Обзор методов реконструкции деревьев.

Сравнение деревьев. Приёмы "бутстрэп" и "складной нож".

Сергей Спирин

2 октября 2014 г.

Династия

Фонд Дмитрия Зимина

ИЛС

ИнтерЛабСервис

Ультраметричность и аддитивность расстояний

Пусть дан набор объектов A, B, C, … и для каждой пары объектов — число d, называемое **расстояние** и обладающее свойствами:
1) $d(A,B) > 0$, если $A \neq B$; $d(A,A) = 0$;
2) $d(A,B) = d(B,A)$
3) $d(A,C) \leq d(A,C) + d(B,C)$

Расстояния называются ультраметрическими, если из трёх чисел d(A,B), d(A,C) и d(B,C) два обязательно равны между собой и ≥ третьему.

Ультраметрические расстояния всегда можно реализовать как расстояния по дереву, причём на таком дереве обязательно найдётся точка, равноудалённая от всех объектов.

Расстояния называются аддитивными, если из трёх чисел d(A,B)+d(C,D), d(A,C)+d(B,D) и d(B,C)+d(A,D) два обязательно равны между собой и ≥ третьему.

Аддитивные расстояния всегда можно реализовать как расстояния по некоторому дереву, причём это дерево единственное.

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:

- бывают дупликации белков без разделения видов:

два родственных белка существуют в одном геноме и эволюционируют (почти) независимо – такие белки называются **паралогами.**

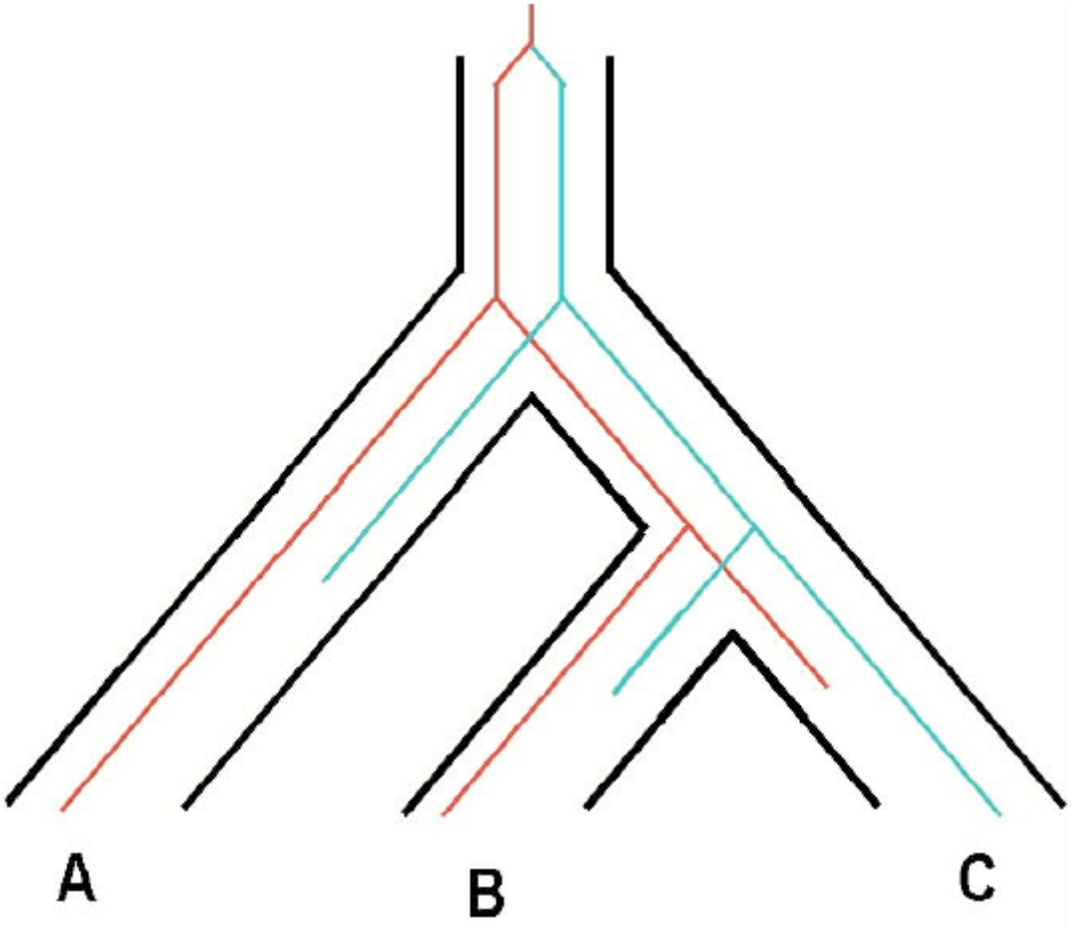
- бывают потери генов.

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.

- бывает, что два белка объединяются в один многодоменный, и наоборот.

Поэтому правильнее говорить об эволюции белковых доменов.

Дерево видов и дерево белков



Исходный материал для реконструкции филогении

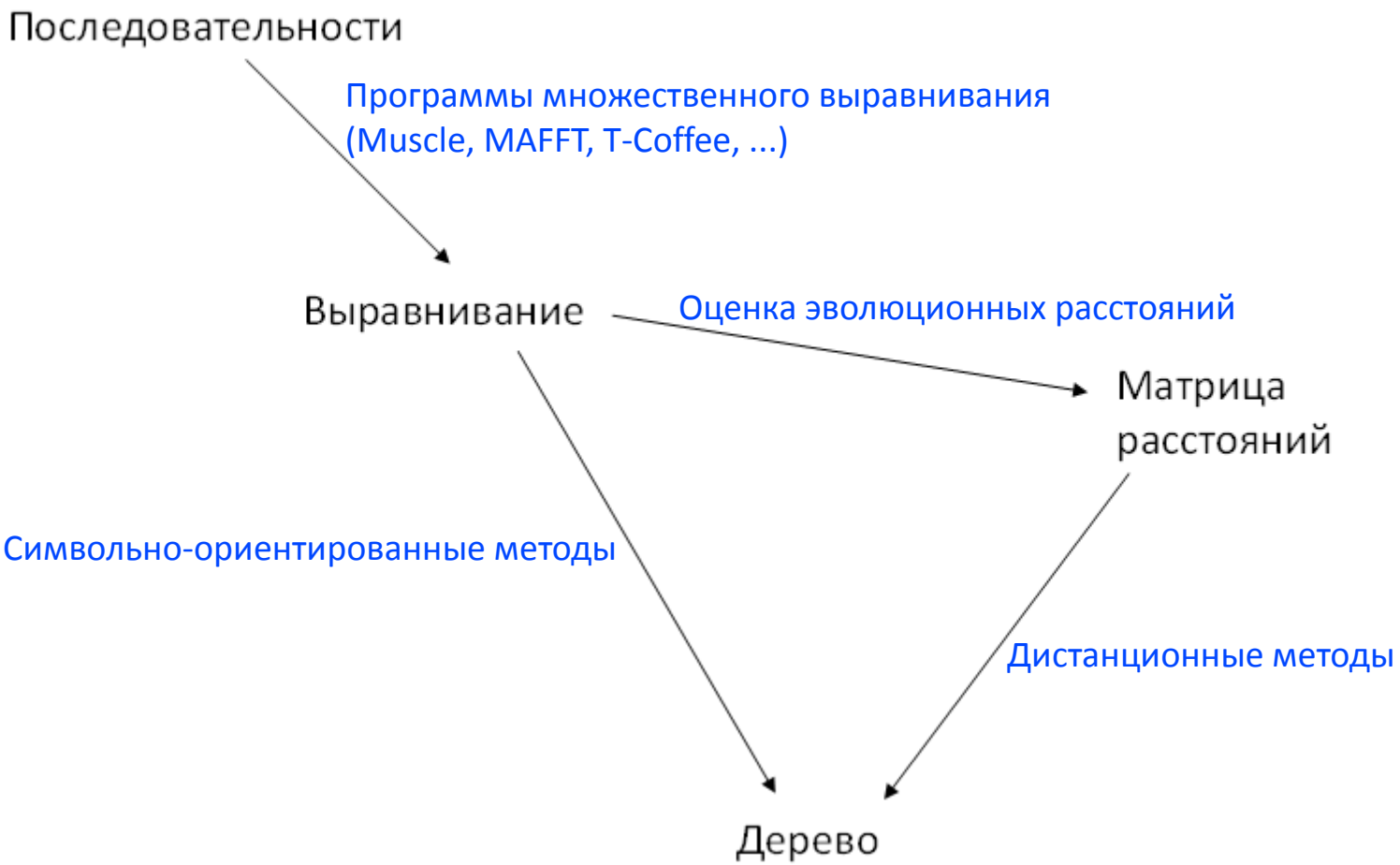
Множественное выравнивание биологических последовательностей

```
CYB5_CHICK      MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
CYB5_HUMAN      ---MAEQSDEAV--KYYTLEEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQA
CYB5_HORSE      ---MAEQSDKAV--KYYTLEEIKKHNHSKSTWLILHHKVYDLTKFLEDHPGGEEVLREQA
CYB5_MUSDO      ------MSSEDV--KYFTRAEVAKNNTKDKNWFIIHNNVYDVTAFLNEHPGGEEVLIEQA
CYB5_DROME      ------MSSEET--KTFTRAEVAKHNTNKDTWLLIHNNIYDVTAFLNEHPGGEEVLIEQA
```

Методы реконструкции филогенетических деревьев:

- символьно ориентированные:
  - * максимальной экономии (maximum parsimony)
  - * наибольшего правдоподобия и байесовские
- дистанционные, т.е. использующие матрицу расстояний
  - * кластерные (UPGMA и др.)
  - * Neighbor-joining
  - * минимальной эволюции
  - * наименьших квадратов
  - * Фитча – Марголиаша
  - * …

Схема реконструкции филогении по последовательностям



Пример матрицы расстояний (выдача программы protdist пакета PHYLIP)

```
   12
YEAST       0.000000  1.226697  1.313711  1.187561  1.143864  1.198065
  1.172131  1.171814  1.178217  1.168880  1.119266  1.136044
TOBAC       1.226697  0.000000  1.130459  1.058881  1.185956  1.085361
  1.168318  1.106069  1.156591  1.173863  1.193820  1.175669
MUSDO       1.313711  1.130459  0.000000  0.281922  0.992650  0.859854
  0.951601  0.863773  0.903579  0.874458  0.922674  0.976793
DROME       1.187561  1.058881  0.281922  0.000000  0.968555  0.881378
  0.935269  0.858443  0.913545  0.884540  0.939652  0.957328
CHICK       1.143864  1.185956  0.992650  0.968555  0.000000  0.346056
  0.281003  0.274432  0.288135  0.283946  0.272658  0.306560
BOVIN       1.198065  1.085361  0.859854  0.881378  0.346056  0.000000
  0.176332  0.107028  0.183987  0.165630  0.177340  0.166408
HUMAN       1.172131  1.168318  0.951601  0.935269  0.281003  0.176332
  0.000000  0.139340  0.133018  0.124695  0.102465  0.108339
PIG         1.171814  1.106069  0.863773  0.858443  0.274432  0.107028
  0.139340  0.000000  0.105960  0.082471  0.123761  0.090834
MOUSE       1.178217  1.156591  0.903579  0.913545  0.288135  0.183987
  0.133018  0.105960  0.000000  0.021980  0.102164  0.123996
RAT         1.168880  1.173863  0.874458  0.884540  0.283946  0.165630
  0.124695  0.082471  0.021980  0.000000  0.092132  0.099815
RABIT       1.119266  1.193820  0.922674  0.939652  0.272658  0.177340
  0.102465  0.123761  0.102164  0.092132  0.000000  0.090807
HORSE       1.136044  1.175669  0.976793  0.957328  0.306560  0.166408
  0.108339  0.090834  0.123996  0.099815  0.090807  0.000000
```

Как оценивать расстояния между последовательностями?

- Просто как долю различий
- На основе доли различий, но с поправкой на возможность повторных мутаций в одной позиции (например, формула Джукса – Кантора)
- На основе принципа наибольшего правдоподобия

Как оценивать расстояния между последовательностями?

- Просто как долю различий
- На основе доли различий, но с поправкой на возможность повторных мутаций в одной позиции (например, формула Джукса – Кантора)
- **На основе принципа наибольшего правдоподобия**

Для оценки по наибольшему правдоподобию нужна **вероятностная модель эволюции.** На основании такой модели можно ответить на вопрос: какова вероятность, что из данной последовательности 1 произошла данная последовательность 2 в результате ровно *N* мутаций?

Получаем функцию $p(N)$. Максимально правдоподобная оценка расстояния есть точка максимума этой функции – то значение *N*, для которого вероятность принимает наибольшее значение.

Оценка — всего лишь оценка!

Не стоит ожидать, что полученные расстояния будут в точности обладать свойством аддитивности.

Если свойство аддитивности не выполняется, невозможно построить дерево, расстояния по которому будут в точности совпадать с посчитанными расстояниями. Остаётся каким-то образом «подогнать» дерево под расстояния. В этом причина большого числа дистанционных методов.

Классификация методов

Название	Переборный / эвристический	Предполагает молекулярные часы	Символьно ориентированный
UPGMA	Эвристический	Да	Нет
Neibour-joining	Эвристический	Нет	Нет
Наименьших квадратов	Переборный	Может	Нет
Фитча – Марголиаша	Переборный	Может	Нет
Минимальной эволюции	Переборный	Может	Нет
Максимальной экономии	Переборный	Нет	Да
Наибольшего правдоподобия	Переборный	Может	Да

Переборные методы

Алгоритм, реализующий переборный метод, должен включать:

а) критерий сравнения деревьев (какая из двух топологий лучше соответствует исходным данным?)

б) алгоритм поиска лучшего по критерию дерева.

Пример критерия
(метод наименьших квадратов, OLS — ordinary least squares)

Пусть дана матрица расстояний и топология дерева;
i, j — две последовательности, тогда мы имеем расстояние $d(i,j)$ из матрицы, и, приписав ветвям длину, будем иметь расстояние $d'(i,j)$ по дереву.
Подберём длины ветвей так, чтобы сумма $(d(i,j) - d'(i,j))^2$ (по всем парам i,j) была наименьшей.
Это наименьшее значение и будет критерием качества — будем считать ту топологию лучшей, для которой это значение получится меньшим.

Переборные методы

Алгоритм, реализующий переборный метод, должен включать:

а) критерий сравнения деревьев (какая из двух топологий лучше соответствует исходным данным?)

б) алгоритм поиска лучшего по критерию дерева.

Название метода совпадает с названием **критерия**.

Критерии бывают символьно-ориентированные: наибольшее правдоподобие и максимальная экономия, и дистанционные: разные варианты критерия минимальной эволюции, наименьшие квадраты (обычные и “улучшенные”), квартетный критерий и др.

Алгоритмы поиска считаются скорее параметрами программы.

Переборные методы

Алгоритм, реализующий переборный метод, должен включать:

а) критерий сравнения деревьев (какая из двух топологий лучше соответствует исходным данным?)

б) алгоритм поиска лучшего по критерию дерева.

Все критерии, кроме максимальной экономии, подразумевают “по ходу дела” вычисление длин ветвей, поэтому и в ответе получается не только топология, но и длины ветвей.

Поиск лучшего дерева

Имеется единственная топология (неукоренённого) дерева с тремя листьями, три разных топологии деревьев с четырьмя листьями,
15 топологий деревьев с пятью листьями,
… … … …
~ 2 млн. топологий деревьев с десятью листьями,
… … … …
~ 8 трлн. топологий деревьев с 15 листьями
… … … …
Триллионы проверок компьютер будет делать слишком долго.
А ведь приходится строить деревья и с сотней листьев...

Поиск лучшего дерева

Все деревья перебрать, как правило, нельзя — число различных деревьев с N листьями равно

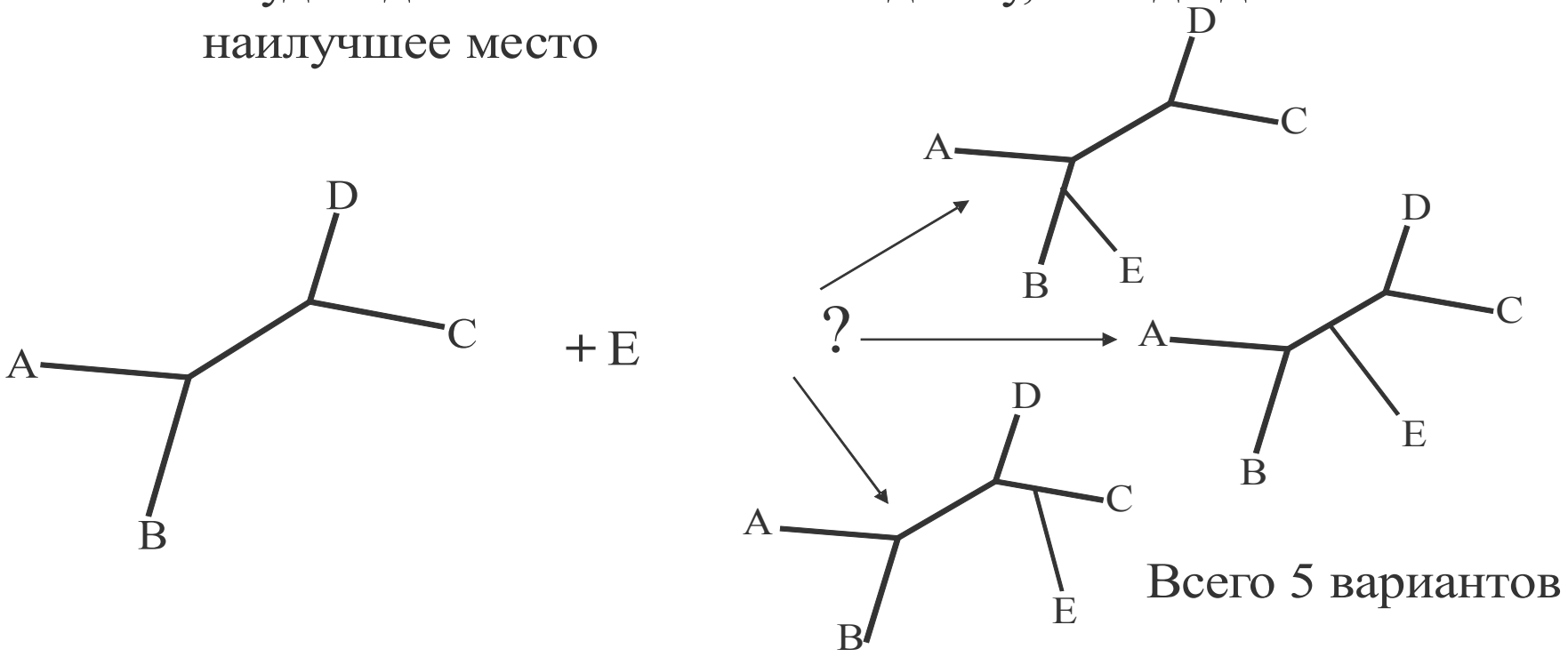
$$(2N - 5)!! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2N - 7) \cdot (2N - 5)$$

Это число очень быстро растёт. Полный перебор возможен, если число листьев не превышает 10–12.

Поэтому применяются (вместе и по отдельности) два приёма — “выращивание” и “жадный” поиск по соседним деревьям.

Поиск лучшего дерева: выращивание

- Найдем лучшее дерево для части последовательностей
- Будем добавлять листья по одному, находя для них наилучшее место



Всего 5 вариантов

Поиск лучшего дерева: выращивание

Разрешённое дерево с N листьями всегда имеет $2N–3$ ветви. Поэтому чтобы вырастить дерево с N листьями, надо проанализировать $3 + 5 + … + 2N–5 = (N–3)(N–5)$ деревьев.
Уже для $N=10$ это число меньше числа всех возможных деревьев в 32175 раз!

Выращивание не гарантирует нахождение лучшего по критерию дерева, но при хороших данных не должно приводить к большим ошибкам.

Поиск лучшего дерева: просмотр соседних деревьев

Построим сначала черновое дерево, а потом попробуем его улучшить.

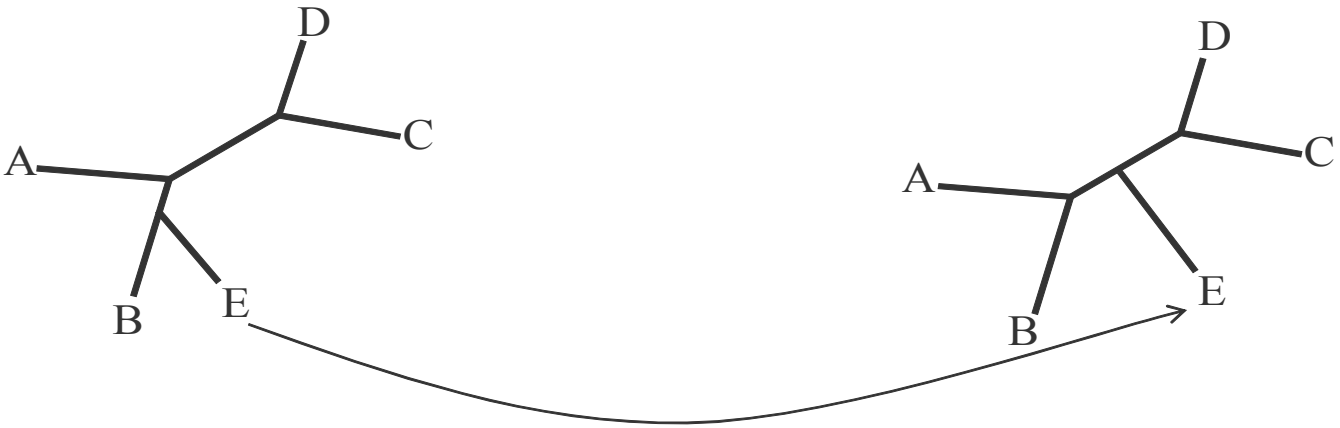
Черновое дерево можно построить одним из эвристических методов или “вырастить”.

Улучшать будем, просматривая соседние деревья — если какое-нибудь соседнее дерево лучше данного, то берём его и просматриваем его соседей, и так пока не получим дерево, превосходящее по нашему критерию всех своих соседей.

Поиск лучшего дерева: просмотр соседних деревьев

Что такое соседние деревья?

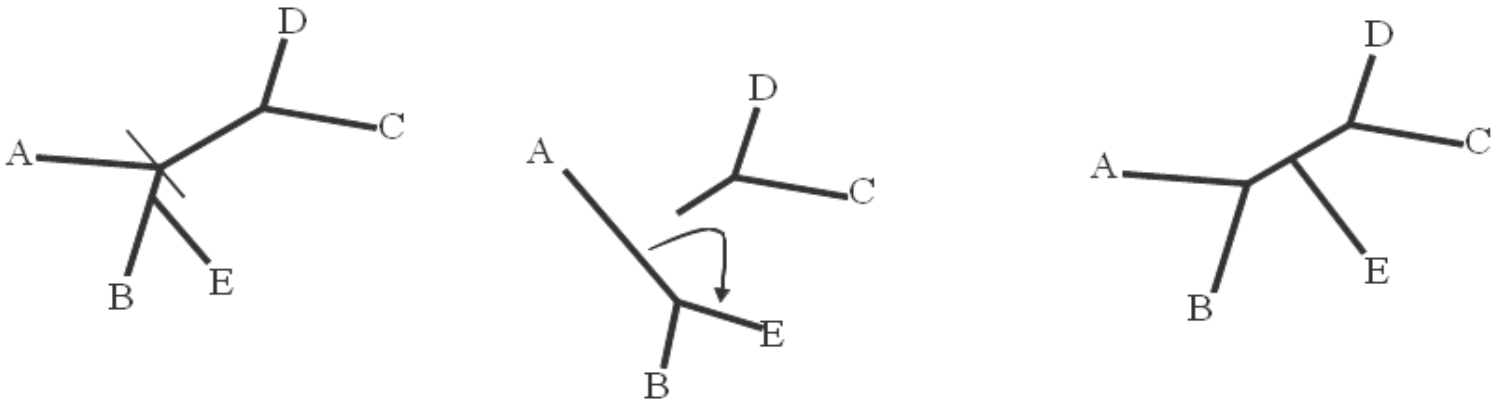
- Оторвём один лист и «привьём» его на другую ветвь



Поиск лучшего дерева: просмотр соседних деревьев

Что такое соседние деревья?

- Можно проделать аналогичную операцию с целой кладой



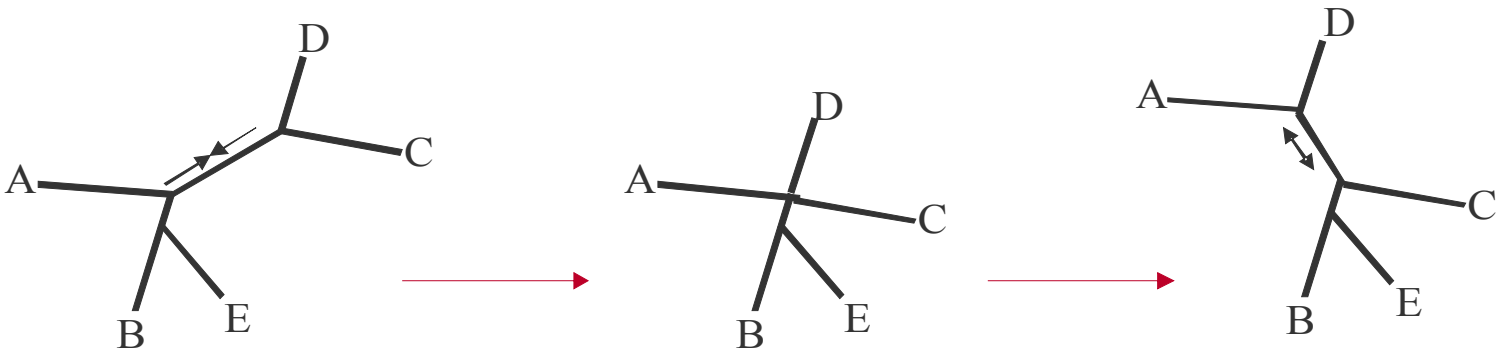
Такая операция обычно называется **SPR** : Subtree Pruning and Regrafting.

В пакете PHYLIP она называется “Global rearrangement”.

Поиск лучшего дерева: просмотр соседних деревьев

Что такое соседние деревья?

- Можно «схлопнуть» одну ветвь и заменить её другой



Такая операция обычно называется **NNI** : Nearest neigbour interchange.

В пакете PHYLIP она называется “Local rearrangement”.

Поиск лучшего дерева: алгоритм

1. Строим черновое дерево
2. Анализируем соседние деревья
3. Если нашлось лучшее, берём за основу его и повторяем п.2
4. Если не нашлось лучшего, выдаём текущее дерево как результат

Параметры алгоритма: способ построения чернового дерева (эвристический метод, выращивание, случайное дерево) и что понимается под соседними деревьями (NNI, SPR, …).
Ну и, конечно, критерий определения лучшего дерева.

Переборные методы

1. Максимальной экономии (или “бережливости”, или даже “парсимонии”, maximum parsimony, MP)
   критерий — наименьшее число мутаций, позволяющее по данному дереву получить данные последовательности
2. Наибольшего правдоподобия (maximum likelihood, ML)
   критерий — вероятность получить данные последовательности по данному дереву
3. Дистанционные
   i. ordinary least squares (OLS)
      критерий — среднее квадратичное отклонение расстояний
   ii. Fitch – Margoliash
      то же, что OLS, но используются относительные отклонения
   iii. Minimum evolution
      критерий — суммарная длина ветвей (сами длины оцениваются по-разному)

Эвристические методы

1. UPGMA = Unweighted Pair Group Method with Arithmetic mean
   Строит укоренённое ультраметрическое дерево.
   Видимо, реально лучший из методов, предполагающих молекулярные часы.
2. Neighbor-joining (NJ)
   строит неукоренённое дерево. Если и уступает некоторым переборным алгоритмам, то не сильно.

UPGMA – схема алгоритма

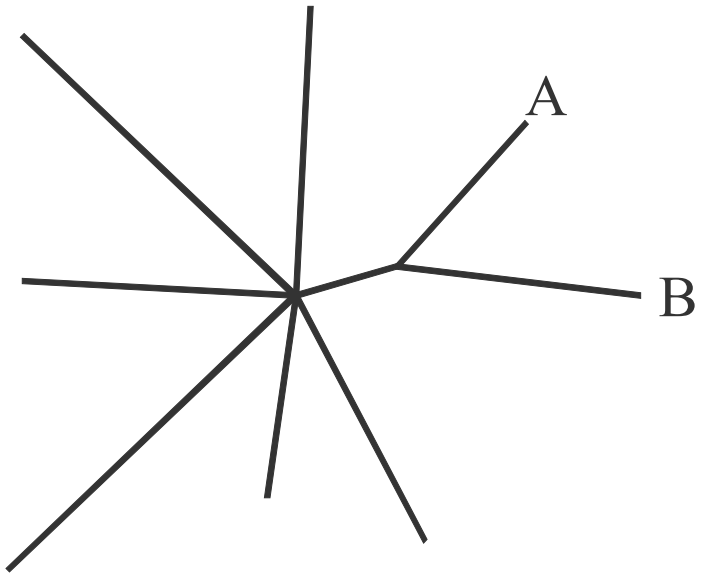
Укоренённое дерево строится «снизу вверх»

- Найдём в матрице расстояний наименьший элемент.
- Объединим два ближайших листа в кластер (это — узел дерева, соединённый ветвями с листьями, образовавшими его).
- Пересчитаем матрицу расстояний, рассматривая кластер как новый лист. Расстоянием до кластера будем считать **среднее арифметическое** расстояний до его элементов (отсюда название метода).
- Повторяем с начала, пока не останется всего два кластера.

К этому прибавляется способ вычисления длин ветвей. Результат — укоренённое ультраметрическое дерево с длинами ветвей.

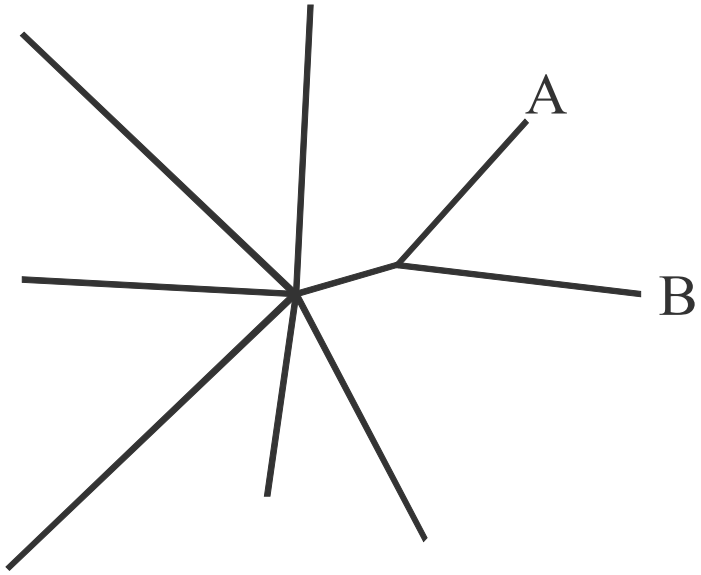
Neighbor-joining

Выбираем пару последовательностей A,B , для которых наименьшее значение имеет величина $(n-2)d(A,B) - s(A) - s(B)$,
где d — расстояние из входной матрицы, n — число последовательностей, а $s(A)$ — сумма расстояний от A до всех остальных последовательностей.
Объединяем пару в кластер, с которым далее обращаемся как с одной последовательностью.



Neighbor-joining (продолжение)

Повторяем объединение, пока не останется три кластера.



В отличие от UPGMA, даже при ультраметрической матрице «соседи» не обязательно объединяются снизу вверх!
Полученное методом Neighbor-joining дерево — неукоренённое!
Метод «по ходу дела» оценивает длины ветвей.
Хотя эти длины иногда получаются отрицательными! :(

Какой метод лучше?

Gonnet *BMC Bioinformatics* 2012, **13**:148
http://www.biomedcentral.com/1471-2105/13/148

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Surprising results on phylogenetic tree building methods based on molecular sequences

Gaston H Gonnet

Abstract

Background: We analyze phylogenetic tree building methods from molecular sequences (PTMS). These are methods which base their construction solely on sequences, coding DNA or amino acids.

Results: Our first result is a statistically significant evaluation of 176 PTMSs done by comparing trees derived from 193138 orthologous groups of proteins using a new measure of quality between trees. This new measure, called the Intra measure, is very consistent between different groups of species and strong in the sense that it separates the methods with high confidence.
The second result is the comparison of the trees against trees derived from accepted taxonomies, the Taxon measure. We consider the NCBI taxonomic classification and their derived topologies as the most accepted biological consensus on phylogenies, which are also available in electronic form. The correlation between the two measures is remarkably high, which supports both measures simultaneously.

Conclusions: The big surprise of the evaluation is that the maximum likelihood methods do not score well, minimal evolution distance methods over MSA-induced alignments score consistently better. This comparison also allows us to rank different components of the tree building methods, like MSAs, substitution matrices, ML tree builders, distance methods, etc. It is also clear that there is a difference between Metazoa and the rest, which points out to evolution leaving different molecular traces. We also think that these measures of quality of trees will motivate the design of new PTMSs as it is now easier to evaluate them with certainty.

Keywords: Phylogenetic trees, Tree building methods, Maximum likelihood, Distance measures, Multiple sequence alignments, Substitution matrices, Molecular sequences

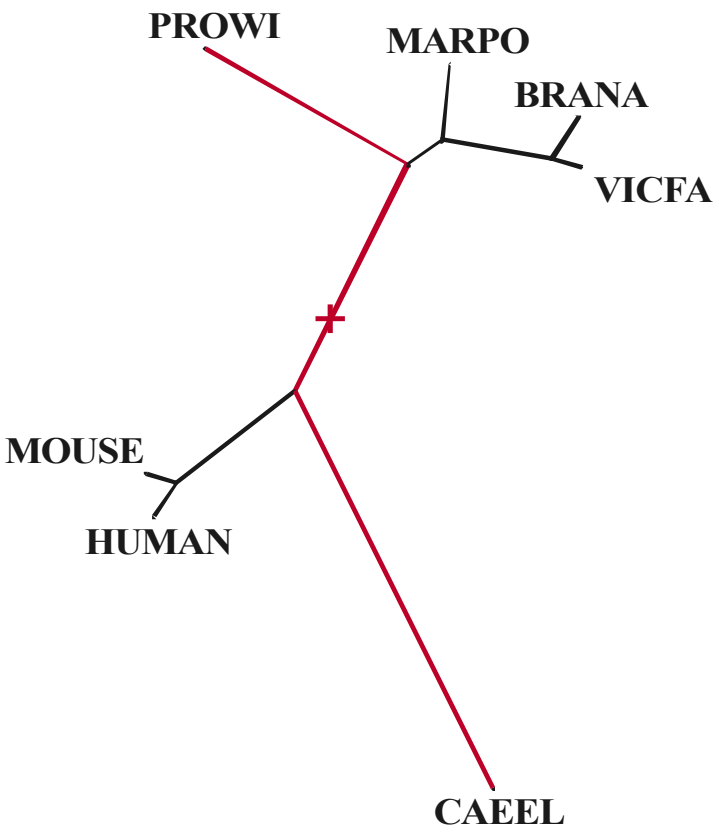
Background

organisms [3], tracing disease [4], finding vectors [5],

Укоренение

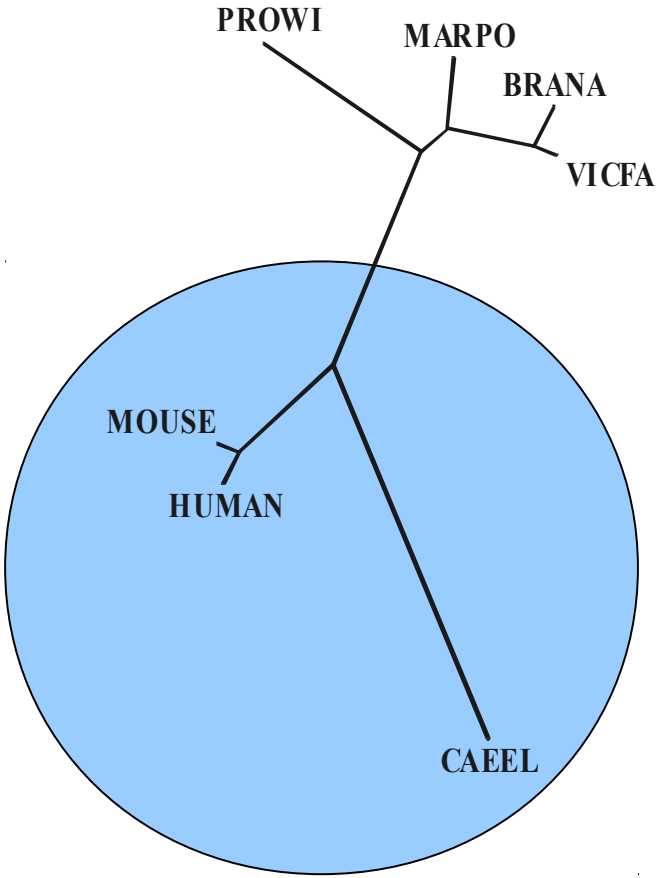
- В среднюю точку:

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)

Сравнение деревьев

- Консенсусное (небинарное) дерево
- Максимальное общее поддерево
- Дерево из ветвей, поддержанных большинством (majority-rule tree)
- Меры сходства деревьев ("расстояние")

  i. Доля общих ветвей
  ii. Расстояние в "пространстве ветвей"
  iii. Доля общих четверок
  iv. Длина пути в пространстве деревьев

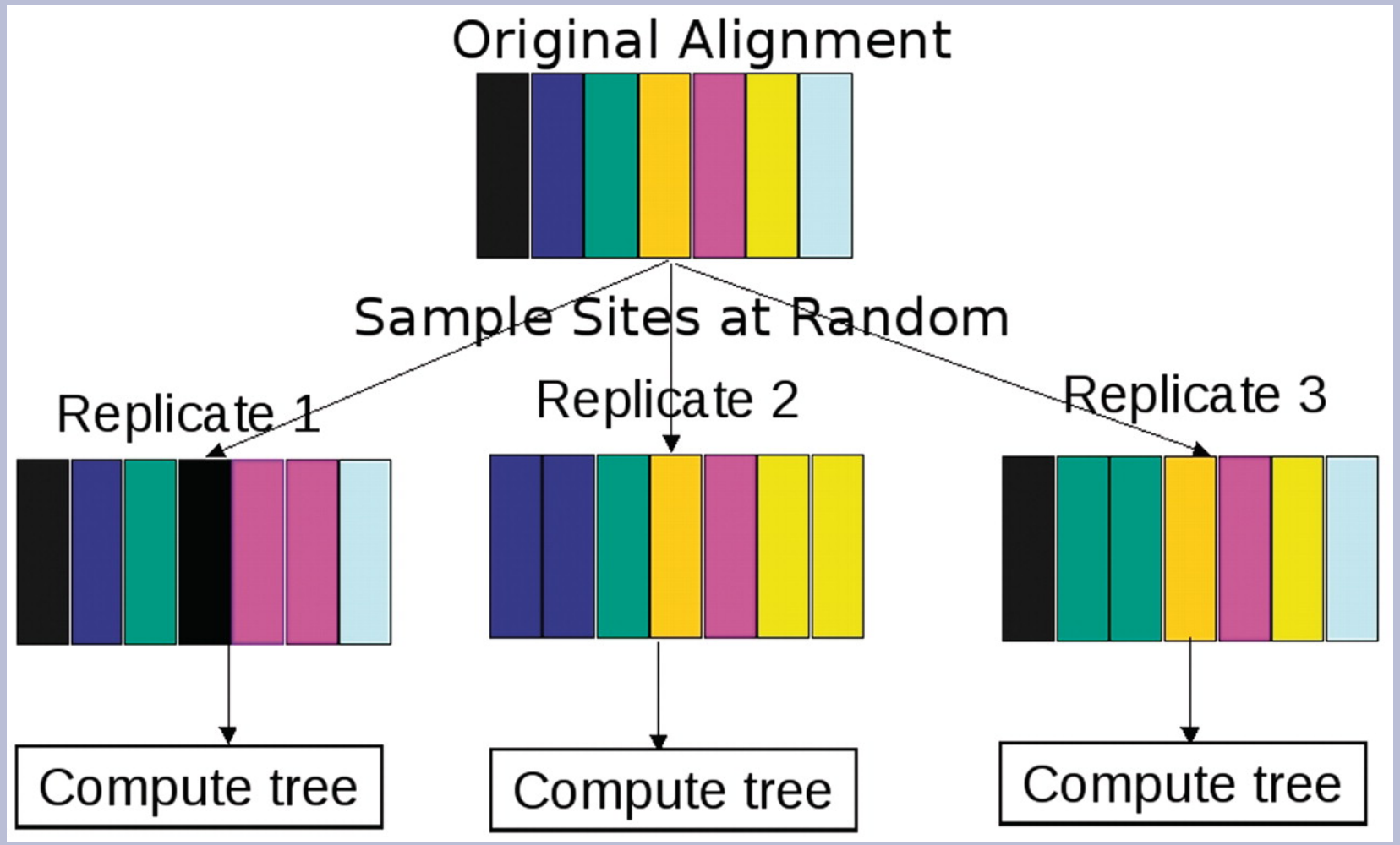
Бутстрэп-анализ

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

Каждая бутстрэп-реплика получается в результате случайного удаления половины столбцов из выравнивания с заменой их копиями других (тоже случайно выбранных) столбцов.

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

Outline of the phylogenetic bootstrap procedure.



Stamatakis A , Izquierdo-Carrasco F Brief Bioinform 2011;12:270-279

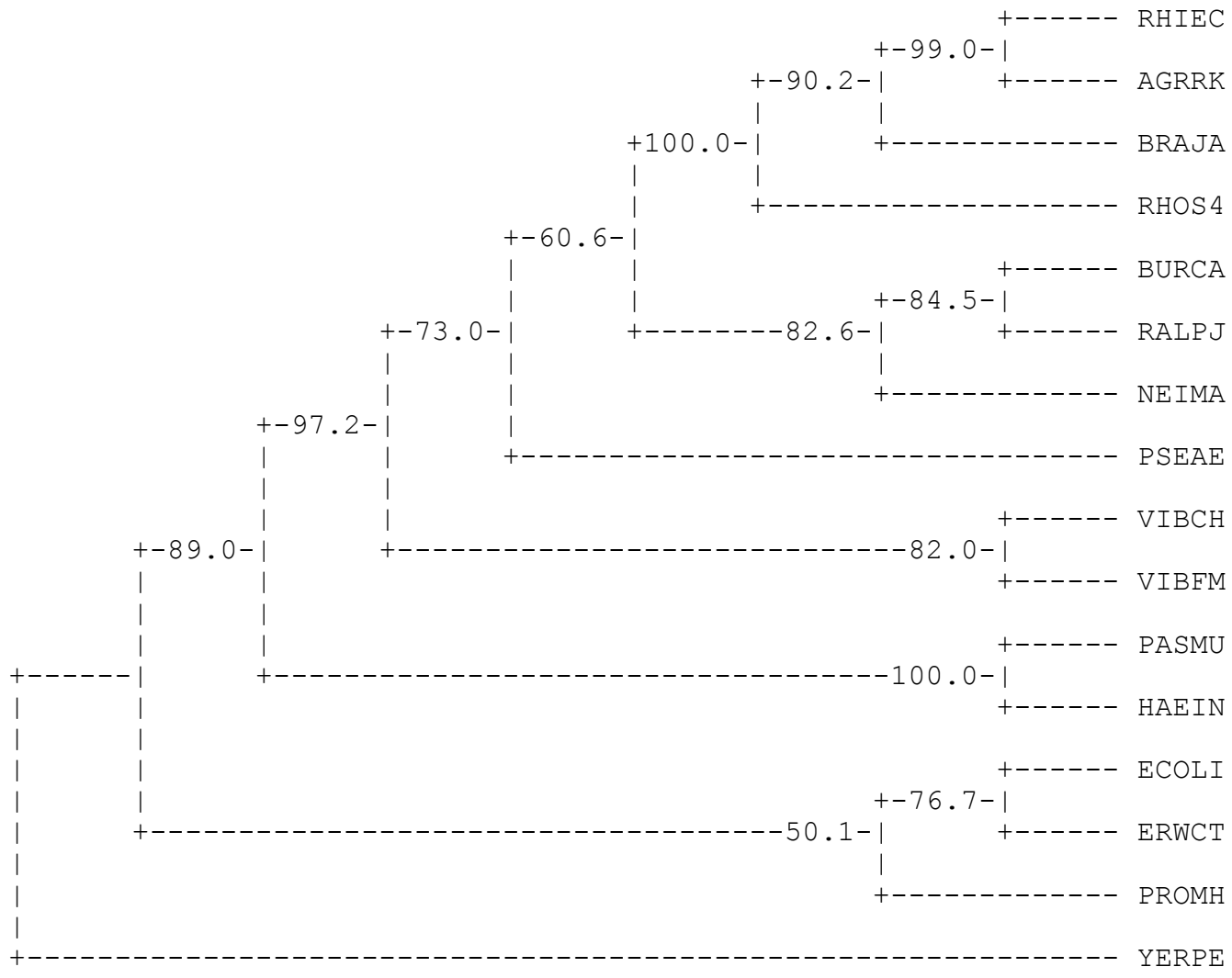
Briefings in Bioinformatics

Бутстрэп

- 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»).

  Помимо того, что (как правило) возрастает качество реконструкции, есть возможность оценить достоверность каждой ветви по т.н. «бутстрэп-поддержке», то есть проценту деревьев, в которых встретилась данная ветвь.

Бутстрэп (пример результата)



Программы

- JalView

http://www.jalview.org/

Работа с выравниваниями.

По невыровненным последовательностям реконструировать эволюцию нельзя!

Может строить деревья (методы UPGMA и Neighbor-Joining)

- MEGA

http://www.megasoftware.net/

Визуализация и построение деревьев (оконный интерфейс)

- Пакет PHYLIP

http://evolution.genetics.washington.edu/phylip.html

Построение и визуализация деревьев (интерактивный интерфейс из командной строки). Имеется веб-интерфейс на портале http://bioweb2.pasteur.fr/

Пакет PHYLIP

- Реализация методов UPGMA и Neighbor-Joining (программа *neighbor*), наименьших квадратов и Фитча – Марголиаша (*fitch* и *kitsch*), максимальной экономии (*dnapars* и *protpars*), наибольшего правдоподобия (*dnaml*, *dnamlk*, *proml*, *promlk*)
- Оценка эволюционных расстояний: программы *dnadist* и *protdist*
- Сравнение деревьев: *consense*, *treedist*, *treedistpair*
- Редактура (включая укоренение в среднюю точку): *retree*
- Бутстрэп: *seqboot*
- Визуализация: *drawtree*, *drawgram*

Что важно помнить при реконструкции филогении по последовательностям

1. Последовательности должны быть гомологичны по всей длине.

2. Если последовательности нуклеотидные, надо убедиться, что часть из них не представлена комплементарными вариантами.

3. Последовательности необходимо выровнять.

Кстати, по виду выравнивания можно оценить, действительно ли последовательности такие, как надо: должно быть много консервативных колонок и мало хаотично расположенных гэпов. Помните: программа выравнивания выдаст результат даже для совершенно неродственных последовательностей, но смысла в этом результате не будет!

4. Большинство программ реконструируют неукоренённое дерево (даже если оно выглядит как укоренённое). Определение положения корня – отдельная задача.

5. Результат реконструкции – не абсолютная истина. Достоверность той или иной ветви можно оценить путём сравнения результатов разных программ и/или бутстрепа.

Как правило, чем короче выравнивание, тем хуже качество реконструкции.