Выравнивание последовательностей. Выбор оптимальной модели эволюции нуклеотидов

Интерфейс MEGA

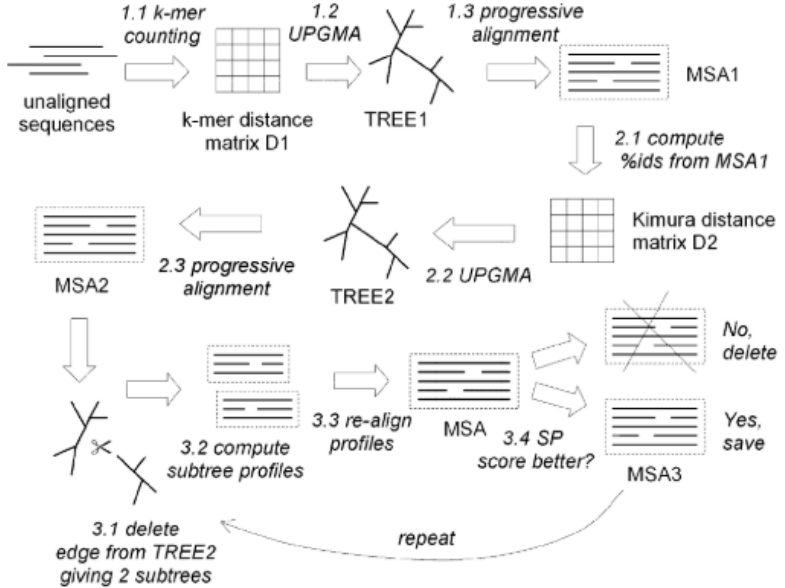
- ALIGN — выравнивание последовательностей
- DATA — загрузка и редактирование данных
- MODELS — выбор оптимальной модели эволюции нуклеотидов
- DISTANCE — расчёт матрицы расстояний по выравниванию
- PHYLOGENY — построение филогенетического дерева

Выравнивание последовательностей

Откройте файл `CoV_task.fa` для выравнивания.

1. DATA
2. Open File/Session
3. Выберите файл
4. Analyze or Align File? — Align

Алгоритм MUSCLE



Edgar, 2004

MUSCLE

- **Align DNA**
- Align Codons

Select all sequences

MUSCLE

- **Align DNA**
- Align Codons

Select all sequences

Option		Setting
GAP PENALTIES		
Gap Open	☑	*-400.00*
Gap Extend	☑	*0.00*
MEMORY/ITERATIONS		
Max Memory in MB	☑	*2048*
Max Iterations	☑	*16*
ADVANCED OPTIONS		
Cluster Method (Iterations 1,2)	☑	*UPGMA*
Cluster Method (Other Iterations)	☑	*UPGMA*
Min Diag Length (Lambda)	☑	*24*

Задание 1. Выполните выравнивание последовательностей с помощью алгоритма MUSCLE. Какова длина полученного выравнивания (порядковый номер последней колонки нуклеотидов, Site)?

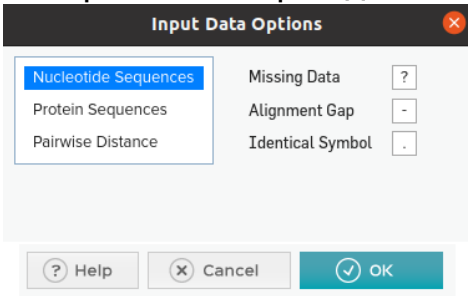
Задание 1. Выполните выравнивание последовательностей с помощью алгоритма MUSCLE. Какова длина полученного выравнивания (порядковый номер последней колонки нуклеотидов, Site)?

5256
Сохраните полученное выранивание в формате fasta:
Data/Export Alignment/FASTA format

Расчёт матрицы расстояний

Compute Pairwise Distance

- Выбор файла с выравниванием
- Настройка импорта данных

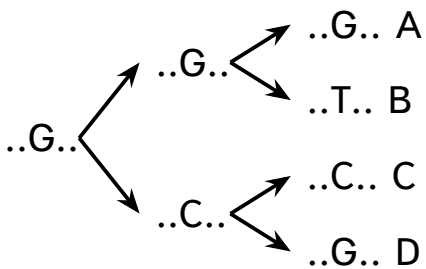


- Уточнение в выравнивании белок кодирующая последовательность или нет (выбор таблицы генетического кода)

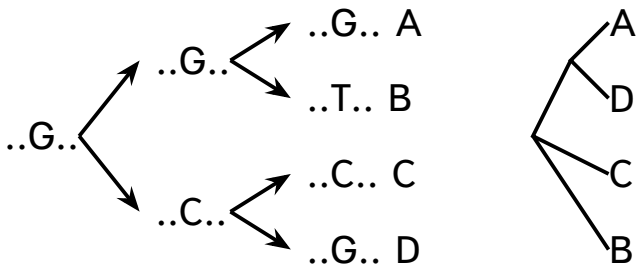
Расчёт матрицы расстояний

ANALYSIS		
Scope	→	*Pairs of taxa*
ESTIMATE VARIANCE		
Variance Estimation Method	→	*None*
No. of Bootstrap Replications	→	Not Applicable
SUBSTITUTION MODEL		
Substitutions Type	→	*Nucleotide*
Genetic Code Table	→	Not Applicable
Model/Method	→	*No. of differences*
Fixed Transition/Transversion Ratio	→	Not Applicable
Substitutions to Include	→	*d: Transitions + Transversions*
RATES AND PATTERNS		
Rates among Sites	→	*Uniform Rates*
Gamma Parameter	→	Not Applicable
Pattern among Lineages	→	*Same (Homogeneous)*
DATA SUBSET TO USE		
Gaps/Missing Data Treatment	→	*Pairwise deletion*
Site Coverage Cutoff (%)	→	Not Applicable
Select Codon Positions	→	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites

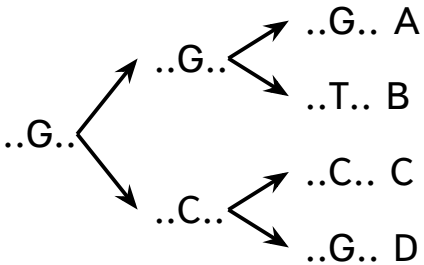
Несоответствие наблюдаемого и действительного расстояния

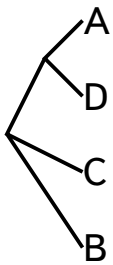


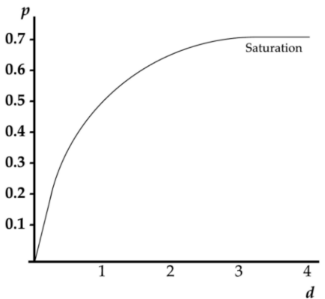
Несоответствие наблюдаемого и действительного расстояния



Несоответствие наблюдаемого и действительного расстояния







Lemey, 2009

Способы расчёта расстояний

- No. of differences (n_d)
- p.distance $p = n_d/L$
- Jukes-Cantor distance (JC69): $d = -\frac{3}{4} log_e(1 - \frac{4}{3}p)$
- Kimura 2-parameter distance (K80):
  $d = -0.5 log_e(w_1) - 0.25 log_e(w_2)$
  $w_1 = 1 - 2P - Q$
  $w_2 = 1 - 2Q$

Задание 2. Рассчитайте матрицу попарных расстояний для получившегося выравнивания. Какое максимальное расстояние между последовательностями (No. of differences)?

Способы расчёта расстояний

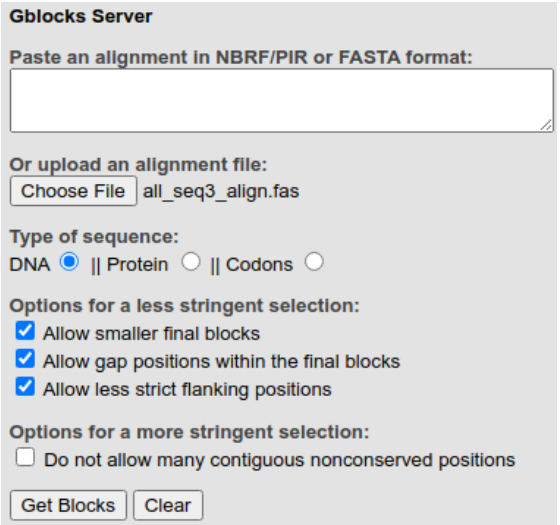
- No. of differences (n_d)
- p.distance $p = n_d/L$
- Jukes-Cantor distance (JC69): $d = -\frac{3}{4} log_e(1 - \frac{4}{3}p)$
- Kimura 2-parameter distance (K80):
  $d = -0.5 log_e(w_1) - 0.25 log_e(w_2)$
  $w_1 = 1 - 2P - Q$
  $w_2 = 1 - 2Q$

Задание 2. Рассчитайте матрицу попарных расстояний для получившегося выравнивания. Какое максимальное расстояние между последовательностями (No. of differences)?

2126

Обрезать или не обрезать выравнивание?

Gblocks



Задание 3. Обрежьте выравнивание с наиболее щадящими настройками.
Перед загрузкой на сервер Gblocks сохраните выравнивание в формате fasta: File/Export alignment/FASTA format

Processed file: **input.fasta**
Number of sequences: **34**
Alignment assumed to be: **DNA**
New number of positions: **2462** (selected positions are underlined in blue)

Сохраните обрезанное выравнивание (Cured alignment in FASTA Format).

Задание 3. Обрежьте выравнивание с наиболее щадящими настройками.
Перед загрузкой на сервер Gblocks сохраните выравнивание в формате fasta: File/Export alignment/FASTA format

Processed file: **input.fasta**
Number of sequences: **34**
Alignment assumed to be: **DNA**
New number of positions: **2462** (selected positions are underlined in blue)

Сохраните обрезанное выравнивание (Cured alignment in FASTA Format).

Задание 4. Повторите процедуру выравнивания для исходного файла, изменив значение Gap Open penalty на -200. Сравните длины выравниваний до и после обрезки Gblocks

Задание 4. Повторите процедуру выравнивания для исходного файла, изменив значение Gap Open penalty на -200. Сравните длины выравниваний до и после обрезки Gblocks

6291, 2120

Форматы выравниваний

FASTA

```
>A
ATGAAATATACACTTTTATTTT--
>B
ATGTTGGTGATATTGTTAATGTTA
>C
ATGTTTTTGATACTTTTAATTTCC
```

CLUSTAL

```
CLUSTAL O(1.2.4) multiple sequence
alignment

A    ATGAAATATACACTTTTATTTT--
B    ATGTTGGTGATATTGTTAATGTTA
C    ATGTTTTTGATACTTTTAATTTCC
     ***      * * * *** * *
```

PHYLIP

```
PHYLIP
3  24
A    ATGAAATATACACTTTTATTTT--
B    ATGTTGGTGATATTGTTAATGTTA
C    ATGTTTTTGATACTTTTAATTTCC
```

NEXUS

```
begin data;
dimensions ntax=3 nchar=24;
format interleave datatype=DNA missing=N gap=-
;
matrix
A  ATGAAATATACACTTTTATTTT--
B  ATGTTGGTGATATTGTTAATGTTA
C  ATGTTTTTGATACTTTTAATTTCC
;
end;
```

Модели эволюции нуклеотидов

Модель	Частоты нуклеотидов	Частоты переходов	Свободные параметры
JC69	равные	равные	0
K80	равные	$Ts \neq Tv$	1
T92	$AT \neq GC$	$Ts \neq Tv$	2
F81	неравные	равные	3
F84	неравные	$Ts \neq Tv$	4
GTR	неравные	неравны для всех переходов	8

Транзиции (*Ts*) — A-G или C-T. **Трансверзии** (*Tv*) — A-T, A-C, G-C, G-T.

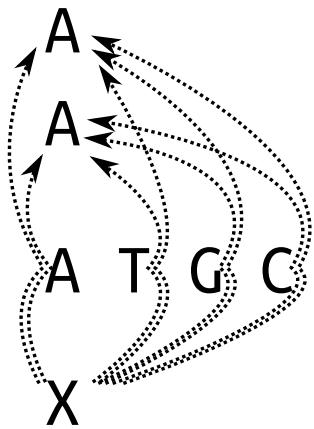
+G — различные частоты замен для сайтов

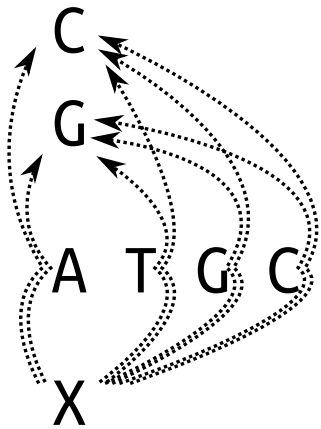
+I — доля инвариантных сайтов

Марковские цепи и выбор модели эволюции

Частоты переходов

Частоты нуклеотидов





Выбор оптимальной модели эволюции

Find best DNA/protein model

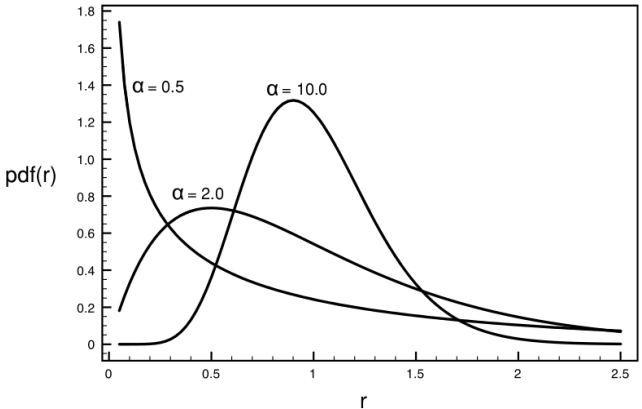
Option		Setting
ANALYSIS		
Tree to Use	→	*Automatic (Neighbor-joining tree)*
User Tree File	→	Not Applicable
Statistical Method	→	*Maximum Likelihood*
SUBSTITUTION MODEL		
Substitutions Type	→	Nucleotide
Genetic Code Table	→	Not Applicable
DATA SUBSET TO USE		
Gaps/Missing Data Treatment	→	*Use all sites*
Site Coverage Cutoff (%)	→	Not Applicable
Select Codon Positions	→	☑ 1st ☑ 2nd ☑ 3rd ☑ Noncoding Sites
Branch Swap Filter	→	*None*
SYSTEM RESOURCE USAGE		
Number of Threads	→	*3*

Просмотр оптимальных параметров

Запустите подбор модели, не меняя настройки по умолчанию. Используйте обрезанное выравнивание ДНК, которое было получено без модификации настроек, его длина 2462.

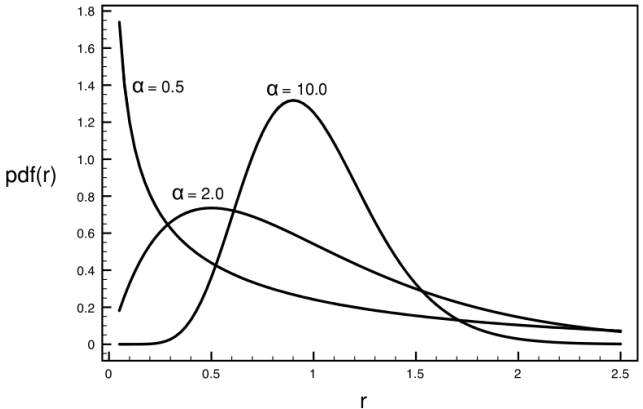
- Model — название модели
- Parameters — общее число параметров
- BIC, AIC — информационные критерии
- lnL — правдоподобие
- (+G)/(+I) — различия в частотах замен между сайтами и доля инвариантных сайтов
- R — Ts/Tv
- f(N) — частоты нуклеотидов
- f(NN) — частоты переходов

Гамма (G) распределение



Задание 5. Какая модель наилучшая?

Гамма (G) распределение



Задание 5. Какая модель наилучшая? GTR+G+I

Задание 6. Откройте последовательности из файла DBY_intron 7_part.fasta. Постройте выравнивание с помощью алгоритма muscle (учтите, что даны последовательности интронов). Обрежьте его в Gblocks при щадящих настройках. Подберите модель эволюции нуклеотидов.

Задание 6. Откройте последовательности из файла DBY_intron 7_part.fasta. Постройте выравнивание с помощью алгоритма muscle (учтите, что даны последовательности интронов). Обрежьте его в Gblocks при щадящих настройках. Подберите модель эволюции нуклеотидов.

T92+G