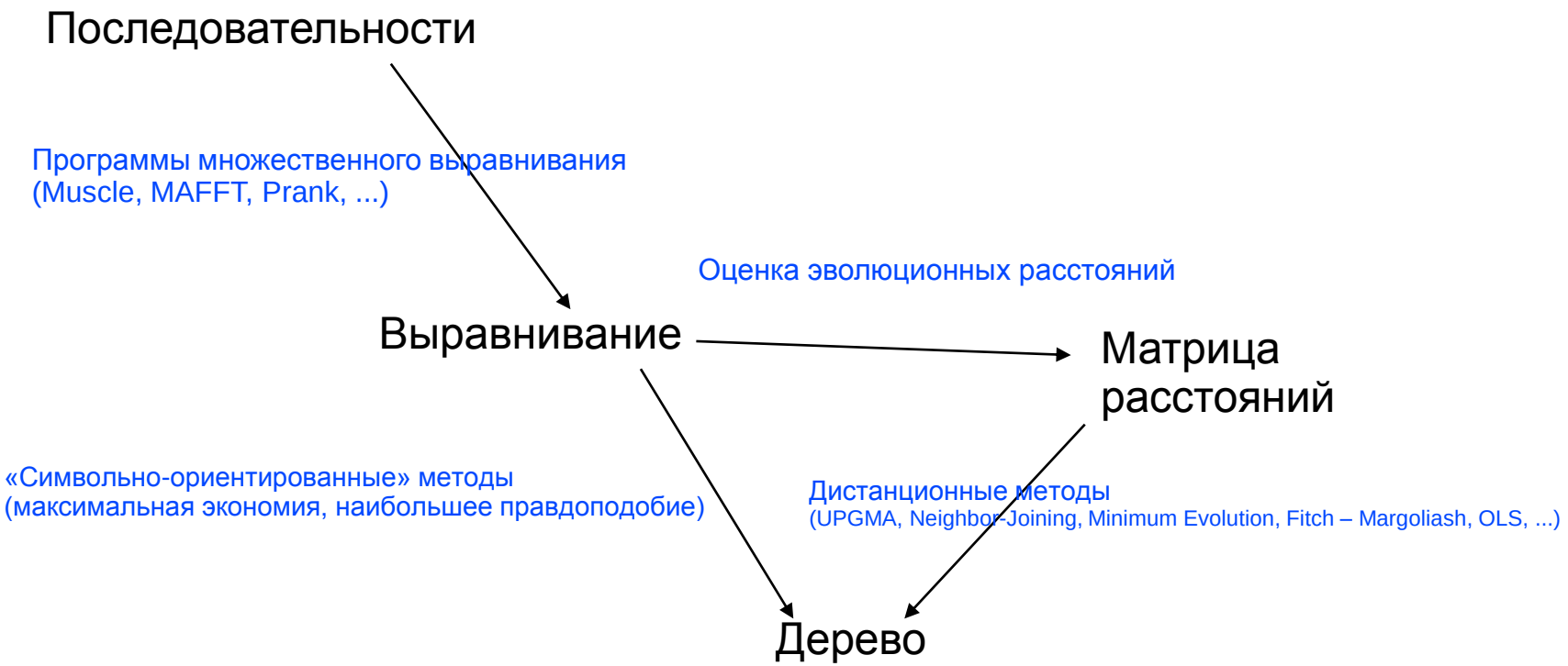
Алгоритмы реконструкции филогении

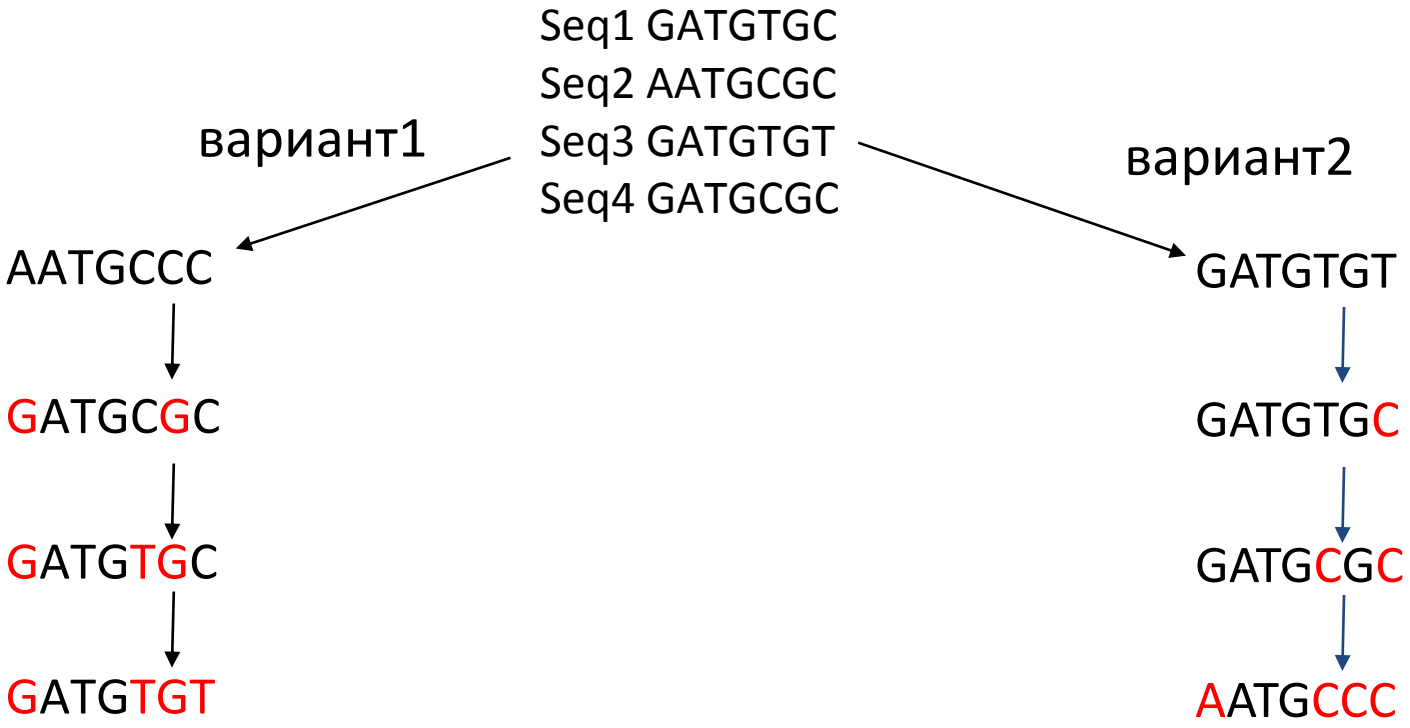
ФББ, IV семестр,
весна 2025

С.А.Спирин

Схема реконструкции филогении по последовательностям

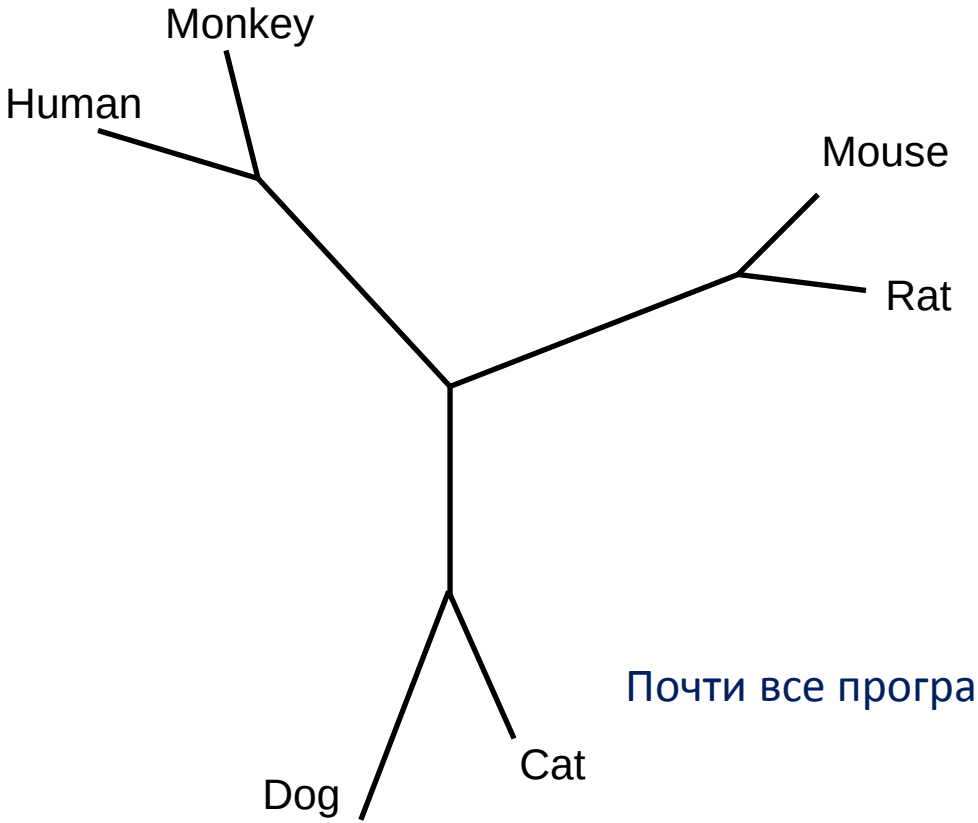


Направление эволюции?



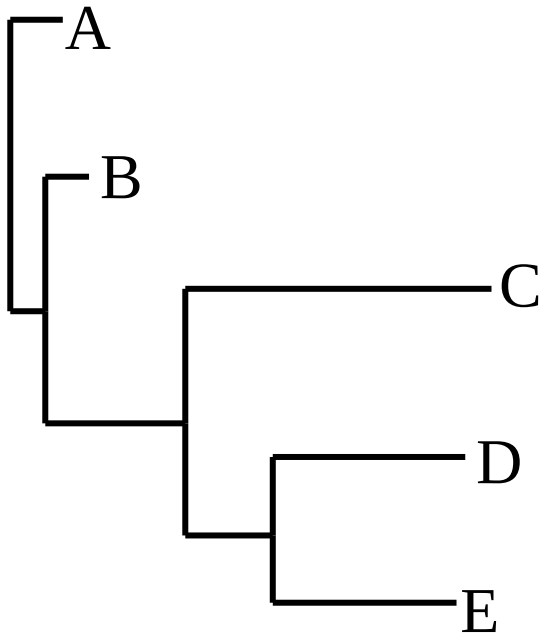
Какие деревья будут соответствовать таким вариантам?

Неукоренённое дерево

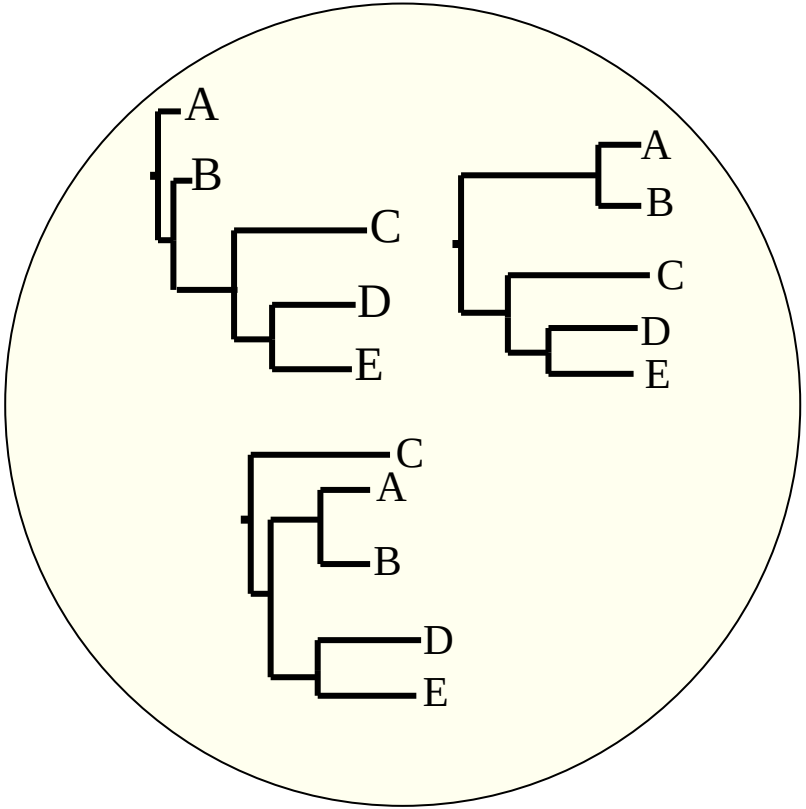


Почти все программы выдают неукоренённые деревья!

Неукоренённое дерево следует понимать как множество возможных укоренений

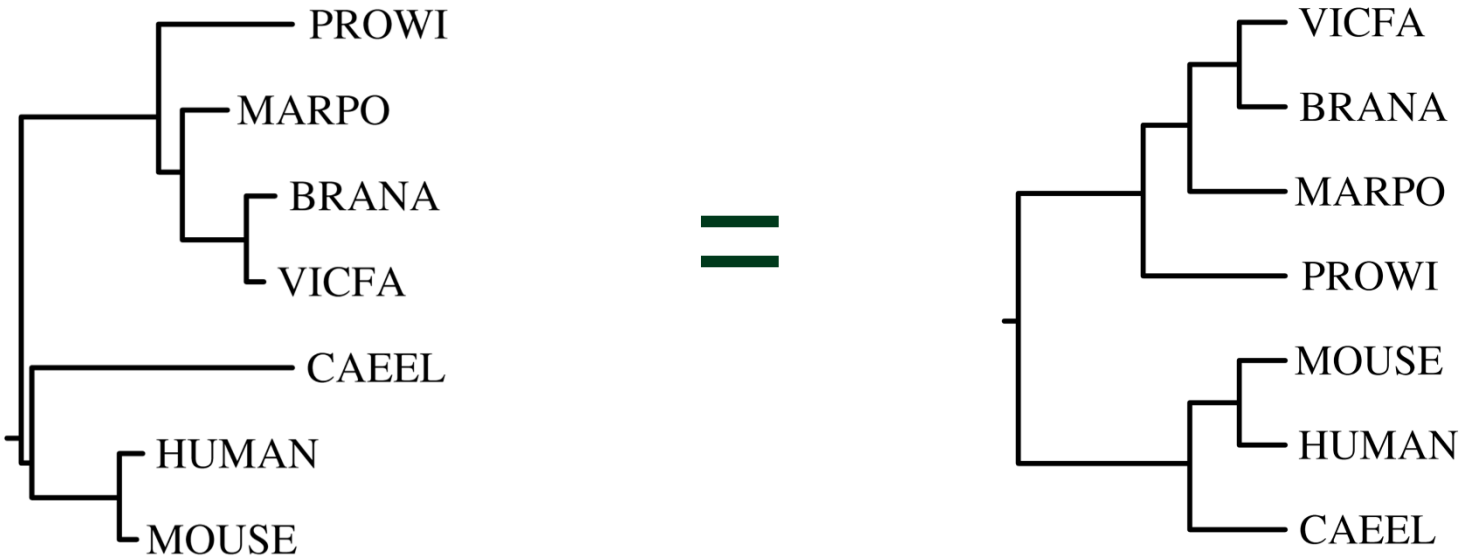


=



Есть еще варианты?

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.

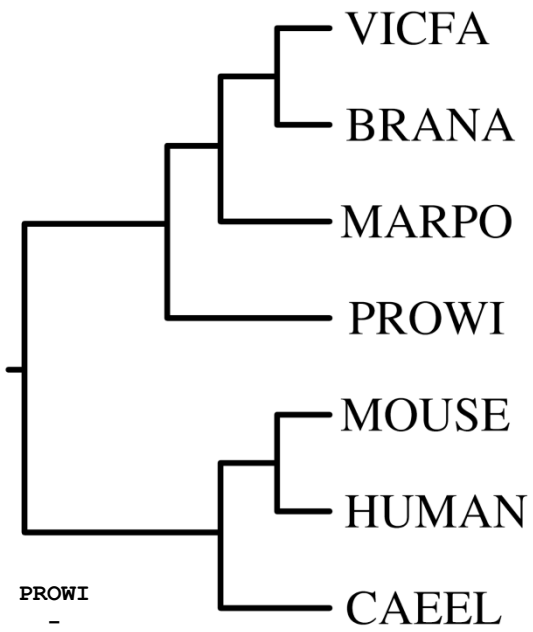
Топологию (неукоренённого) дерева можно однозначно записать набором нетривиальных разбиений. Например:

{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
+	+	-	-	-	-	-
+	+	+	-	-	-	-
+	+	+	-	-	-	+
+	+	+	-	-	+	+

Классификация методов

Название метода	Переборный / прямой	Использует молекулярные часы	Символьный/ дистанционный	Реконструирует длины ветвей
UPGMA	Прямой	Да	Дистанционный	Да
Neighbor-Joining (и BioNJ)	Прямой	Нет	Дистанционный	Да
Наименьших квадратов	Переборный	Может	Дистанционный	Да
Фитча – Марголиаша	Переборный	Может	Дистанционный	Да
Минимальной эволюции	Переборный	Может	Дистанционный	Да
Максимальной экономии	Переборный	Нет	Символьный	Нет
Наибольшего правдоподобия	Переборный	Может	Символьный	Да
Байесовский	Переборный	Может	Символьный	Да

Классификация методов

Дистанционные: берут на вход матрицу расстояний
Символьные: берут на вход выравнивания

Прямые: строят дерево по некоторому алгоритму
Переборные: перебирают некоторое количество деревьев и выбирают лучшее (согласно некоторой метрике)

Методы, предполагающие молекулярные часы, строят укоренённые **ультраметрические** деревья.
Дерево (с длинами ветвей) называется ультраметрическим, если расстояния от корня до всех листьев по дереву одинаковы.

Методы, не предполагающие молекулярные часы, строят неукоренённые деревья. Их приходится укоренять отдельно.

Матрица расстояний

	MUSDO	CHICK	BOVIN	HUMAN
MUSDO	0	9.5	8.9	9.2
CHICK	9.5	0	3.4	2.8
BOVIN	8.9	3.4	0	1.7
HUMAN	9.2	2.8	1.7	0

Как оценить расстояние?

Seq1 GATGTGC
Seq2 AATGCGC
Seq3 GATGTGT
Seq4 GATGCGC

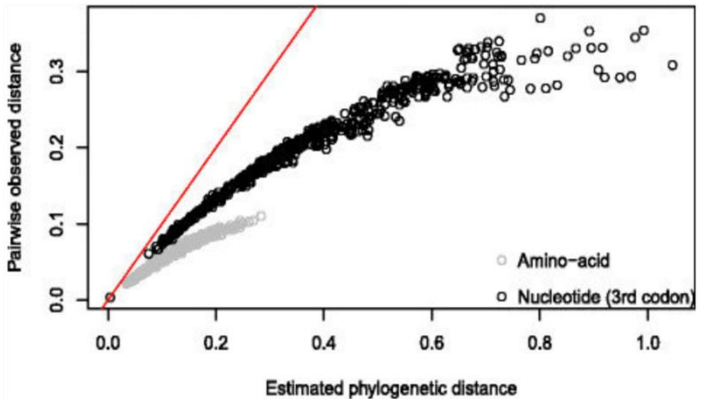
1. Число отличий
2. Число отличий / длину последовательности
   (это называется **p-distance**)
   Чем плохо?

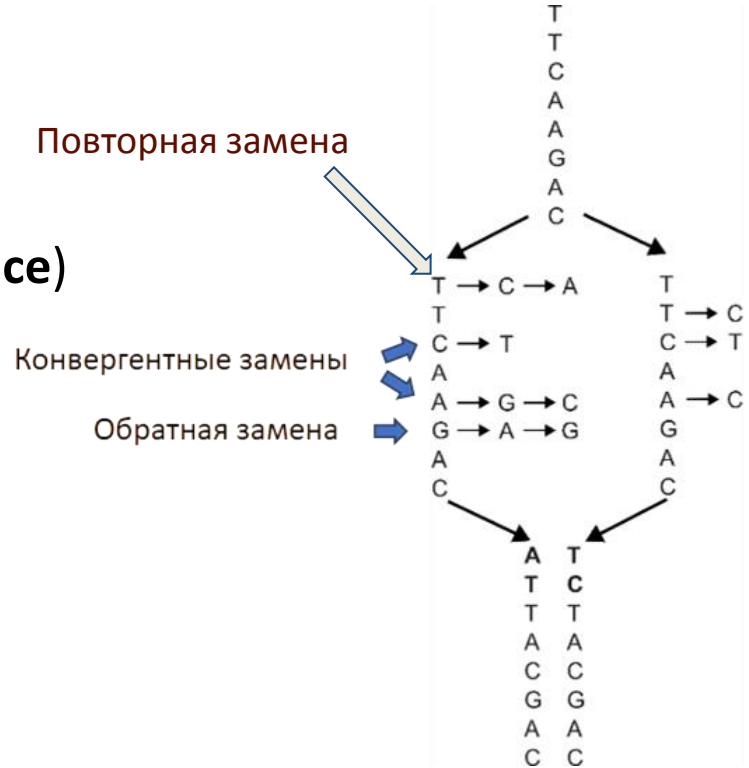
Как оценить расстояние?

Seq1 GATGTGC
Seq2 AATGCGC
Seq3 GATGTGT
Seq4 GATGCGC

1. Число отличий
2. Число отличий / длину последовательности (это называется **p-distance**)

Чем плохо?





https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0542-8

Как оценить расстояние?

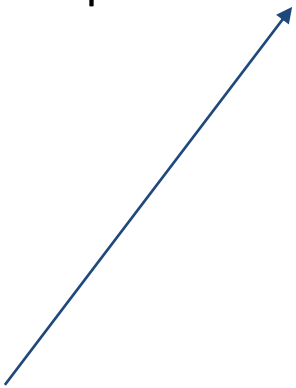
Seq1 GATGTGC
Seq2 AATGCGC
Seq3 GATGTGT
Seq4 GATGCGC

1. Число отличий
2. p-distance
3. + поправка на вероятность повторных замен
   + поправка на разные вероятности разных замен

Как оценить расстояние?

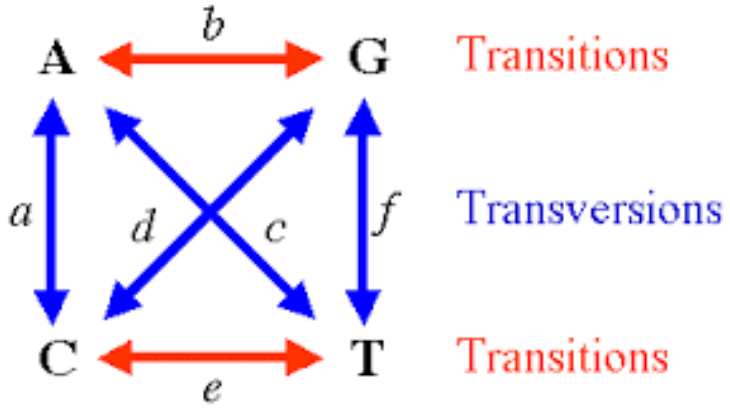
Seq1 GATGTGC
Seq2 AATGCGC
Seq3 GATGTGT
Seq4 GATGCGC

1. Число отличий
2. p-distance
3. + поправка на вероятность повторных замен
   + поправка на разные вероятности разных замен



Откуда брать эти вероятности?

Модели нуклеотидных замен



На стрелках — скорости замен

Их понимают как вероятности замен при заданном условии (обычно при условии одной произошедшей замены на каждые 100 позиций выравнивания)

Jukes-Cantor (JC69)
a=b=c=d=e=f
A=C=G=T=1/4

Kimura 2-parameter (K80)
b=e, a=c=d=f
A=C=G=T=1/4

Hasegawa-Kishino-Yano(HKY85)
b=e, a=c=d=f
A!=C!=G!=T

GTR (General Time Reversible)
a,b,c,d,e,f
A!=C!=G!=T

Модели замен

Table 1.
Substitution Models and Algorithms Available in FastME 2.0.

Models			
		Target	**Method**
DNA	p-distance	General	Analytical formula
	RY symmetric		
	RY		
	JC69 (Jukes, *Mam. Prot. Metab.*, 1969)		
	K2P (Kimura, *J. Mol. Evol.*, 1980)		
	F81 (Felsenstein, *J. Mol. Evol.*, 1981)		
	F84 (Felsenstein, *Evolution*, 1984)		
	TN93 (Tamura, *MBE*, 1993)		
	LogDet (Lockhart, *MBE*, 1994)		
Protein	p-distance	General	Analytical formula
	F81-like	General	Analytical formula
	LG (Le, *MBE*, 2008)	General	ML estimation
	WAG (Whelan, *MBE*, 2001)	General	ML estimation
	JTT (Jones, *CABIOS*, 1992)	General	ML estimation
	Dayhoff (Dayhoff, *A. Prot. Seq. Struct.*, 1978)	General	ML estimation
	DCMut (Kosiol, *MBE*, 2004)	General	ML estimation
	CpRev (Adachi, *J. Mol. Evol.*, 2000)	Chloroplast	ML estimation
	MtREV (Adachi, *J. Mol. Evol.*, 1996)	Mitochondria	ML estimation
	RtREV (Dimmic, *J. Mol. Evol.*, 2002)	Retrovirus	ML estimation
	HIVb/w (Nickle, *PLoS One*, 2007)	HIV	ML estimation
	FLU (Dang et al., *BMC Evol. Biol.*, 2010)	Flu	ML estimation

Vincent Lefort, Richard Desper, Olivier Gascuel. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, Volume 32, Issue 10, October 2015, Pages 2798–2800

Принцип наибольшего правдоподобия (maximum likelihood, ML)

Оцениваем причины по последствиям.

Принимаем как наиболее обоснованную гипотезу тот вариант причины, при котором вероятность наблюдаемых последствий наибольшая.

В нашем случае «причина» – это эволюционное расстояние, а «последствия» – наблюдаемые замены букв. Эволюционная модель (вероятности замен для всех пар букв при заданном расстоянии) предполагается фиксированной.

Для каждого расстояния (= общего числа мутаций) считаем вероятность получить из первой последовательности вторую. За оценку расстояния (ML estimation) принимаем то, при котором эта вероятность максимальна.

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. Приписав ветвям дерева длину, будем иметь расстояние $d'(i,j)$ «по дереву».

Подберём длины ветвей так, чтобы сумма величин $(d(i,j) - d'(i,j))^2$
(по всем парам листьев i,j) была наименьшей.

Это наименьшее значение и будет критерием качества: будем считать ту топологию лучшей, для которой это значение получится меньшим.