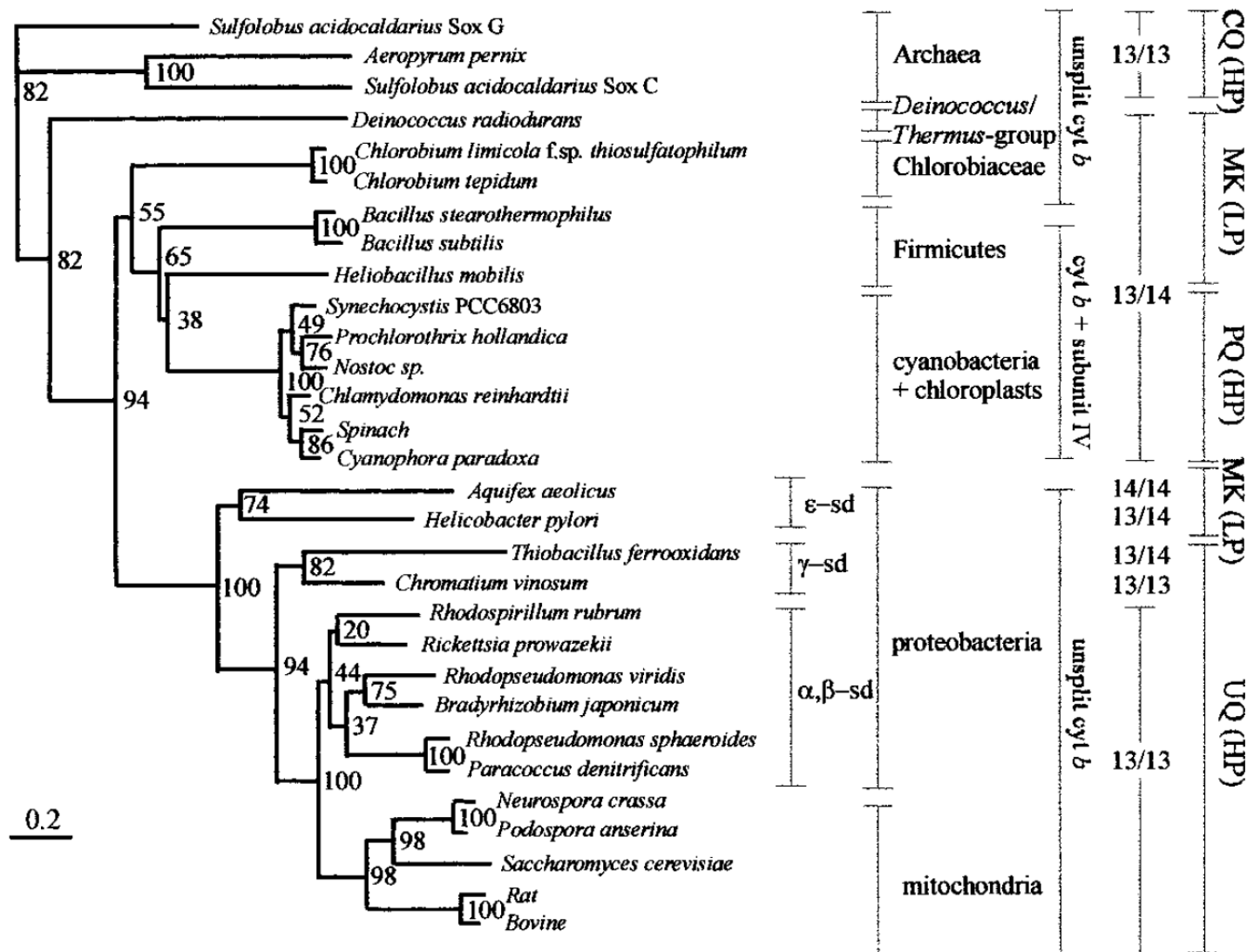
Укоренение дерева



От
положения
корня
зависит, как
видится
эволюция
фермента!

Укоренение

(Schütz *et al.*, 2000)



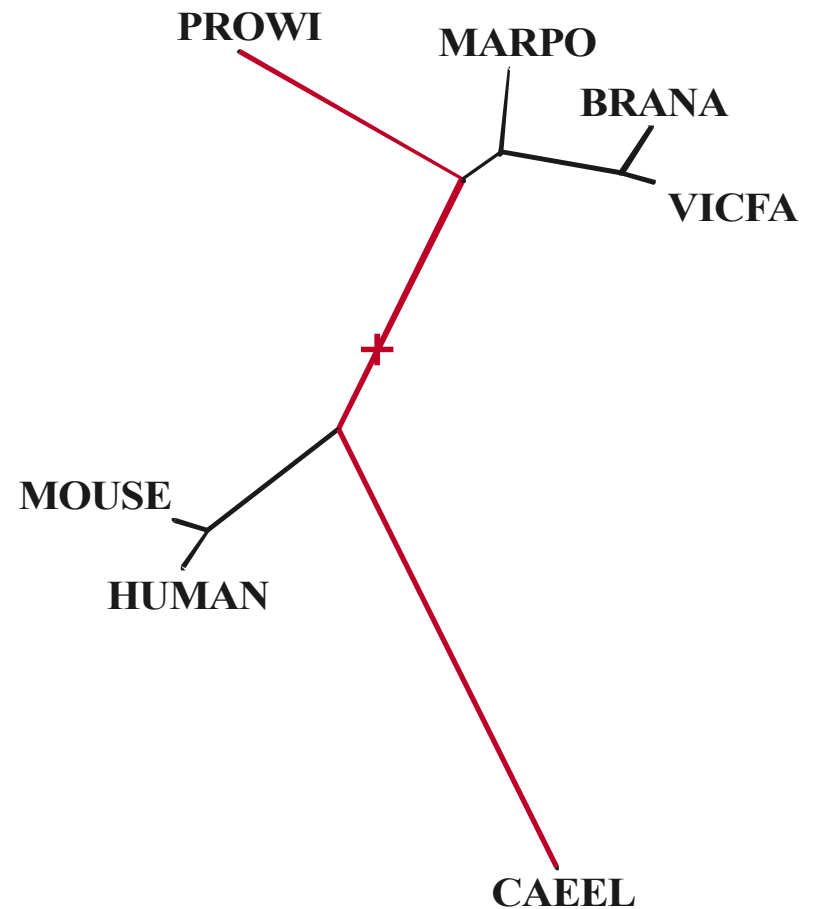
Для подтверждения своей гипотезы о древности некоторого фермента авторы работы укоренили дерево так, чтобы корневая ветвь разделяла археи и эубактерии.

Однако такое укоренение ничем не обосновано!

Укоренение

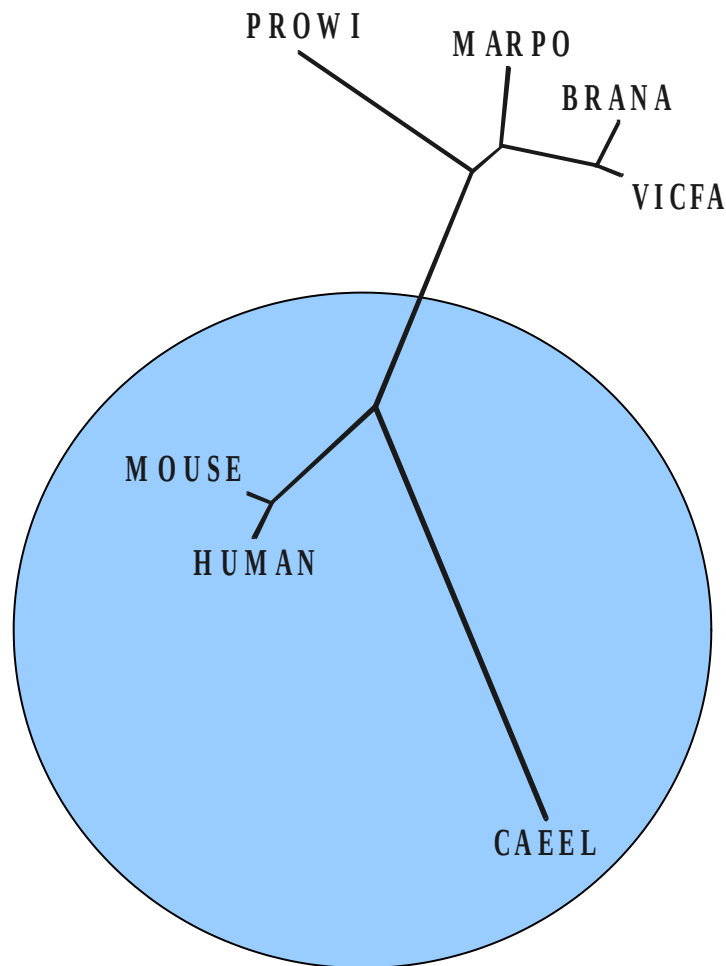
- В среднюю точку:

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



Укоренение

- Используя внешнюю группу («аутгруппу», outgroup):



В данном случае укоренено дерево четырёх растений, для чего пришлось построить дерево с участием внешней группы — трёх животных (в синем круге)

BOOTSTRAP

Bootstrap



Из Википедии

[https://ru.wikipedia.org/wiki/Бутстрэп_\(статистика\)](https://ru.wikipedia.org/wiki/Бутстрэп_(статистика)) :

Бутстрэп (англ. bootstrap) в статистике — практический компьютерный метод исследования распределения статистик вероятностных распределений, основанный на многократной генерации выборок методом Монте-Карло на базе имеющейся выборки.

Слово происходит от выражения: «To pull oneself over a fence by one's bootstraps.» (дословно — «перебраться через ограду, потянув за ремешки на ботинках»). Для русскоязычных людей ближе будет история барона Мюнхгаузена, который, потянув себя за волосы, вытащил себя и свою лошадь из болота.

Сам англицизм «бутстрэп» используется во многих областях знаний, где нужно передать смысл того, что вы получаете что-то «бесплатно» или магическим образом из ничего получаете нечто стоящее. В области статистики ближайший по этимологии аналог термина — «самовытягивание».

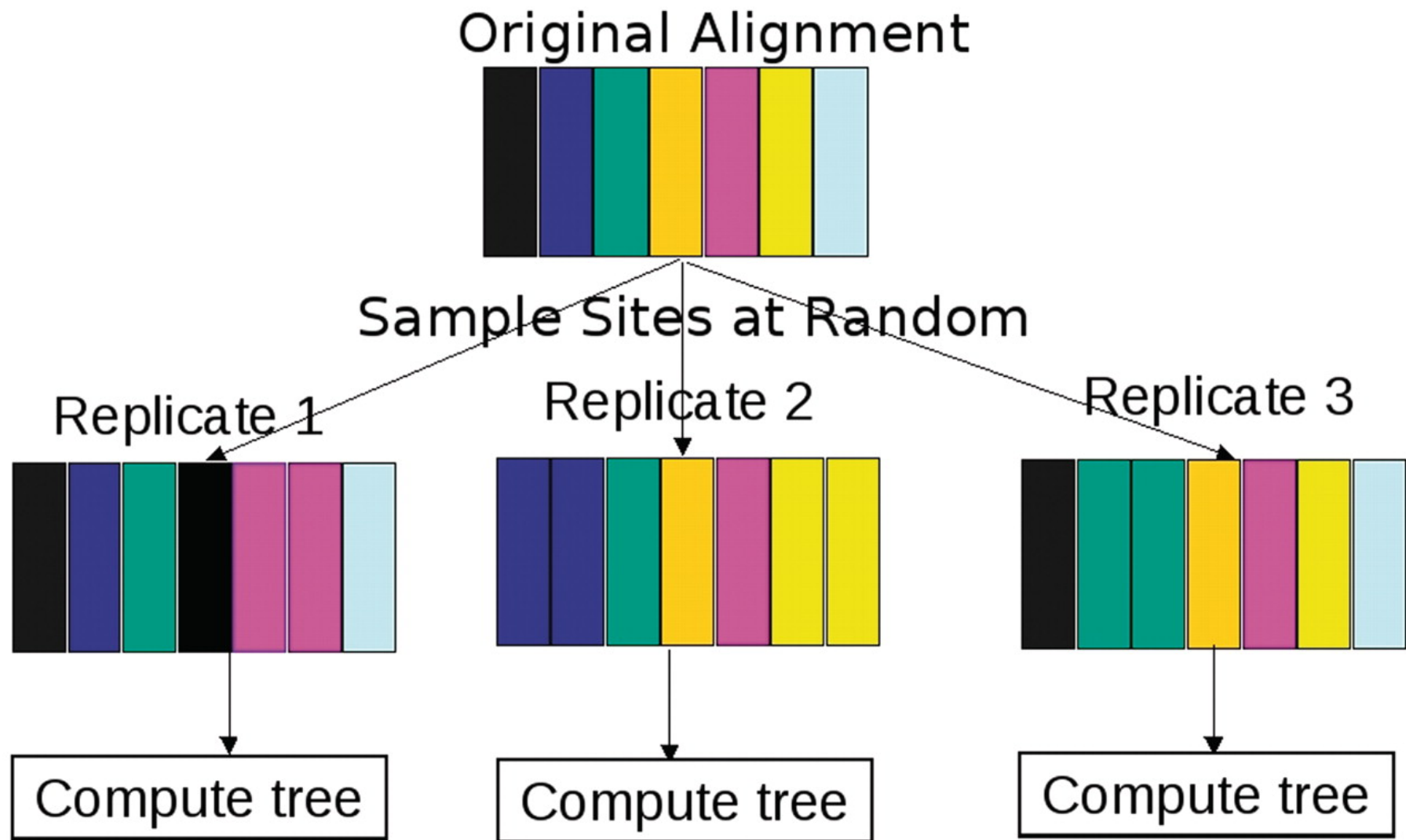
Бутстрэп-анализ в филогении

Из входного выравнивания делается много (например, 100) так наз. «бутстрэп-реплик».

Каждая бутстрэп-реплика получается случайным выбором столбцов исходного выравнивания, в количестве, равном длине выравнивания. В результате часть столбцов не будет выбрана ни разу, часть — один раз, часть — два раза и т.д.

Смысл в том, чтобы построить дерево по части данных и затем сравнить результаты от по разному выбранных частей.

Outline of the phylogenetic bootstrap procedure.



Бутстрэп-анализ

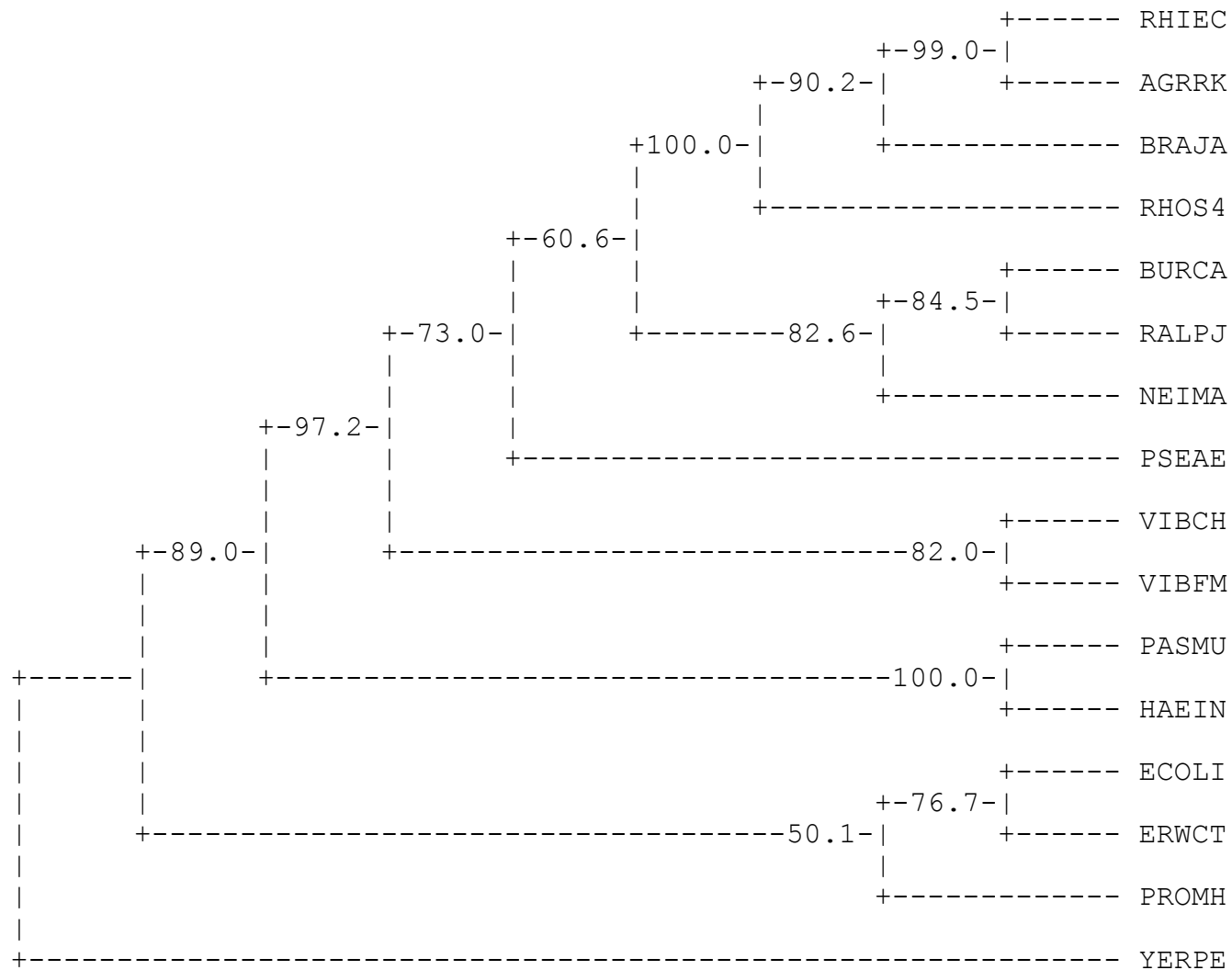
- создаём из входного выравнивания 100 бутстрэп-реплик;
- для каждой из реплик строим по дереву;
- из 100 деревьев строим дерево по методу расширенного большинства («Extended majority-rule tree»).

Качество реконструкции часто возрастает.

Но главное, есть возможность оценить достоверность каждой ветви по т.н. «бутстрэп-поддержке», то есть проценту деревьев, в которых встретилась данная ветвь.

Бутстрэп-анализ

(пример результата)



АНАЛИЗ ДЕРЕВА, СОДЕРЖАЩЕГО ПАРАЛОГИ

Ортологи и паралоги

Белки-гомологи: белки, имеющие общее происхождение.

Среди гомологов выделяют **ортологи** и **паралоги**

- Белки-ортологи разошлись в результате видообразования.
- Белки-паралоги:
 - а) разошлись в результате генной дупликации;
 - б) присутствуют в одном и том же организме.

Вообще-то из б) следует а), ну да ладно :)

Анализ дерева группы гомологичных белков помогает выявить ортологи и реконструировать эволюционные события: видообразования и генной дупликации.

Пример дерева с паралогами

