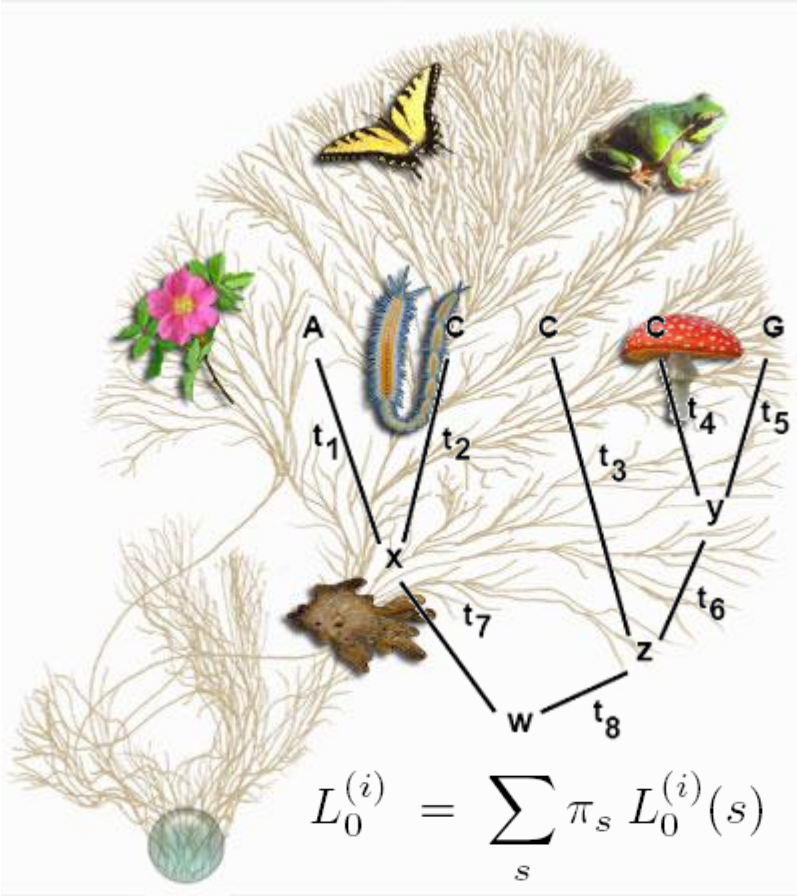
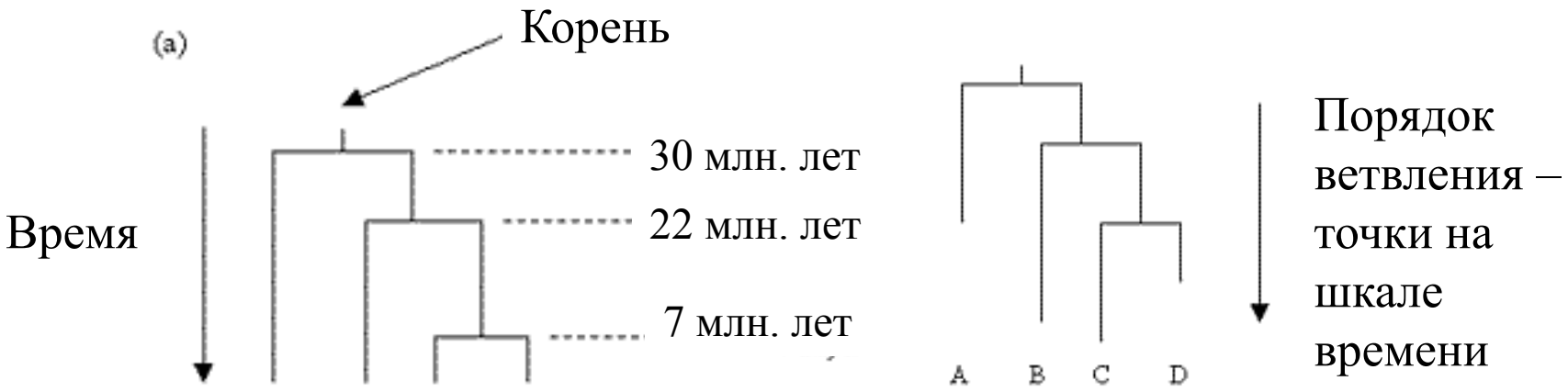
Эволюционная биоинформатика и реконструкция филогении



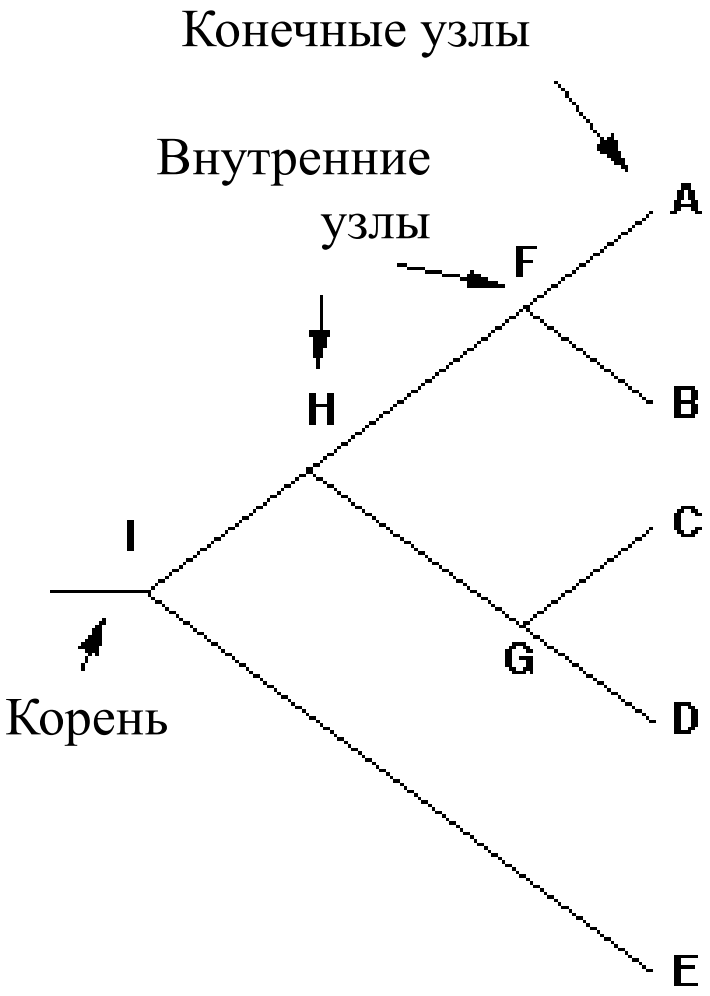
Афонников Д.А., к.б.н.
Лаборатория эволюционной биоинформатики и теоретической генетики

Филогенетические деревья



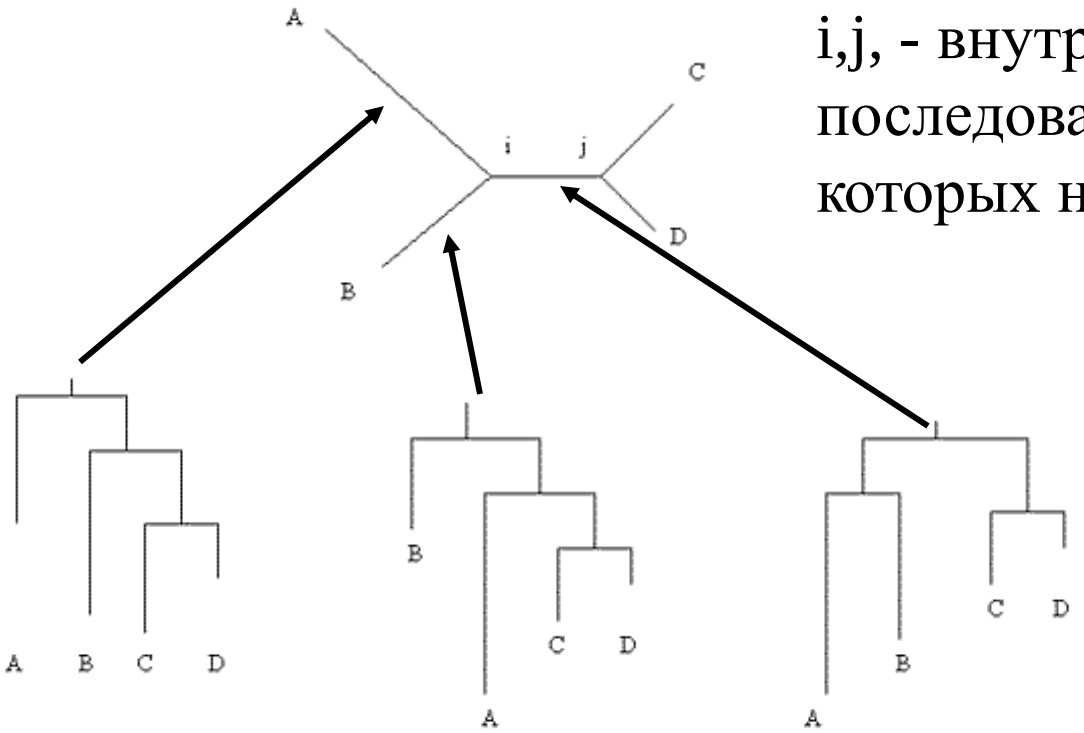
- Описывают эволюционные отношения
- Необходимы для реконструкции эволюционных событий
- Используются для функциональной аннотации белков

Филогенетические деревья



A-E – конечные узлы (листья), соответствуют таксономическим единицам (OTU);
F-I внутренние узлы (предковые)
Таксономические единицы: виды, популяции, особи, гены, белки.
Потомки эволюционируют независимо.
Топология – порядок ветвления узлов дерева.

Не все деревья имеют корень



i,j, - внутренние узлы, последовательности для которых неизвестны

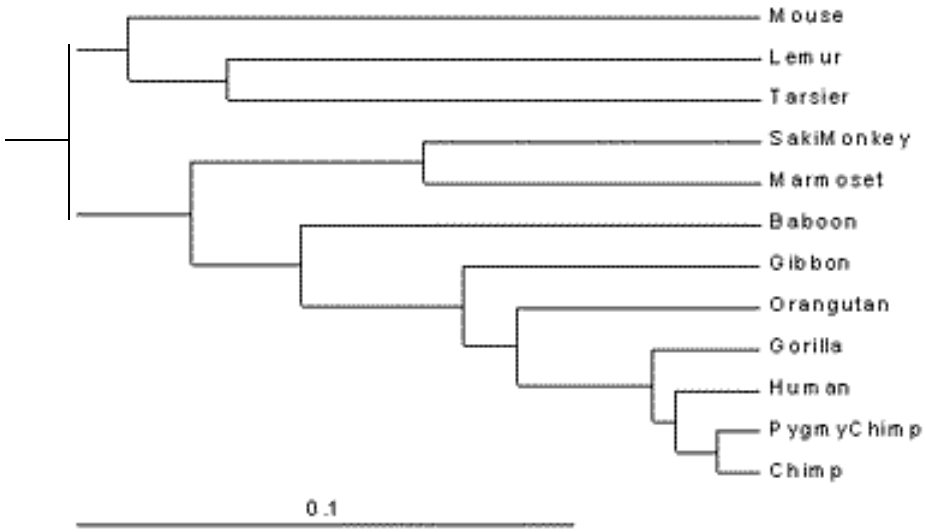
- Не все методы построения деревьев могут давать положение корня

Группы методов построения деревьев по молекулярным данным

Основанные на эволюционных расстояниях (UPGMA, объединения соседей)

Основанные на наблюдаемых признаках - нуклеотидах, аминокислотах (метод максимальной экономии, максимального правдоподобия).

Методы построения деревьев: UPGMA

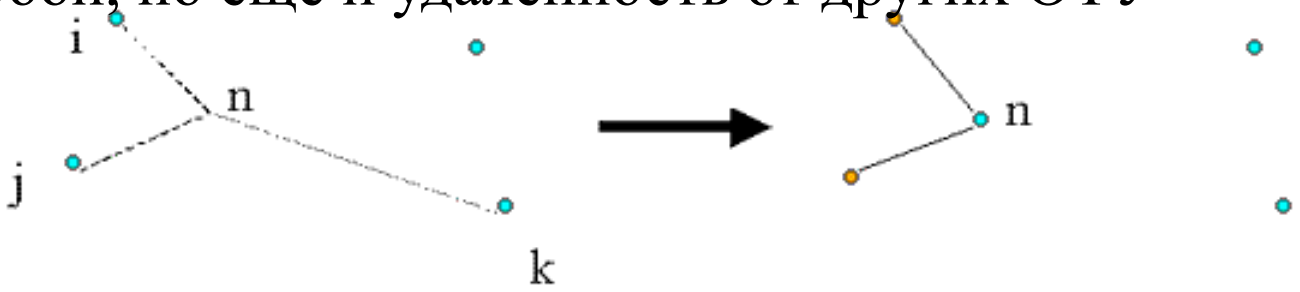


- Расстояние между кластером X и кластером Y равно среднему от парных расстояний между последовательностями этих кластеров

- Предполагает равномерность замен (молек.часы) во всех таксонах (ультраметрическое дерево)
- Расстояние=2*длину ветви
- Дает всегда дерево с корнем
- Искажает топологию дерева если скорости замен на разных ветвях различны

Метод объединения соседей (Neighbor joining)

При формировании OTU учитывается не только их близость между собой, но еще и удаленность от других ОТУ



При построении дерева два ближайших узла i,j заменяются новым узлом n; расстояния пересчитываются по следующим правилам:

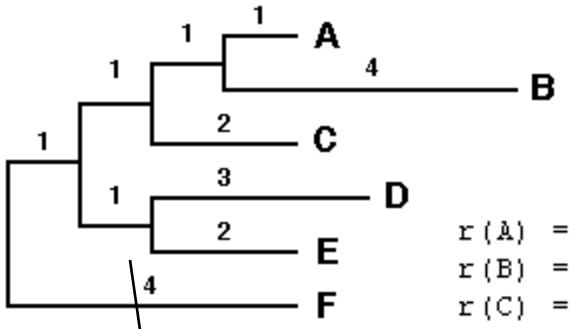
$$d_{in} + d_{nk} = d_{ik}; \qquad d_{jn} + d_{nk} = d_{jk}; \qquad d_{in} + d_{jn} = d_{ij};$$

$$d_{nk} = (d_{ik} + d_{jk} - d_{ij})/2 ;$$ Для каждого к

$$r_i = \frac{1}{N-2}\sum_k d_{ik} \qquad r_j = \frac{1}{N-2}\sum_k d_{jk}$$

$$d_{in} = (d_{ij} + r_i - r_j)/2 ; \quad d_{jn} = (d_{ij} + r_j - r_i)/2 .$$

На следующем шаге выбирается пара i,j, для которых d_{ij} минимально ($d_{ij}=d_{ij}-r_i-r_j$)



	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

```
r(A) =  5+4+7+6+8=30
r(B) = 42
r(C) = 32
r(D) = 38
r(E) = 34
r(F) = 44
```

Средняя удаленность

$$M(ij)=d(ij) - [r(i) + r(j)]/(N-2)$$

Для пары A,B

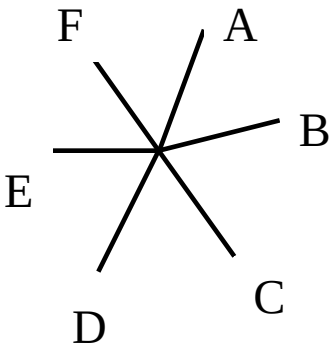
$$M(AB)=d(AB) -[(r(A) + r(B)]/(N-2) = -13$$

	A	B	C	D	E
B	-13				
C	-11.5	-11.5			
D	-10	-10	-10.5		
E	-10	-10	-10.5	-13	
F	-10.5	-10.5	-11	-11.5	-11.5

	U	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

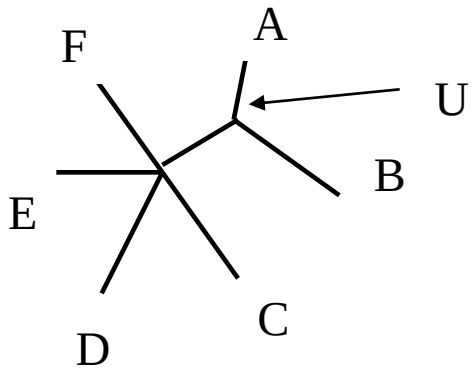
Первый шаг – стартуем с звездного дерева

Второй шаг - A,B образуют новую единицу, U




```
S(AU) =d(AB) / 2 + [r(A)-r(B)] / 2(N-2) = 1
S(BU) =d(AB) -S(AU) = 4

d(CU) = d(AC) + d(BC) - d(AB) / 2 = 3
d(DU) = d(AD) + d(BD) - d(AB) / 2 = 6
d(EU) = d(AE) + d(BE) - d(AB) / 2 = 5
d(FU) = d(AF) + d(BF) - d(AB) / 2 = 7
```



Программы реализующие данный подход

- Neighbor в пакете Phylip
- ClustalW
- Distnj в пакете Protml
- BioNJ
- QuickTree

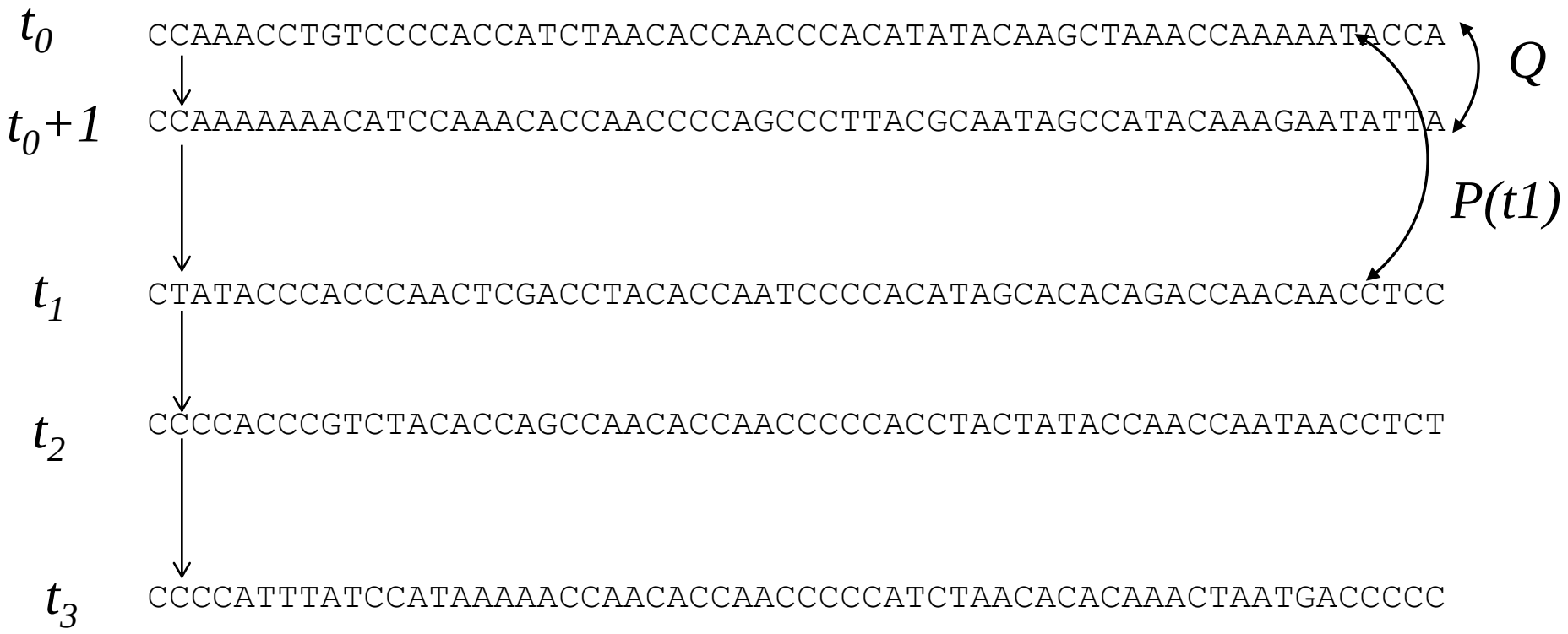
Учитывает неравномерность скоростей эволюции на ветвях дерева

Быстрый, может использоваться для больших семейств (тысячи последовательностей - QuickTree)

Иногда могут встречаться отрицательные расстояния.

Часто используется для генерации стартовой топологии дерева в методах макс. правдоподобия

Марковская модель эволюции

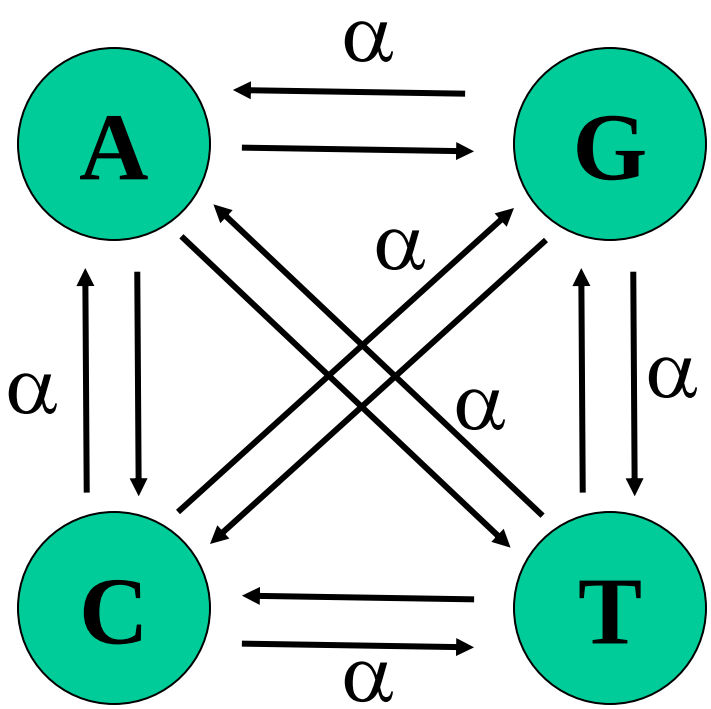


Замены в позициях независимы и происходят по одним (вероятностным) законам

Вероятности замен за время 1 (или Δt) описываются матрицей скоростей замен Q (постоянна во времени);

Вероятности замен на разных временах описываются матрицей P и меняются со временем.

Модель Джукса-Кантора



Замены случайны, независимы о других позиций, равновероятны с вероятностью α

Дискретное приближение

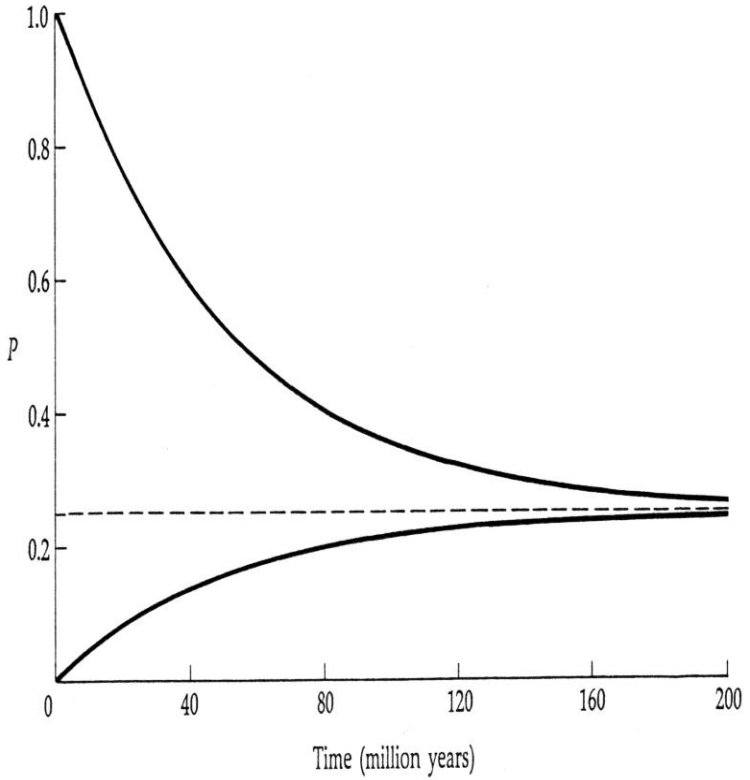
$$P_A(t+1) = (1-3\alpha)P_A(t) + \alpha[1-P_A(t)] =$$

$$= -4\alpha P_A(t) + \alpha$$

Непрерывное приближение

$$\frac{dP_A(t)}{dt} = -4\alpha P_A(t) + \alpha$$

Зависимость частоты нуклеотида ‘A’ от времени



$$P_A(t) = 0.25 + (P_A(0) - 0.25)e^{-4\alpha t}$$

$$P_A(t) = 0.25 + 0.75e^{-4\alpha t} \quad (P_A(0)=1)$$

$$P_A(t) = 0.25 - 0.25e^{-4\alpha t} \quad (P_A(0)=0)$$

Особенности:

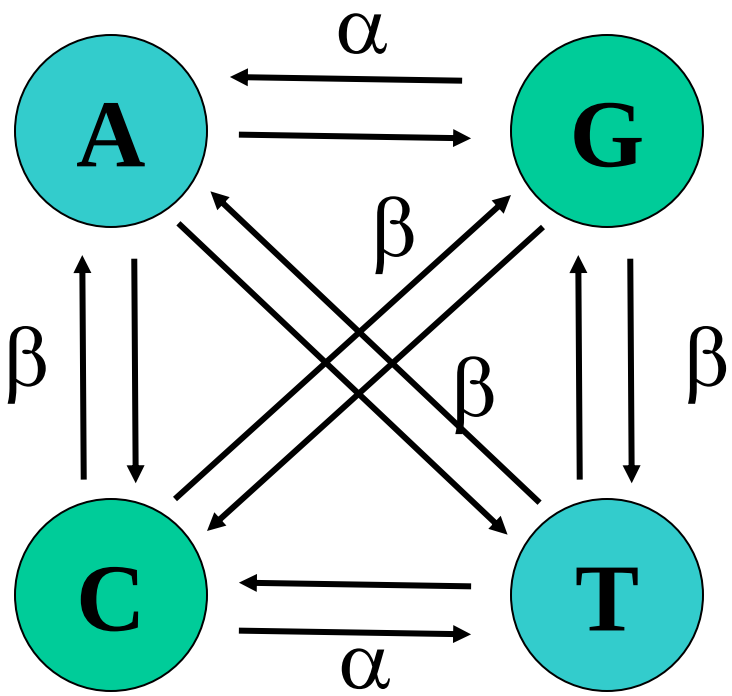
$$P_A(\infty) = P_T(\infty) = P_G(\infty) = P_C(\infty) = 1/4$$

$$P_{AA}(t) = P_{ii}(t) = 0.25 + 0.75e^{-4\alpha t}$$

$$P_{AG}(t) = P_{AT}(t) = P_{AC}(t) =$$

$$= P_{ij}(t) = 0.25 - 0.25e^{-4\alpha t}$$

Модель Кимуры



Учет разных частот
Для транзиций – (T-C,A-G) α
Для трансверсий (остальные) - β

$$P_{ii}(t) = 0.25 + 0.25e^{-4\beta t} + \\ + 0.5e^{-2(\alpha+\beta)t}$$

Однако P_{ij} различаются по парам замен (транзиции Y, трансверсии Z):

$$Y(t) = 0.25 + 0.25e^{-4\beta t} - 0.5e^{-2(\alpha+\beta)t}; Z(t) = 0.25 - 0.25e^{-4\beta t}$$

Матрица скоростей замен и ее свойства

$$p_A' = (1-\lambda_A)p_A + \lambda_{TA}p_T + \lambda_{GA}p_G + \lambda_{CA}p_C$$
$$p_T' = \lambda_{AC}p_A + (1-\lambda_T)p_T + \lambda_{GT}p_G + \lambda_{CT}p_C$$
$$p_G' = \lambda_{AG}p_A + \lambda_{TG}p_T + (1-\lambda_G)p_G + \lambda_{CG}p_C$$
$$p_C' = \lambda_{AT}p_A + \lambda_{TC}p_T + \lambda_{GC}p_G + (1-\lambda_C)p_C$$

$$\mathbf{p}' = \mathbf{Q}\mathbf{p}$$

Из:

$$\mathbf{Q} = \begin{array}{c} \begin{matrix} \text{A} & \text{T} & \text{G} & \text{C} \end{matrix} \\ \begin{bmatrix} (1-\lambda_A) & \lambda_{TA} & \lambda_{GA} & \lambda_{CA} \\ \lambda_{AT} & (1-\lambda_T) & \lambda_{GT} & \lambda_{CT} \\ \lambda_{AG} & \lambda_{TG} & (1-\lambda_G) & \lambda_{CG} \\ \lambda_{AC} & \lambda_{TC} & \lambda_{CG} & (1-\lambda_C) \end{bmatrix} \end{array}$$

Свойства:

- Сумма по столбцам равна 1
- Эволюция за n шагов эквивалентна умножению на $\mathbf{Q}^n$.
- Если **Q**=const, то существуют равновесные частоты, которые находятся из уравнения

$$\mathbf{p}=\mathbf{Q}\mathbf{p}$$

Матрица замен: непрерывное время

$$\mathbf{p}(\Delta t) = \mathbf{p}(0)\exp(\Delta t \cdot \mathbf{Q})$$

Различные меры расстояний последовательностей ДНК

Определение расстояния зависит от типа модели. Примеры:

Модель Джукса-Кантора:

$$d = -\frac{3}{4}\log_e\left(1-\frac{4}{3}p\right)$$

Модель Кимуры (P-частота транзиций, Q-частота трансверсий):

$$d = -\frac{1}{2}\log_e\left[(1-2P-Q)\sqrt{1-2Q}\right]$$

Бестиарий моделей нуклеотидных замен

Pietro Liò and Nick Goldman *Genome Res. 1998 8: 1233-1244*

Kimura, 1980

$$\mathbf{Q} = \begin{bmatrix} \cdot & \beta & \beta & \alpha \\ \beta & \cdot & \alpha & \beta \\ \beta & \alpha & \cdot & \beta \\ \alpha & \beta & \beta & \cdot \end{bmatrix}$$

π_i=1/4 [2 параметра]

Blaisdell, 1985

$$\mathbf{Q} = \begin{bmatrix} \cdot & \gamma & \gamma & \alpha \\ \delta & \cdot & \alpha & \delta \\ \delta & \beta & \cdot & \delta \\ \beta & \gamma & \gamma & \cdot \end{bmatrix}$$

π_i=1/4 [4]

Felsenstein, 1981

$$\mathbf{Q} = \begin{bmatrix} \cdot & \mu\pi_T & \mu\pi_C & \mu\pi_G \\ \mu\pi_A & \cdot & \mu\pi_C & \mu\pi_G \\ \mu\pi_A & \mu\pi_T & \cdot & \mu\pi_G \\ \mu\pi_A & \mu\pi_T & \mu\pi_C & \cdot \end{bmatrix}$$

[4 парам.]

Hasegawa, 1981

$$\mathbf{Q} = \begin{bmatrix} \cdot & \beta\pi_T & \beta\pi_C & \alpha\pi_G \\ \beta\pi_A & \cdot & \alpha\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha\pi_T & \cdot & \beta\pi_G \\ \alpha\pi_A & \beta\pi_T & \beta\pi_C & \cdot \end{bmatrix}$$

[5 парам.]

General time reversible

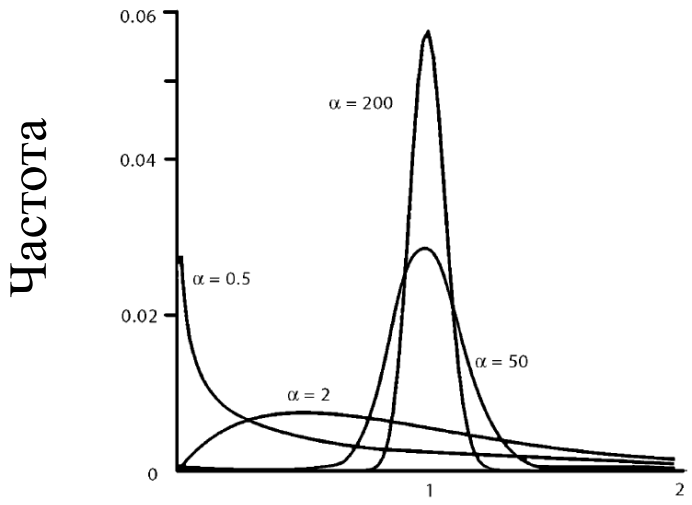
$$\mathbf{Q} = \begin{bmatrix} \cdot & \alpha\pi_T & \beta\pi_C & \gamma\pi_G \\ \alpha\pi_A & \cdot & \rho\pi_C & \sigma\pi_G \\ \beta\pi_A & \rho\pi_T & \cdot & \tau\pi_G \\ \gamma\pi_A & \sigma\pi_T & \tau\pi_C & \cdot \end{bmatrix}$$

[9 парам.]

Codon (61x61)

$$Q_{ij} = \begin{cases} 0 & \text{if 2 or 3 of the pairs } i_k, j_k \text{ are different} \\ \mu\pi_j e^{-d_{aa_i,aa_j}/V} & \text{if one pair differs by a transversion} \\ \mu\kappa\pi_j e^{-d_{aa_i,aa_j}/V} & \text{if one pair differs by a transition} \end{cases}$$

Учет неравномерности скоростей замен в последовательности



Скорость фиксации замен

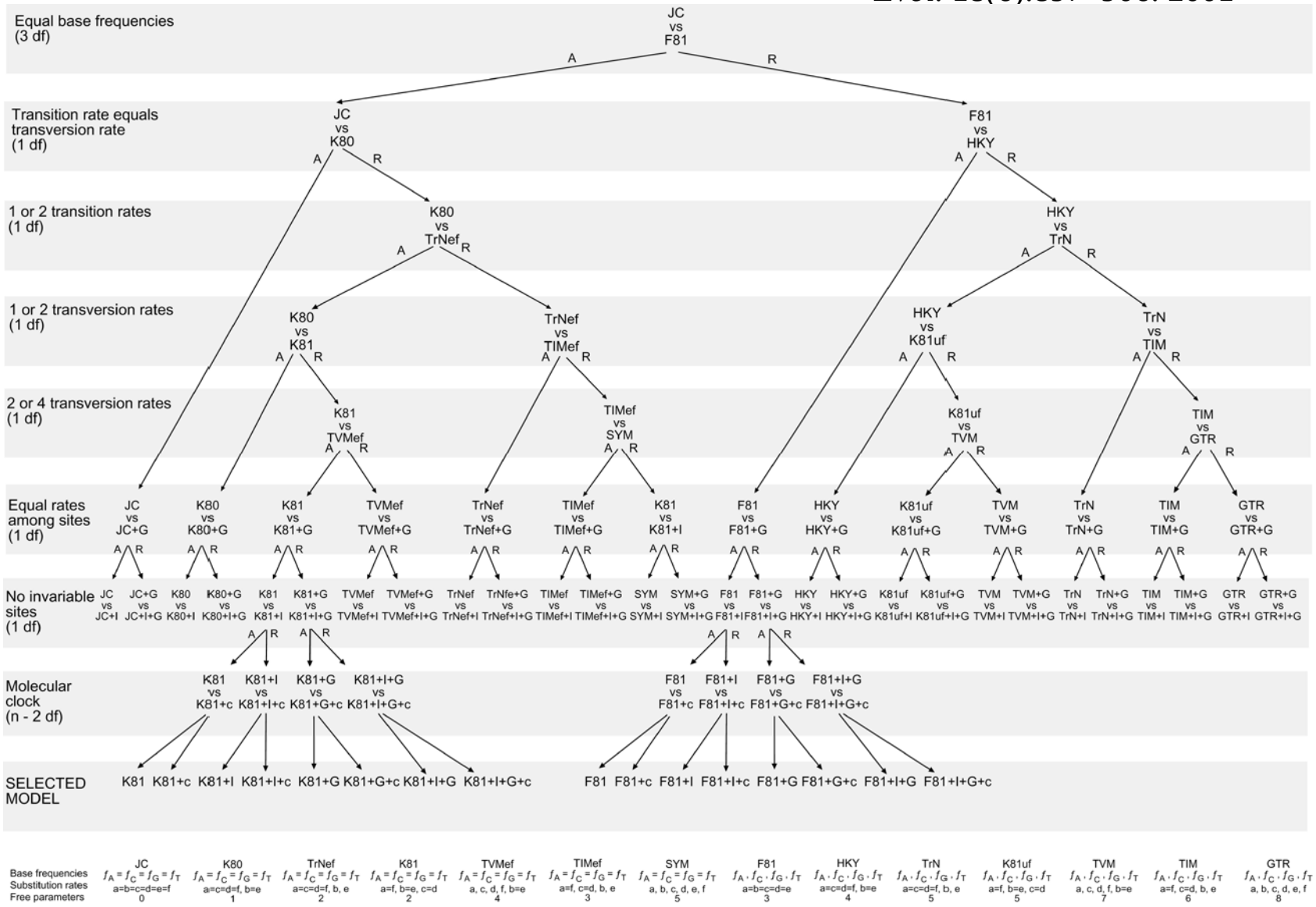
Зависимость частоты встречаемости позиций от их вариабельности описывается гамма-распределением для которого параметр β принимается равным 1, а форма зависит только от параметра α.

При анализе последовательностей форма описывается дискретным приближением (бинами гистограммы, высота которых соответствует доле позиций, с определенной скоростью замен). Обозначение модели - Г

Дополнительно можно определить долю инвариантных позиций (число от 0 до 1). Обозначение модели - I

Взаимосвязь моделей нуклеотидных замен

Posada and Crandall *Mol. Biol. Evol.* 18(6):897–906. 2001



Модель замен в белках

20 аминокислот. Замены в позициях независимы и определяются матрицей одинаковой для всех белков и всех позиций **M**(20x20). Матрица замен **M** была определена эмпирически на основе анализа нескольких семейств гомологичных белков Дайхофф и сотр. (1978).

Свойства матрицы Дайхофф:

•Равновесные частоты равны частотам встречаемости аминокислот в последовательностях белков.

•Наиболее часты замены аминокислот на аминокислоты, сходные по физико-химическим свойствам.

•Исходная матрица нормирована на время, эквивалентное 1 замене на 100 позиций (1PAM).

•Для оценки вероятности замен через время t=n надо матрицу 1PAM возвести в степень n.

PAM1:

Исходная аминокислота (j)

Аминокислота после замены (i)

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Вероятность не измениться (ALA на ALA):0.9867; измениться: 0.0133 (1.33%)

Вероятность замены ALA на GLU равна 10/10000=0.1%

Вероятность замены ALA на SER равна 28/10000=0.28%

Вероятность замены ALA на ARG равна 1/10000=0.01%

PAM250.

По мере увеличения времени эволюции элементы матрицы Mij по столбцам будут стремиться к величинам f_i – долям аминокислот в банке данных

Исходная аминокислота

Аминокислота после замены

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Вероятность не измениться (ALA на ALA):0.13; измениться: 0.87 (87%)
Вероятность замены ALA на GLU равна 5/100=5% (PAM1: 0.1%)
Вероятность замены ALA на SER равна 9/100=9% (PAM1: 0.28%)
Вероятность замены ALA на ARG равна 3/100=3% (PAM1: 0.01%)

Оценка матрицы замен на более современных данных: матрица JTT.

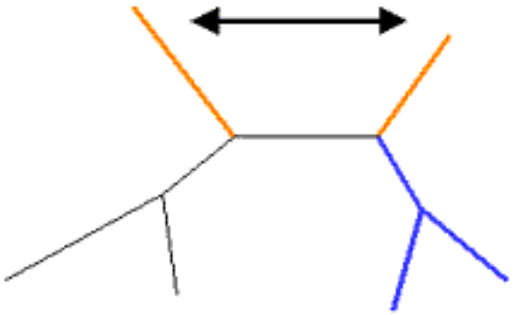
В 1992 году Jones, Taylor and Thornton оценили матрицу замен, аналогичную Дайхофф, но по большему числу последовательностей. Реконструкция предковых последовательностей не использовалась. Полученная модель замен названа JTT. Она считается лучше модели Дайхофф.

Jones DT, Taylor WR & Thornton JM (**1992**) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**: 275-282.

В 1994 г. Jones построил аналогичную матрицу для трансмембранных сегментов белков. Она ситльно отличалась от матрицы JTT, но выравнивание трансмембранных сегментов выполненное с ее использованием оказалось лучшим.

Дерево можно перестраивать и оценивать

Перестановка соседних узлов



Сокращение и перестройка поддеревьев

Число различных топологий дерева

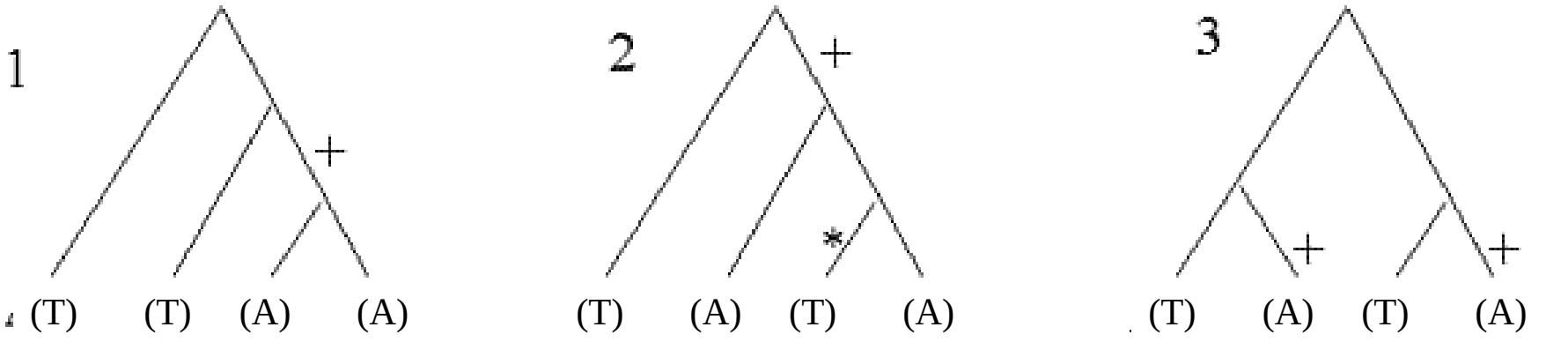
N	Без корня	С корнем
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395

$U_N = (2N-5)U_{N-1}$ $R_N = (2N-3)R_{N-1}$

- Каждому дереву можно присвоить числовую характеристику и сравнивать их

Метод парсимонии

Пример построения дерева для набора из четырех последовательностей (с одной позицией)



1) В первом дереве изменения происходят только один раз (+)
2) Во втором дереве А появляется (+) и теряется (*)
3) В третьем дереве А появляется независимо два раза (+)

Дерево (1) содержит минимальное число эволюционных событий – его и выбираем.

Программы реализующие данных подход

- Protpars (Felsentein, пакет Phylip)
- Paup (David Swofford)

Приемлем для последовательностей с высокой гомологией.
Нельзя использовать для сильно дивергировавших последовательностей!

Функция правдоподобия (ФП)

Имеются n наблюдений случайной величины x – вектор наблюдений $\mathbf{x}=(x_1,x_2,\ldots x_n)$;

Вероятность наблюдать значение x зависит от некоторого параметра θ: $p(x|\theta)$.

Тогда вероятность наблюдать n значений $\mathbf{x}=(x_1,x_2,\ldots x_n)$ равна

$$L(\mathbf{x}|\theta)=p(x_1|\theta)\ p(x_2|\theta)\ \ldots.\ p(x_n|\theta)$$

$L(X|\theta)$ называют функцией правдоподобия. Ее удобно использовать при оценке параметров распределений $p(x|\theta)$.

Идея: выбрать такой параметр, который максимизирует вероятность наблюдать набор значений $\mathbf{x}=(x_1,x_2,\ldots x_n)$.

Пример : бросание монетки

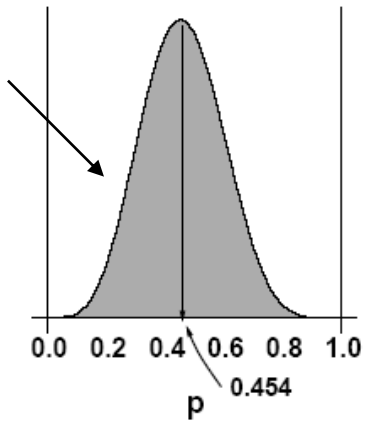
Бросаем монету, вероятность орла (О) – p, вероятность решки (Р) – 1-p. В данном случае параметр, от которого зависит вероятность наблюдать событие О – p.

Наблюдаем 11 бросаний монет: **ООРОРРРОРРО**

Функция правдоподобия:

$L=pp(1-p)p(1-p)(1-p)(1-p)p(1-p)(1-p)p$

$L=p^5(1-p)^6$



Оценка параметра p

$$\frac{\partial L}{\partial p} = \left(\frac{5}{p} - \frac{6}{1-p}\right) p^5 (1-p)^6 = 0$$

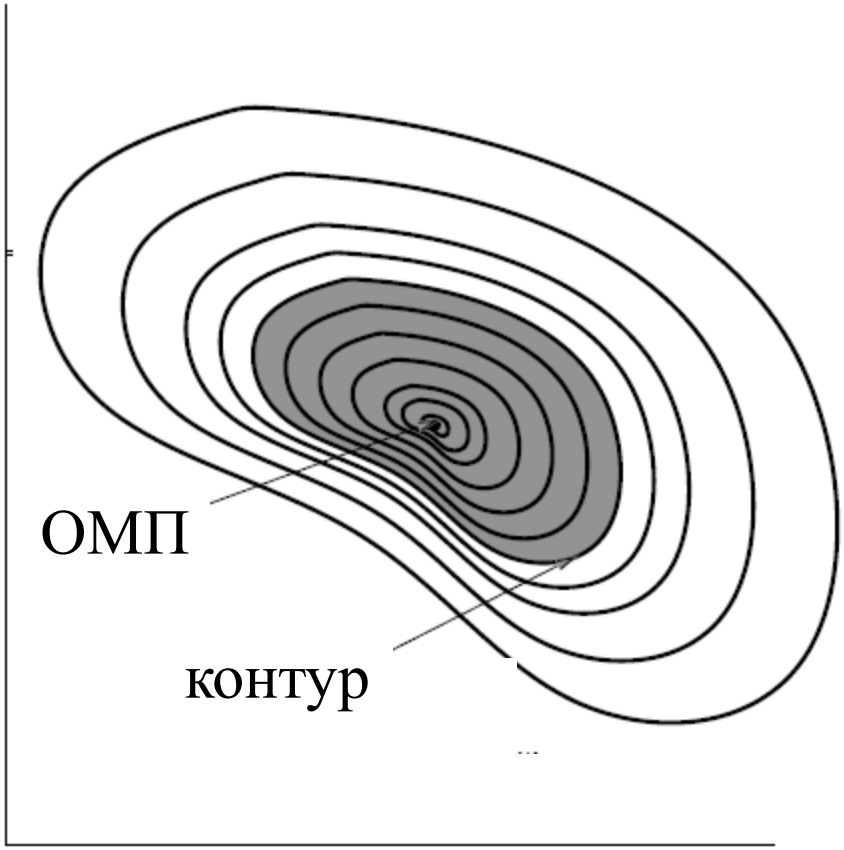
$$5 - 11p = 0$$

$$\hat{p} = \frac{5}{11}$$

Обычно используют логарифм Ф.П. (логарифмирование не меняет положение максимума)

Если много параметров

В случае нескольких параметров определяется поверхность правдоподобия. Поиск ее максимума – задача численной оптимизации.



Функция правдоподобия и проверка гипотез

D Данные

H_1 Гипотеза 1

H_2 Гипотеза 2

$|$ Условная вероятность

$$\underbrace{\frac{\text{Prob}\,(H_1 \mid D)}{\text{Prob}\,(H_2 \mid D)}}_{\text{Posterior odds ratio}} = \underbrace{\frac{\text{Prob}\,(D \mid H_1)}{\text{Prob}\,(D \mid H_2)}}_{\text{Likelihood ratio}} \underbrace{\frac{\text{Prob}\,(H_1)}{\text{Prob}\,(H_2)}}_{\text{Prior odds ratio}}$$

Выбирается гипотеза, при которой вероятность наблюдать набор данных выше .

Пример: тест на модель частот нуклеотидов

Пусть имеется последовательность нуклеотидов, в которых частоты их встречаемости π_A π_C π_G π_T

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$L = \pi_G\pi_A\pi_A\pi_G\pi_T\pi_C\pi_C\pi_T\pi_T\pi_G\pi_A\pi_G\pi_A\pi_A\pi_A\pi_T\pi_A\pi_A\pi_A\pi_C\pi_T\pi_G\pi_C\pi_A\pi_C\pi_A\pi_C\pi_A\pi_C\pi_T\pi_G\pi_G$$
$$= \pi_A^{12}\pi_C^{7}\pi_G^{7}\pi_T^{6}$$

$$\ln L = 12\ln(\pi_A) + 7\ln(\pi_C) + 7\ln(\pi_G) + 6\ln(\pi_T)$$

Гипотеза H1: частоты встречаемости π_A π_C π_G π_T равны их оценкам в последовательности

Гипотеза H2: частоты встречаемости $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$ (модель Джукса-Кантора)

Сравнение значений ФП

Гипотеза 1

$$\begin{aligned}\ln L &= 12\ln(\pi_A) + 7\ln(\pi_C) + 7\ln(\pi_G) + 6\ln(\pi_T) \\ &= 12\ln(0.375) + 7\ln(0.21875) + 7\ln(0.21875) + 6\ln(0.1875) \\ &= -43.1\end{aligned}$$

Гипотеза 2

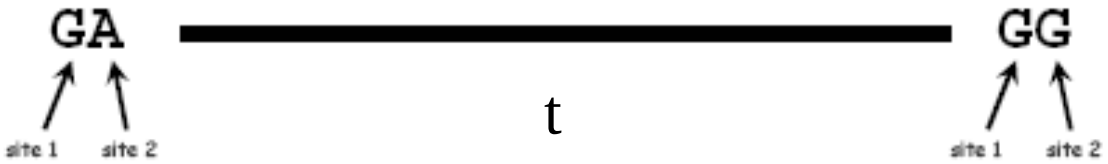
$$\begin{aligned}\ln L &= 12\ln(\pi_A) + 7\ln(\pi_C) + 7\ln(\pi_G) + 6\ln(\pi_T) \\ &= 12\ln(0.25) + 7\ln(0.25) + 7\ln(0.25) + 6\ln(0.25) \\ &= -44.4\end{aligned}$$

Принять гипотезу 1 выгоднее, первая модель более вероятна.

Пример сравнения двух последовательностей

Сравниваются две последовательности, эволюционировавшие в течении времени t со скоростью замен α

Простейший случай – 2 нуклеотида, модель Джукса-Кантора

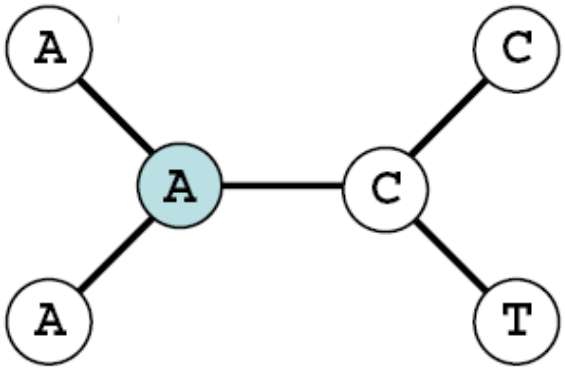


$$L = L_1 L_2$$

$$= \left[\Pr(G)\Pr(G \to G)\right]\left[\Pr(A)\Pr(A \to G)\right]$$

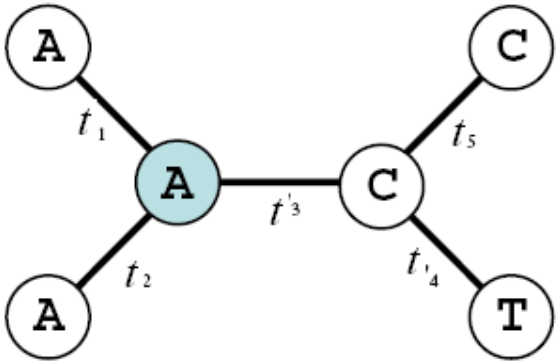
$$= \left[\tfrac{1}{4}\right]\left[\tfrac{1}{4} + \tfrac{3}{4}e^{-4\alpha t}\right]\left[\tfrac{1}{4}\right]\left[\tfrac{1}{4} - \tfrac{1}{4}e^{-4\alpha t}\right]$$

Оценка ФП для топологии дерева при фиксированной модели замен



Если нуклеотиды во внутренних узлах известны

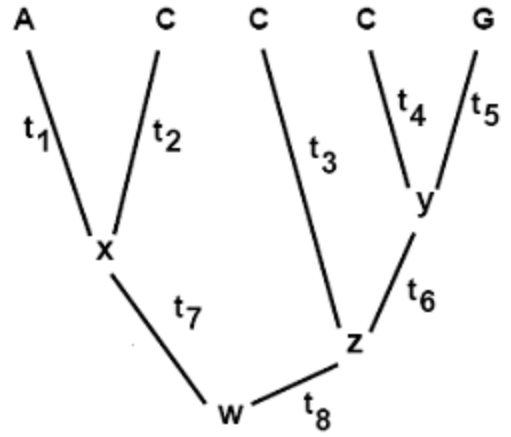
$$L_k = \Pr(A)\Pr(A \to A)\Pr(A \to A)\Pr(A \to C)\Pr(C \to T)\Pr(C \to C)$$



$$L_k = \tfrac{1}{4}\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4t_1/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4t_2/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4t_3/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4t_4/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4t_5/3}\right]$$

Но они неизвестны

Необходимо просуммировать по всем нуклеотидам во внутренних узлах дерева (усреднить)

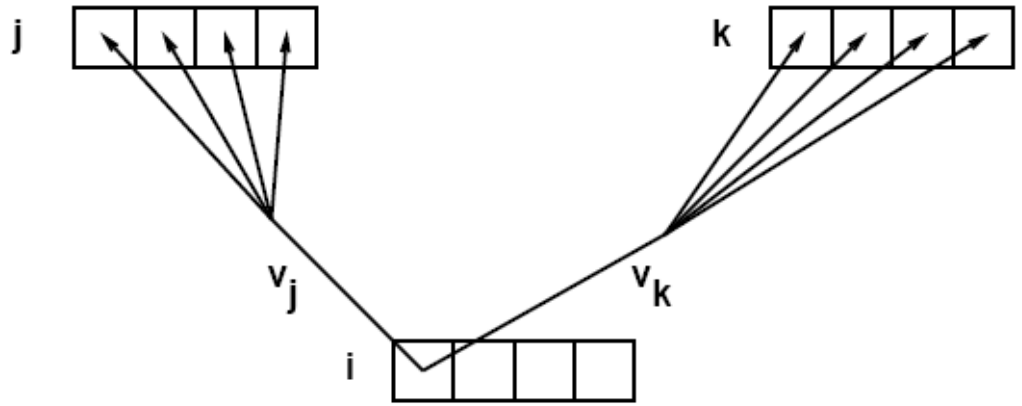


$$L^{(i)} = \sum_x \sum_y \sum_z \sum_w \text{Prob}\,(w)\ \text{Prob}\,(x \mid w, t_7)\ \text{Prob}\,(x \mid w, t_7)$$

$$\times\ \text{Prob}\,(A \mid x, t_1)\ \text{Prob}\,(C \mid x, t_2)\ \text{Prob}\,(z \mid w, t_8)$$

$$\times\ \text{Prob}\,(C \mid z, t_3)\ \text{Prob}\,(C \mid y, t_4)\ \text{Prob}\,(G \mid y, t_5)$$

Подсчет методом сокращения Felsenstein, 1981



$$L_\ell^{(i)}(s) = \left[\sum_{s_j} \text{Prob}\,(s_j \mid s, v_j) L_j^{(i)}(s_j)\right] \times \left[\sum_{s_k} \text{Prob}\,(s_k \mid s, v_k) L_k^{(i)}(s_k)\right]$$

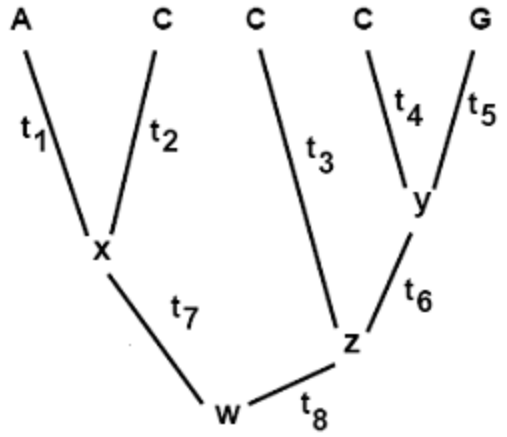
$\text{Prob}(s_j|s,v_j)$ – Вероятность наблюдать нуклеотид типа sj в дочернем узле j при условии, что в родительском узле i находится символ s и время эволюции составило v_j.

Итоговое значение

Итоговое значение усредняется по частотам нуклеотидов (аминокислот) общей предковой последовательности

$$L_0^{(i)} = \sum_s \pi_s \, L_0^{(i)}(s)$$

$$L = \prod_{i=1}^{\text{sites}} L_0^{(i)}$$



$L_0^{(i)}$

Итоговое значение Ф.П. : перемножаются для всех позиций (независимость мутаций в позициях)

Программы реализующие данный подход

- DNAML (пакет Phylip, ДНК)
- FastDNAML (ДНК)
- ProtML (ДНК и белки, Adachi and Hasegawa)
- Puzzle (Днк и белки, Strimmer and von Haeseler)
- **Phyml (ДНК, белки ; Guindon, Gasquel)**
- **RaXML (ДНК, белки (если очень много последовательностей – более 1000, быстрый эффективный), Stamatakis)**

Методы максимального правдоподобия

- С помощью методов МП можно оценивать и другие параметры: так как и матрица замен, и скорости замен могут быть такими параметрами.

$$L=L(T,M,t\ldots..).$$

- Можно усложнять модель, добавляя новые параметры.
- Метод имеет статистическое обоснование
- Но требует большого количества вычислений

Матрица замен, оцененная методом МП – WAG [Whealan and Goldman] (лучше Dayhoff, JTT).

Оценка устойчивости топологии дерева: бутстрэп

Original Data Set

Taxa	Characters
	1 2 3 4 5 6 7 8 9 10
A	C G A A C C A C T T
B	C G A A C C G G T T
C	G G T A C C G G A T
D	G C T A G C G C A T

Bootstrap Data Sets

Bootstrap Pseudoreplicate 1:

Taxa	Characters
	8 10 7 4 1 10 2 8 5 3
A	C T A A C T G C C A
B	G T G A C T G G C A
C	G T G A G T G G C T
D	C T G A G T C C G T

Bootstrap Pseudoreplicate 2:

Taxa	Characters
	1 8 10 4 2 9 2 8 5 6
A	C C T A G T G C C C
B	C G T A G T G G C C
C	G G T A G A G G C C
D	G C T A C A C C G C

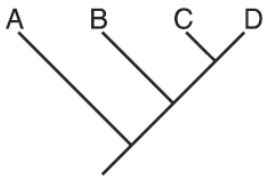
Bootstrap Pseudoreplicate 3:

Taxa	Characters
	3 2 5 7 1 6 9 4 4 10
A	A G C A C C T A A T
B	A G C G C C T A A T
C	T G C G G C A A A T
D	T C G G G C A A A T

Bootstrap Pseudoreplicate 4:

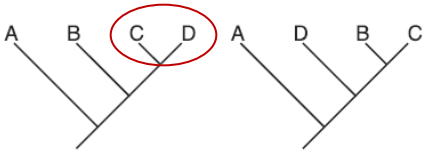
Taxa	Characters
	7 8 5 8 9 6 4 10 1 5
A	A C C C T C A T C C
B	G G C G T C A T C C
C	G C G A C C A T G C
D	G C G C A C A T G C

Tree from Original Data Set

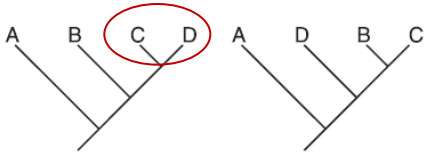


Trees from Bootstrapped Data Sets

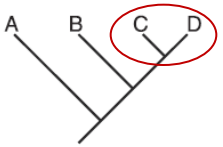
Bootstrap Pseudoreplicate 1:



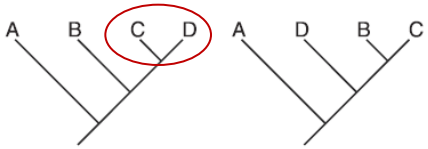
Bootstrap Pseudoreplicate 2:



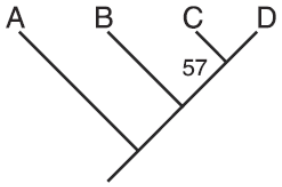
Bootstrap Pseudoreplicate 3:



Bootstrap Pseudoreplicate 4:



Bootstrap Consensus Tree:



Выбор моделей эволюции

Posada and Crandall *Mol. Biol. Evol.* 18(6):897–906. 2001

Модель должна обеспечивать максимальное значение Ф.П. при минимальном числе параметров

Если есть две модели $\mathbf{M}_1$ и $\mathbf{M}_2$, с числом свободных параметров p_1 и p_2 (p_2-p_1=k; M_1 «вложена» в M_2) для которых были получены оптимальные значения L_1 и L_2, тогда

$$\delta = 2(\ln L_1 - \ln L_0)$$

Распределена по закону χ^2 с k степенями свободы.

Если модели не являются вложенными, сравнивать их можно на основе критерия Акаике:

$$\mathrm{AIC} = -2 \ln L + 2p$$

Меньшие значения соответствуют лучшей модели. Для нескольких моделей можно сравнить с лучшей:

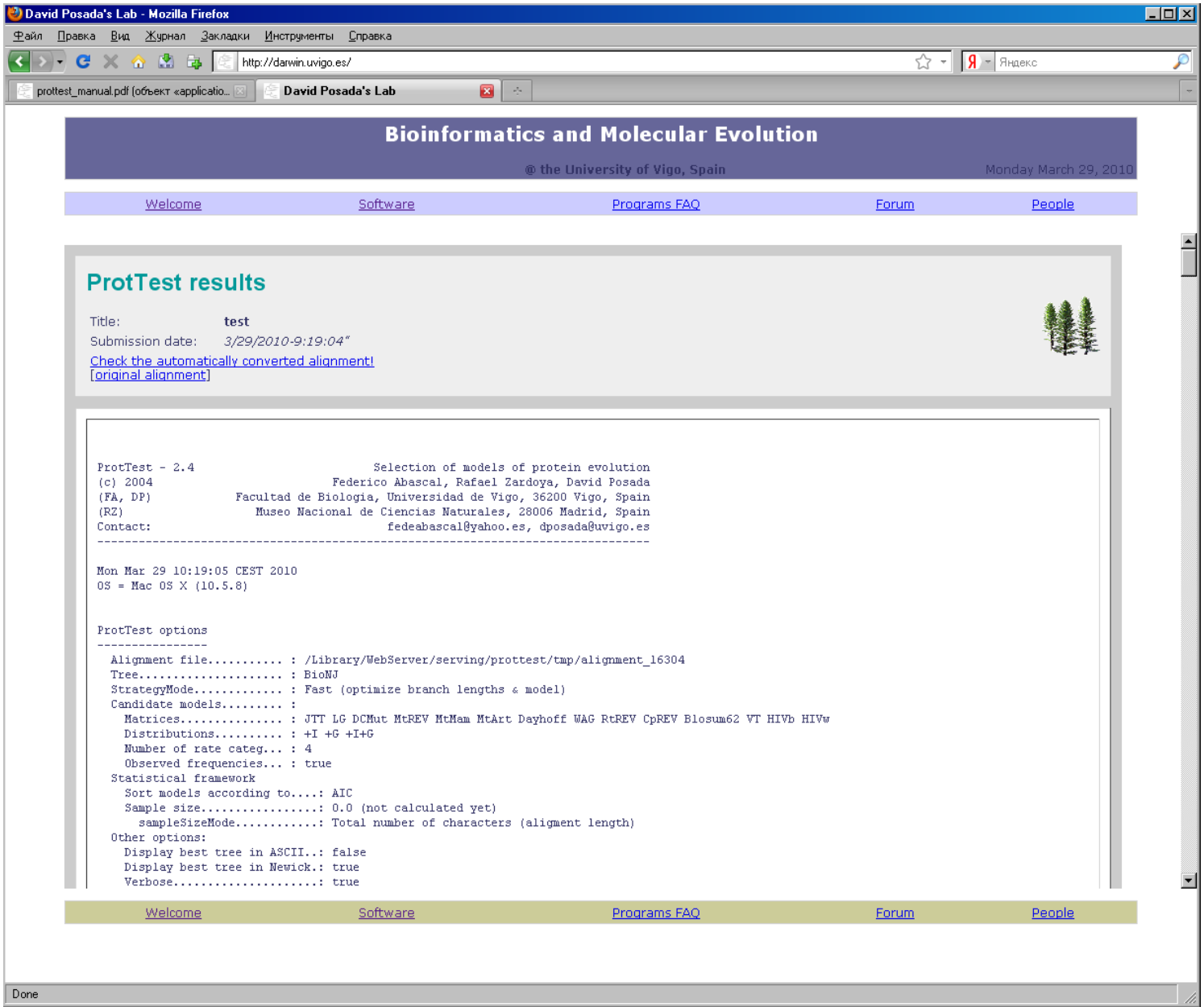
$$\Delta_i = AIC_i - \min AIC$$

Программы реализующие данный подход

Нуклеотидные последовательности:
Modelestimator
jModeltest

Аминокислотные последовательности
Prottest
Modelestimator

Пример анализа:



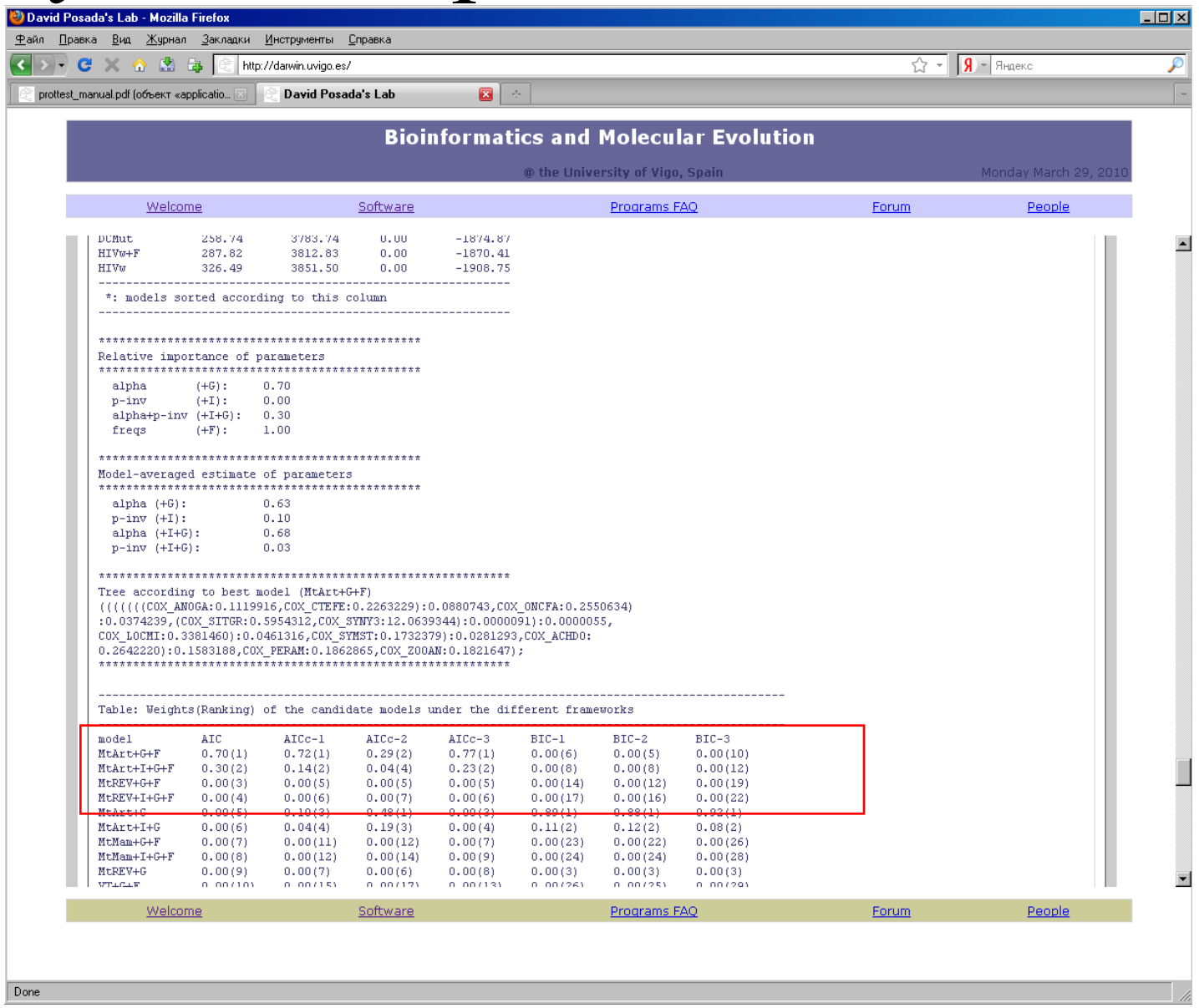
Bioinformatics and Molecular Evolution

@ the University of Vigo, Spain

Monday March 29, 2010


```
************************************************************
Best model according to AIC: MtArt+G+F
************************************************************
Model           deltaAIC*    AIC          AICw       -lnL
------------------------------------------------------------
MtArt+G+F       0.00         3525.00      0.70       -1725.50
MtArt+I+G+F     1.74         3526.75      0.30       -1725.37
MtREV+G+F       20.58        3545.58      0.00       -1735.79
MtREV+I+G+F     22.58        3547.58      0.00       -1735.79
MtArt+G         24.05        3549.05      0.00       -1756.53
MtArt+I+G       25.19        3550.20      0.00       -1756.10
MtMam+G+F       40.47        3565.47      0.00       -1745.74
MtMam+I+G+F     41.81        3566.81      0.00       -1745.41
MtREV+G         49.39        3574.39      0.00       -1769.19
VT+G+F          49.60        3574.61      0.00       -1750.30
MtREV+I+G       51.38        3576.38      0.00       -1769.19
VT+I+G+F        51.60        3576.61      0.00       -1750.30
CpREV+G         53.55        3578.55      0.00       -1771.28
CpREV+G+F       55.13        3580.13      0.00       -1753.07
CpREV+I+G       55.55        3580.55      0.00       -1771.28
CpREV+I+G+F     57.13        3582.13      0.00       -1753.07
LG+G+F          60.57        3585.58      0.00       -1755.79
JTT+G+F         61.15        3586.16      0.00       -1756.08
WAG+G+F         61.17        3586.18      0.00       -1756.09
LG+I+G+F        62.57        3587.57      0.00       -1755.79
JTT+I+G+F       63.15        3588.16      0.00       -1756.08
WAG+I+G+F       63.17        3588.18      0.00       -1756.09
Dayhoff+G+F     63.55        3588.56      0.00       -1757.28
DCMut+G+F       63.94        3588.94      0.00       -1757.47
MtMam+G         65.05        3590.05      0.00       -1777.03
Dayhoff+I+G+F   65.55        3590.56      0.00       -1757.28
DCMut+I+G+F     65.94        3590.94      0.00       -1757.47
MtMam+I+G       65.95        3590.95      0.00       -1776.48
LG+G            69.96        3594.96      0.00       -1779.48
LG+I+G          71.96        3596.96      0.00       -1779.48
JTT+G           74.87        3599.87      0.00       -1781.94
JTT+I+G         76.87        3601.87      0.00       -1781.94
VT+G            78.34        3603.34      0.00       -1783.67
VT+I+G          80.34        3605.34      0.00       -1783.67
RtREV+G+F       81.56        3606.57      0.00       -1766.28
Blosum62+G+F    81.62        3606.63      0.00       -1766.31
RtREV+I+G+F     83.56        3608.57      0.00       -1766.28
```

Результаты сравнения моделей



Bioinformatics and Molecular Evolution

@ the University of Vigo, Spain — Monday March 29, 2010

LG	0.00(94)	0.00(88)	0.00(87)	0.00(91)	0.00(76)	0.00(77)	0.00(73)
RtREV+I	0.00(95)	0.00(90)	0.00(89)	0.00(93)	0.00(79)	0.00(79)	0.00(75)
MtMam+I	0.00(96)	0.00(91)	0.00(90)	0.00(94)	0.00(80)	0.00(80)	0.00(76)
HIVb+I+F	0.00(97)	0.00(99)	0.00(100)	0.00(98)	0.00(106)	0.00(106)	0.00(106)
RtREV+F	0.00(98)	0.00(101)	0.00(101)	0.00(99)	0.00(107)	0.00(107)	0.00(107)
HIVb+I	0.00(99)	0.00(97)	0.00(95)	0.00(97)	0.00(85)	0.00(87)	0.00(81)
MtMam+F	0.00(100)	0.00(102)	0.00(103)	0.00(100)	0.00(108)	0.00(108)	0.00(108)
Dayhoff+I	0.00(101)	0.00(98)	0.00(98)	0.00(101)	0.00(91)	0.00(92)	0.00(86)
DCMut+I	0.00(102)	0.00(100)	0.00(99)	0.00(102)	0.00(92)	0.00(93)	0.00(87)
HIVw+I+F	0.00(103)	0.00(105)	0.00(106)	0.00(104)	0.00(109)	0.00(109)	0.00(109)
RtREV	0.00(104)	0.00(103)	0.00(102)	0.00(103)	0.00(95)	0.00(96)	0.00(92)
HIVb+F	0.00(105)	0.00(110)	0.00(110)	0.00(107)	0.00(110)	0.00(110)	0.00(110)
MtMam	0.00(106)	0.00(104)	0.00(104)	0.00(105)	0.00(99)	0.00(101)	0.00(96)
HIVb	0.00(107)	0.00(106)	0.00(105)	0.00(106)	0.00(102)	0.00(102)	0.00(99)
HIVw+I	0.00(108)	0.00(108)	0.00(108)	0.00(108)	0.00(105)	0.00(105)	0.00(103)
Dayhoff	0.00(109)	0.00(107)	0.00(107)	0.00(109)	0.00(103)	0.00(103)	0.00(100)
DCMut	0.00(110)	0.00(109)	0.00(109)	0.00(110)	0.00(104)	0.00(104)	0.00(101)
HIVw+F	0.00(111)	0.00(111)	0.00(111)	0.00(111)	0.00(112)	0.00(112)	0.00(112)
HIVw	0.00(112)	0.00(112)	0.00(112)	0.00(112)	0.00(111)	0.00(111)	0.00(111)

Relative importance of

parameters	AIC	AICc-1	AICc-2	AICc-3	BIC-1	BIC-2	BIC-3
+G	0.70	0.82	0.77	0.77	0.89	0.88	0.92
+I	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+I+G	0.30	0.18	0.23	0.23	0.11	0.12	0.08
+F	1.00	0.86	0.33	1.00	0.00	0.00	0.00

Model-averaged estimate of

parameters	AIC	AICc-1	AICc-2	AICc-3	BIC-1	BIC-2	BIC-3
alpha (+G)	0.63	0.62	0.59	0.62	0.58	0.58	0.58
p-inv (+I)	0.10	0.10	0.10	0.10	0.10	0.10	0.10
alpha (+I+G)	0.68	0.68	0.67	0.68	0.66	0.66	0.66
p-inv (+I+G)	0.03	0.03	0.04	0.03	0.04	0.04	0.04

AIC : Akaike Information Criterion framework.
AICc-x: Second-Order Akaike framework.
BIC-x : Bayesian Information Criterion framework.
AICc/BIC-1: sample size as: number of sites in the alignment (149.0)
AICc/BIC-2: sample size as: Sum of position's Shannon Entropy over the whole alignment (127.6)
AICc/BIC-3: sample size as: align. length x num sequences x averaged (0-1)Sh. Entropy (295.2)

Рекомендации для публикации в хороший журнал

(1) Не рекомендуется использовать ClustalW

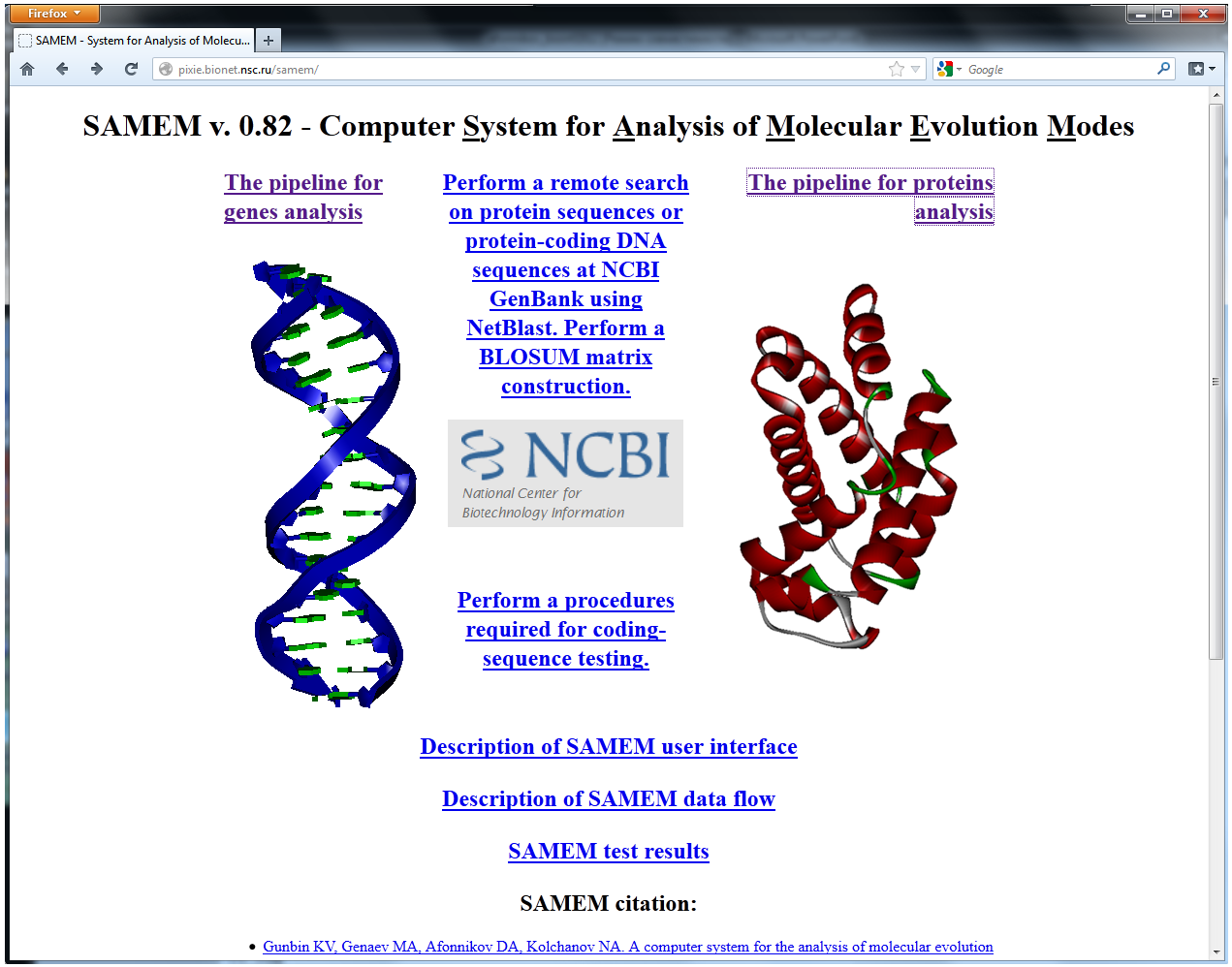
(2) Рекомендуется использовать методы максимального правдоподобия (или байесовские), особенно для сильно различающихся последовательностей

(3) Перед построением филогенетического дерева необходимо провести выбор наилучшей модели эволюции

(4) Всегда использовать бутстрэп и приводить оценки надежности узлов

Пакет программ SAMEM

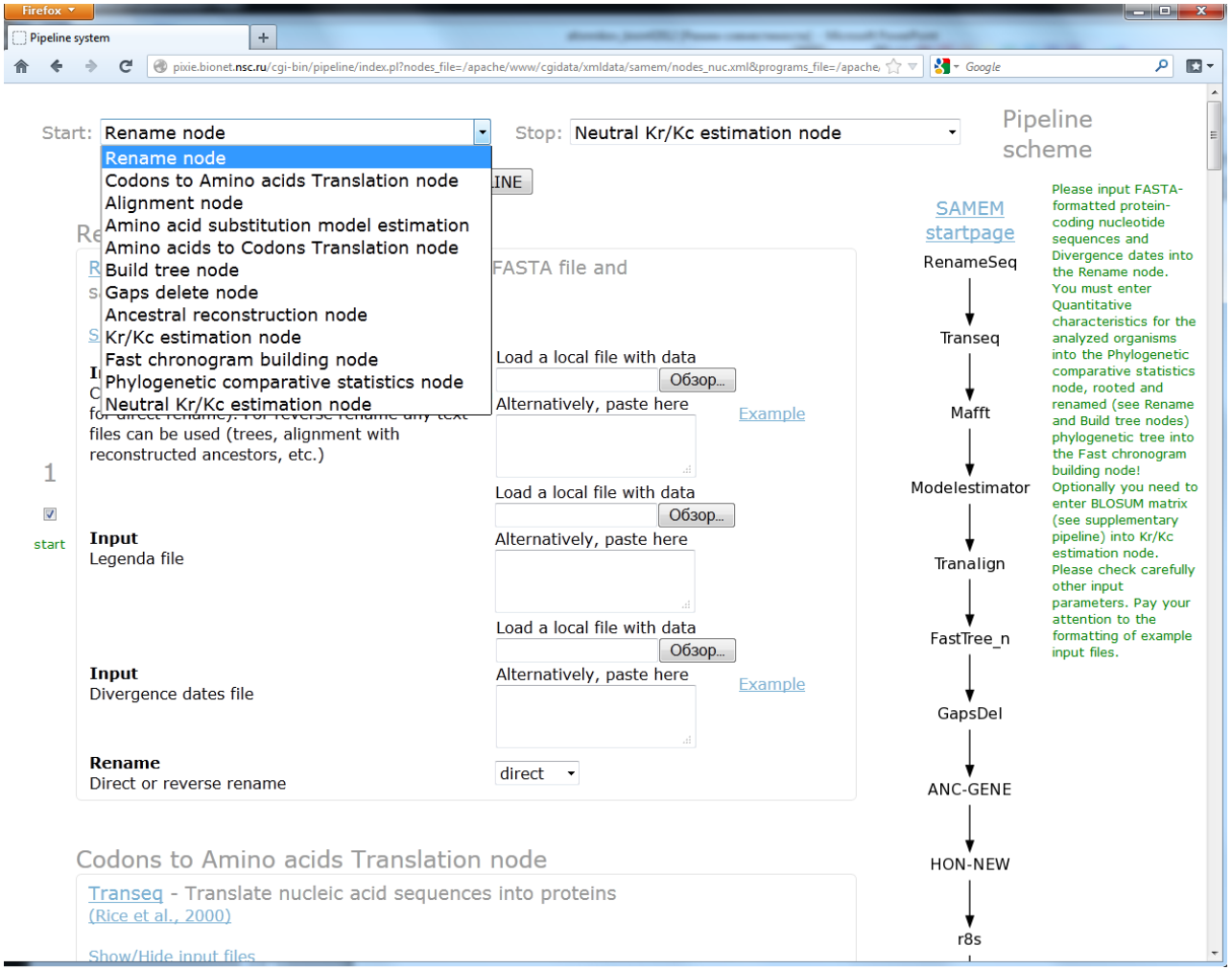
Гунбин К.В., Генаев М.А.



http://pixie.bionet.nsc.ru/samem/

Пакет программ SAMEM

Выполняет обработку данных по цепочке



http://pixie.bionet.nsc.ru/samem/