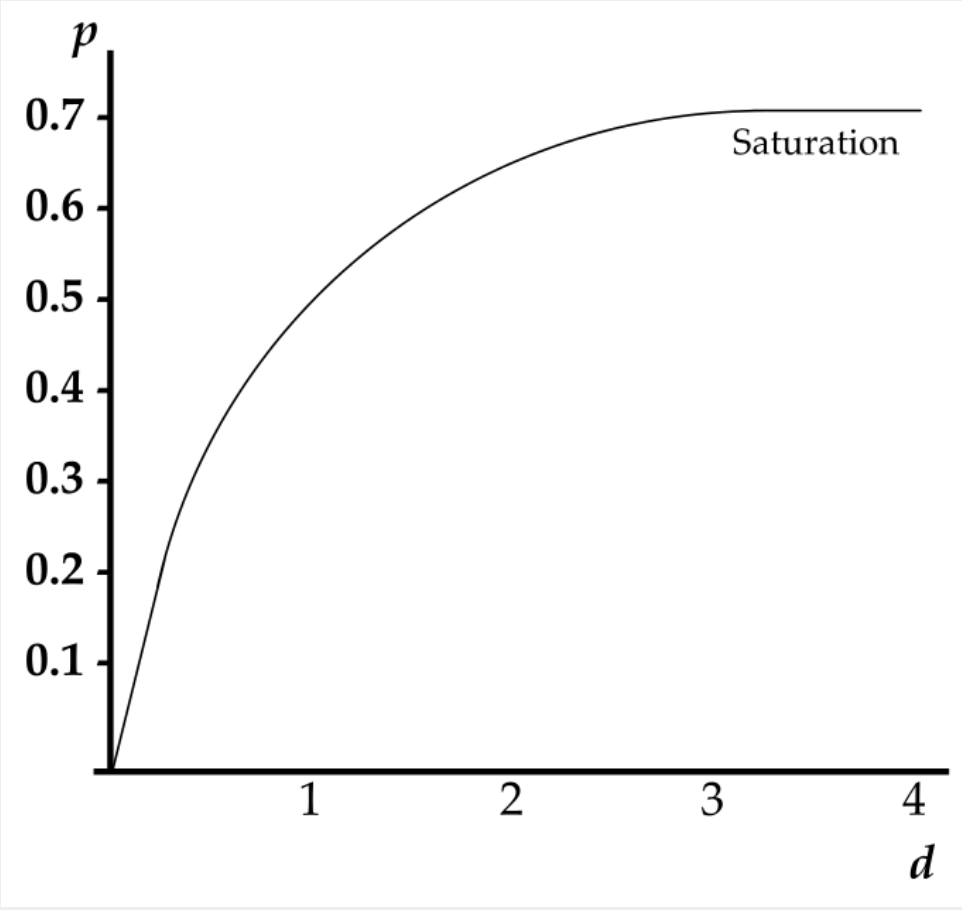
Фильтрация выравнивания перед построением дерева

1. Gblocks: популярный, (но) давно не обновлённый и консервативный
2. trimAl: можно несколько выравниваний, лучше для большого объёма данных
3. GUIDANCE2: только веб-форма, сам выравнивает.
4. Aliscore: локальное качество, счёт для каждой позиции.
5. AL2CO: только белки, чаще используют для оценки консервативности, а не для фильтрации.
6. Zorro aka Probmask: тоже оценивает каждую позицию, более сложная модель оценки.

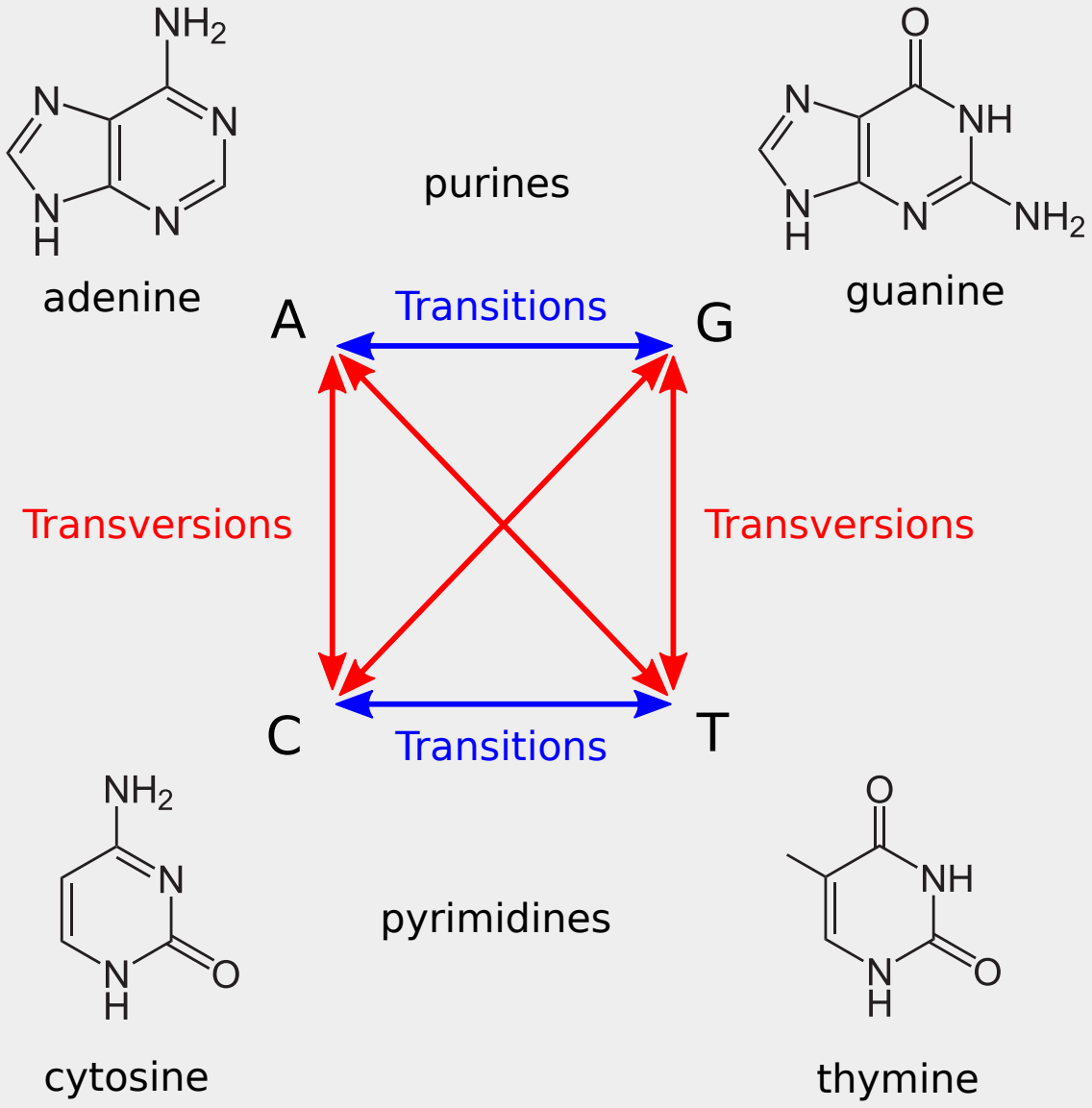
Определение расстояния между последовательностями

- p-distance (наблюдаемое расстояние)



Lemey P., Salemi M., Vandamme A-M. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press; 2009.

Модели эволюции нуклеотидов



https://en.wikipedia.org/wiki/Transversion

Модели эволюции нуклеотидов

Автор(ы), название	Частоты нуклеотидов	Частоты переходов	Свободных параметров
Jukes-Cantor (JC69)	равные	равные	0

Модели эволюции нуклеотидов

Автор(ы), название	Частоты нуклеотидов	Частоты переходов	Свободных параметров
Jukes-Cantor (JC69)	равные	равные	0
Kimura (K80=K2P)	равные	$Ts \neq Tv$	1

Модели эволюции нуклеотидов

Автор(ы), название	Частоты нуклеотидов	Частоты переходов	Свободных параметров
Jukes-Cantor (JC69)	равные	равные	0
Kimura (K80=K2P)	равные	$Ts \neq Tv$	1
Felsenstein (F81)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	равные	3

Модели эволюции нуклеотидов

Автор(ы), название	Частоты нуклеотидов	Частоты переходов	Свободных параметров
Jukes-Cantor (JC69)	равные	равные	0
Kimura (K80=K2P)	равные	$Ts \neq Tv$	1
Felsenstein (F81)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	равные	3
Felsenstein (F84) & Hasegawa-Kishono-Yano (HKY85)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts \neq Tv$	4

Модели эволюции нуклеотидов

Автор(ы), название	Частоты нуклеотидов	Частоты переходов	Свободных параметров
Jukes-Cantor (JC69)	равные	равные	0
Kimura (K80=K2P)	равные	$Ts \neq Tv$	1
Felsenstein (F81)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	равные	3
Felsenstein (F84) & Hasegawa-Kishono-Yano (HKY85)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts \neq Tv$	4
Tamura-Nei (TN93)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts_R \neq Ts_Y \neq Tv$	5

Модели эволюции нуклеотидов

Автор(ы), название	Частоты нуклеотидов	Частоты переходов	Свободных параметров
Jukes-Cantor (JC69)	равные	равные	0
Kimura (K80=K2P)	равные	$Ts \neq Tv$	1
Felsenstein (F81)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	равные	3
Felsenstein (F84) & Hasegawa-Kishono-Yano (HKY85)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts \neq Tv$	4
Tamura-Nei (TN93)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts_R \neq Ts_Y \neq Tv$	5
Tavaré (GTR)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	6 переходов	8

Модели эволюции нуклеотидов

Автор(ы), название	Частоты нуклеотидов	Частоты переходов	Свободных параметров
Jukes-Cantor (JC69)	равные	равные	0
Kimura (K80=K2P)	равные	$Ts \neq Tv$	1
Felsenstein (F81)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	равные	3
Felsenstein (F84) & Hasegawa-Kishono-Yano (HKY85)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts \neq Tv$	4
Tamura-Nei (TN93)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	$Ts_R \neq Ts_Y \neq Tv$	5
Tavaré (GTR)	$\pi_A \neq \pi_C \neq \pi_T \neq \pi_G$	6 переходов	8

jModelTest2

Классификация методов

1. Методы расстояний:
   - (невзвешенного попарного среднего, UPGMA);
   - присоединения ближайшего соседа, NJ;
   - (наименьших квадратов, LS);
   - (минимальной эволюции, ME).
2. Дискретные методы:
   - максимальной парсимонии, MP;
   - максимального правдоподобия, ML;
   - Байесовский подход, BI.

Методы расстояний: Unweighted pair group method with arithmetic mean (UPGMA)

1. Находим два самых близких значения в матрице.
2. Объединяем соответствующие узлы в группу.
3. Перерасчитываем расстояния до остальных узлов как среднее расстояние до первого и второго членов группы.
4. Строим новую матрицу (-1 строка, -1 столбец).
5. Если в матрице >1 значения, см. п. 1.

	A	B	C	D	E
A	-				
B	2	-			
C	9	9	-		
D	10	10	3	-	
E	12	12	13	13	-

Методы расстояний: UPGMA

1. Находим два самых близких значения в матрице.
2. Объединяем соответствующие узлы в группу.
3. Перерасчитываем расстояния до остальных узлов как среднее расстояние до первого и второго членов группы.
4. Строим новую матрицу (-1 строка, -1 столбец).
5. Если в матрице >1 значения, см. п. 1.

	AB	C	D	E
AB	-			
C	9	-		
D	10	3	-	
E	12	13	13	-

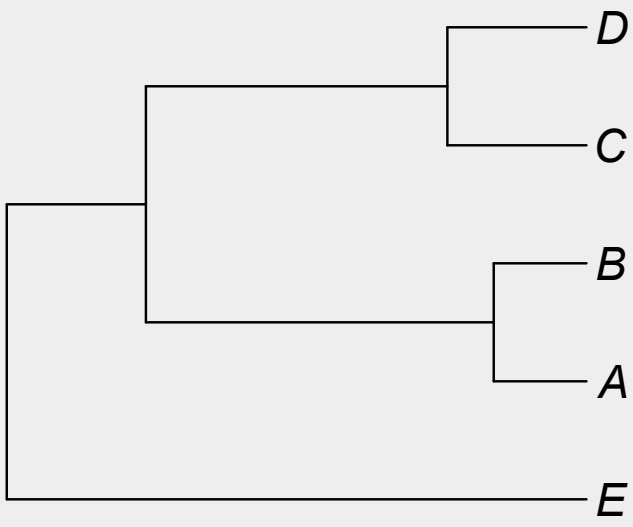
Методы расстояний: UPGMA

1. Находим два самых близких значения в матрице.
2. Объединяем соответствующие узлы в группу.
3. Перерасчитываем расстояния до остальных узлов как среднее расстояние до первого и второго членов группы.
4. Строим новую матрицу (-1 строка, -1 столбец).
5. Если в матрице >1 значения, см. п. 1.

	AB	CD	E
AB	-		
CD	9.5	-	
E	12	13	-

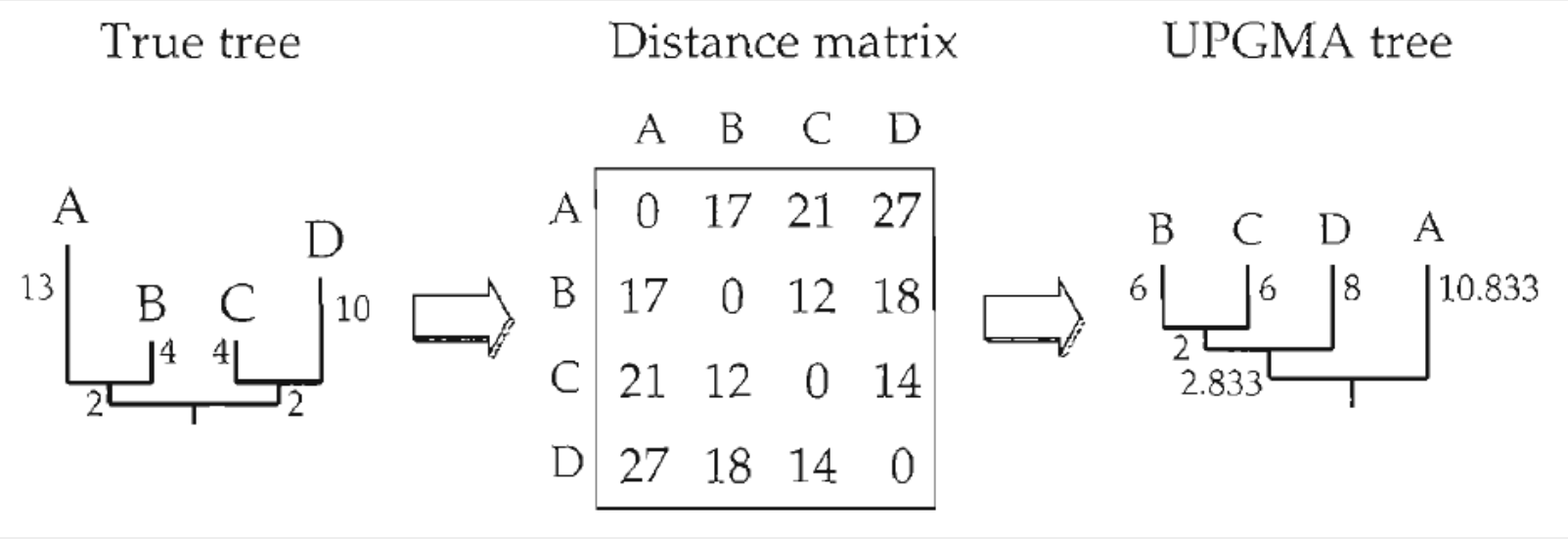
Методы расстояний: UPGMA

1. Находим два самых близких значения в матрице.
2. Объединяем соответствующие узлы в группу.
3. Перерасчитываем расстояния до остальных узлов как среднее расстояние до первого и второго членов группы.
4. Строим новую матрицу (-1 строка, -1 столбец).
5. Если в матрице >1 значения, см. п. 1.



Иногда UPGMA ошибается

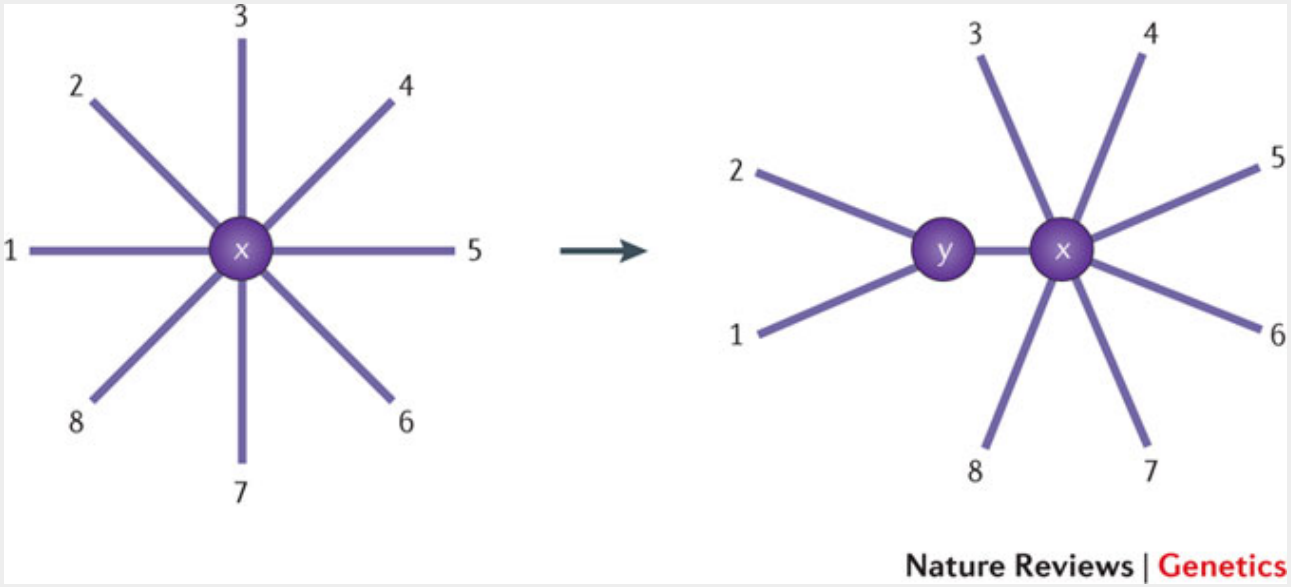
UPGMA не подходит, если скорость эволюции в разных ветвях разная!



Felsenstein, Joseph. Inferring phylogenies. Vol. 2. Sunderland: Sinauer associates, 2004

Методы расстояний: Neighbor joining (NJ)

1. Выбираем два случайных узла, объединяем их, всё остальное — второй узел.

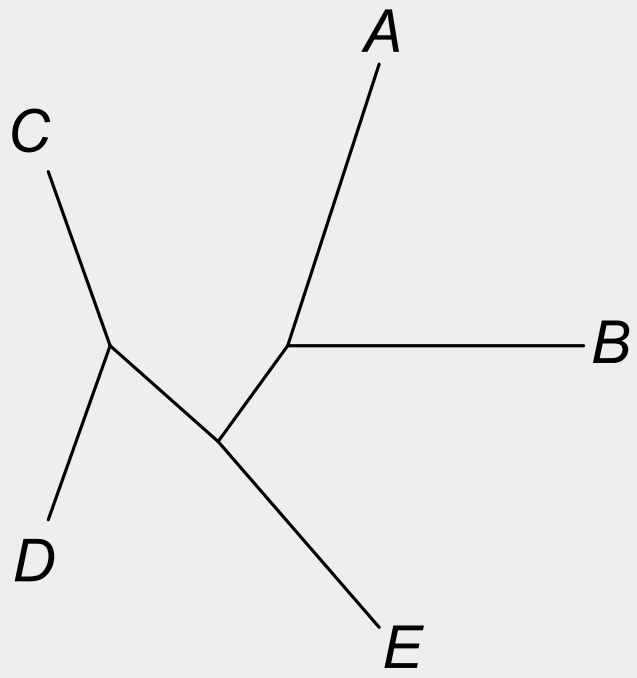


2. Определяем расстояния в матрице.
3. Повторяем со всеми остальными вариантами.
4. Минимальное расположение фиксируем.
5. Строим новую матрицу (-1 строка, -1 столбец).
6. Если в матрице >1 значения, см. п. 1.

Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nature Reviews Genetics. 2012 May 1;13(5):303-14.

Методы расстояний: NJ

1. Выбираем два случайных узла, объединяем их, всё остальное — второй узел.
2. Определяем расстояния в матрице.
3. Повторяем со всеми остальными вариантами.
4. Минимальное расположение фиксируем.
5. Строим новую матрицу (-1 строка, -1 столбец).
6. Если в матрице >1 значения, см. п. 1.



Методы расстояний: NJ

1. Выбираем два случайных узла, объединяем их, всё остальное — второй узел.
2. Определяем расстояния в матрице.
3. Повторяем со всеми остальными вариантами.
4. Минимальное расположение фиксируем.
5. Строим новую матрицу (-1 строка, -1 столбец).
6. Если в матрице >1 значения, см. п. 1.

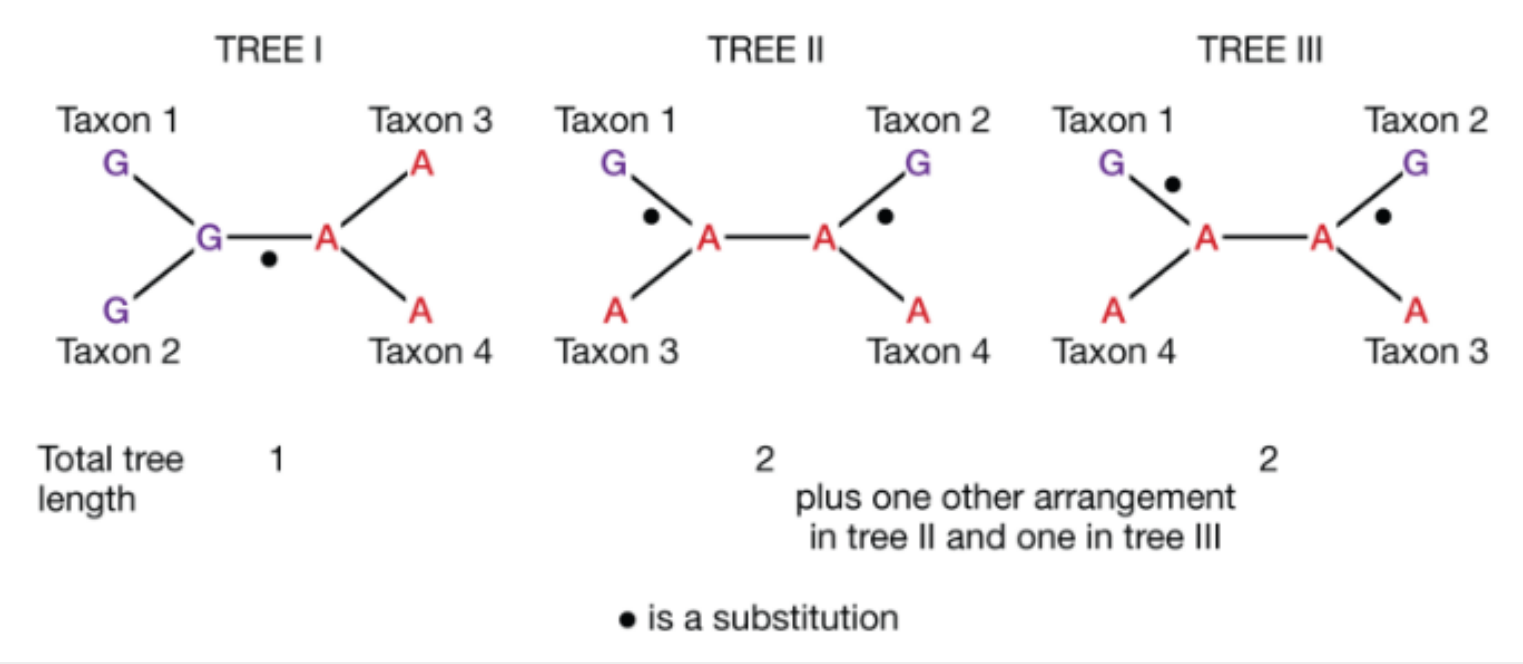
Наиболее используемая модификация: BioNJ

Основное достоинство: высокая скорость работы.

Дискретные методы: метод максимальной парсимонии (MP)

Идея: эволюция экономна.

Ищем дерево, в котором как можно меньше изменений.



Mount DW. Maximum parsimony method for phylogenetic prediction. Cold Spring Harbor Protocols. 2008 Apr 1;2008(4):pdb-top32.

Эвристики для поиска в пространстве деревьев

1. Последовательное добавление (stepwize addition).
2. Обмен ближайшими соседями (nearest neighbor interchange, NNI).
3. «Обрезка и прививка» ветви (subtree pruning and regrafting, SPR).
4. Деление дерева пополам с последующим воссоединением (tree bisection and reconnection, TBR).

Актуально для MP и ML

Дискретные методы: MP

Достоинства:

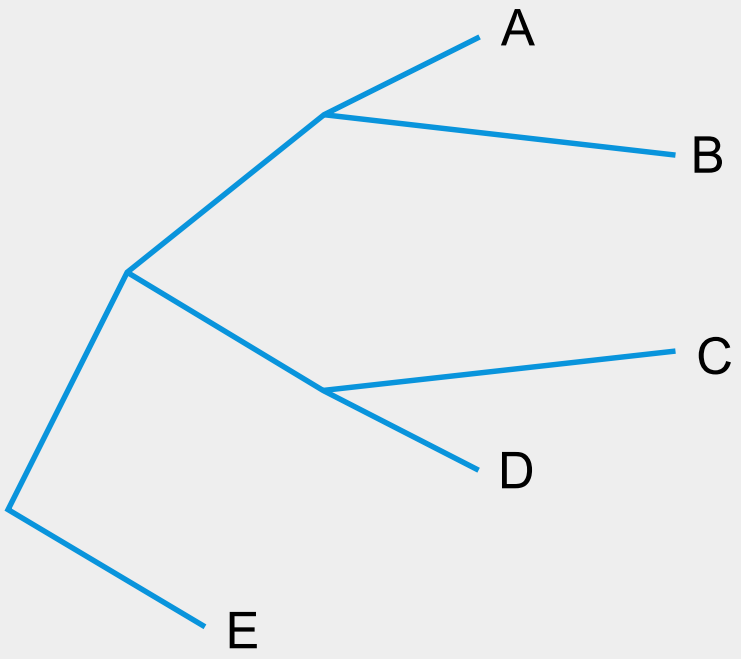
1. используем больше данных, чем в случае NJ;
2. всё ещё достаточно быстро.

Недостатки:

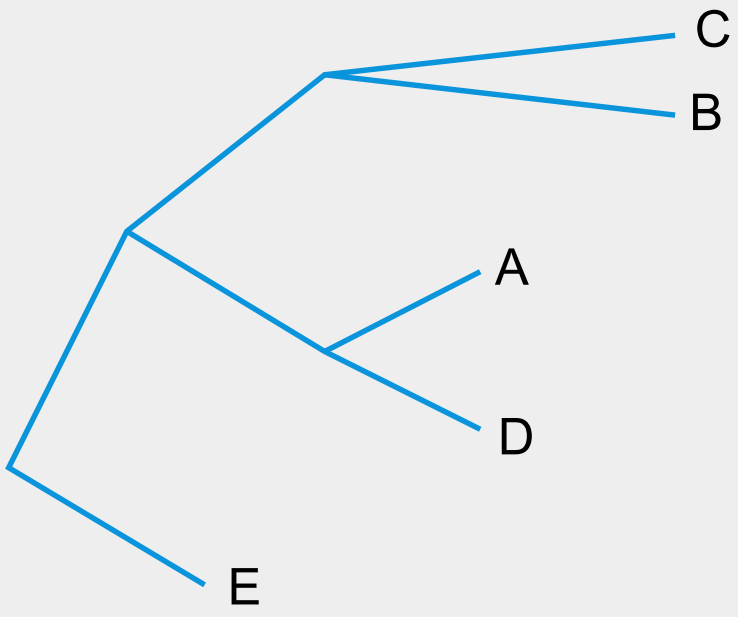
1. не самая адекватная модель с точки зрения биологии;
2. нет информации о длинах ветвей;
3. не учитываем двойные и более замены => эффект притяжения длинных ветвей.

Проблема притяжения длинных ветвей (long branch attraction)

Настоящее дерево:

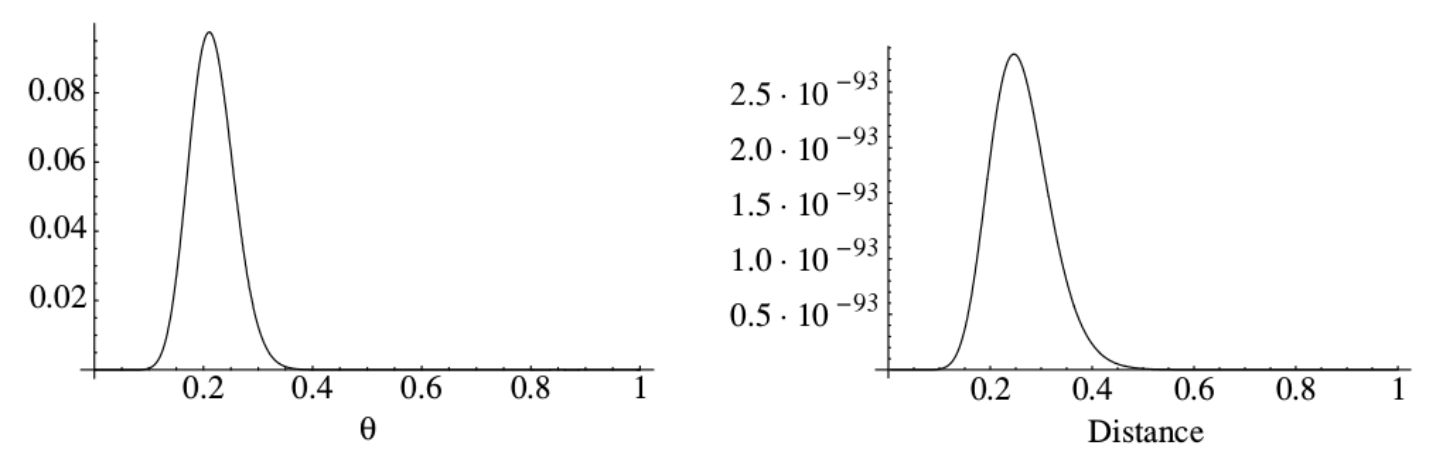


Предполагаемое дерево (inferred tree):



По http://evolution-textbook.org/content/free/figures/ch27.html
Актуально и для MP, и для ML.

Дискретные методы: метод максимального правдоподобия (ML)



Функция правдоподобия для монетки и выравнивания (JC69).

$$L(\tau, \Sigma) = Pr(Data|\tau, \Sigma) = Pr(alignment|tree, model)$$

Schmidt HA, von Haeseler A. Phylogenetic inference using maximum likelihood methods. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. 2009.

Дискретные методы: ML

$$L(\tau, \Sigma) = Pr(Data|\tau, \Sigma) = Pr(alignment|tree, model)$$

Достоинства:

- наиболее биологически оправданный.

Недостатки:

1. время- и ресурсоёмкость;
2. сильно зависит от выбранной модели.

Программы

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

- Mesquite
- MEGA
- PhyML
- RAxML
- Garli