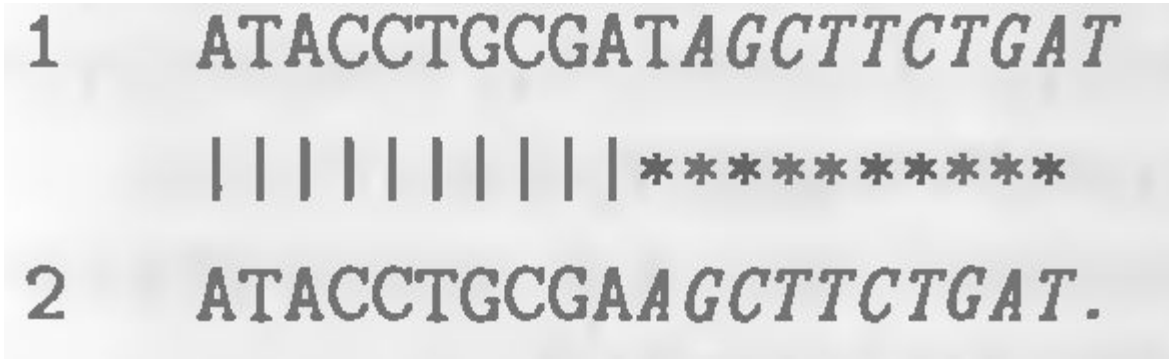
Лекция 9

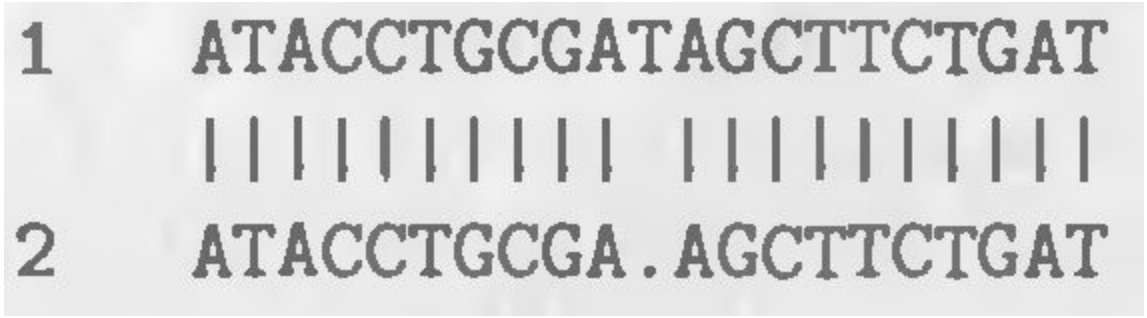
Алгоритмы и методы глобального выравнивания последовательностей. Множественное выравнивание. Рекомбинационный анализ

Выравнивание последовательностей —метод, основанный на размещении двух или более последовательностей ДНК, РНК или белков друг под другом таким образом, чтобы легко увидеть сходные участки в этих последовательностях. Сходство первичных структур двух молекул может отражать их функциональные, структурные или эволюционные взаимосвязи. Выровненные последовательности оснований нуклеотидов или аминокислот обычно представляются в виде строк матрицы. Добавляются разрывы между основаниями таким образом, чтобы одинаковые или похожие элементы были расположены в следующих друг за другом столбцах матрицы

Глобальное выравнивание

- Полагается, что последовательности обладают достаточным сходством по всей длине
- Можно разделить на: попарное (выравнивание двух последовательностей) и множественное (выравнивание трех и более)
- Самые распространенные: Clustal, T-Coffee, MAFFT и MUSCLE


1 ATACCTGCGATAGCTTCTGAT
|||||||||||**********
2 ATACCTGCGAAGCTTCTGAT.


1 ATACCTGCGATAGCTTCTGAT
|||||||||||.|||||||||||
2 ATACCTGCGA.AGCTTCTGAT

```
1   AGATCCGACTCTACG
    ||*||***|******
2   AGGTCGTTCAGACGT
```

```
1   AGATCCGACTCT.ACG.
    ||*||  **||* |||
2   AGGTC..GTTCAGACGT
```

```
1   AGATCCGACTCTACG.
    ||*||***|** |||
2   AGGTCGTTCAG.ACGT
```

```
1   AGATCCGACTCT.ACG.
    ||*|| |* ||* |||
2   AGGTC.GT.TCAGACGT
```

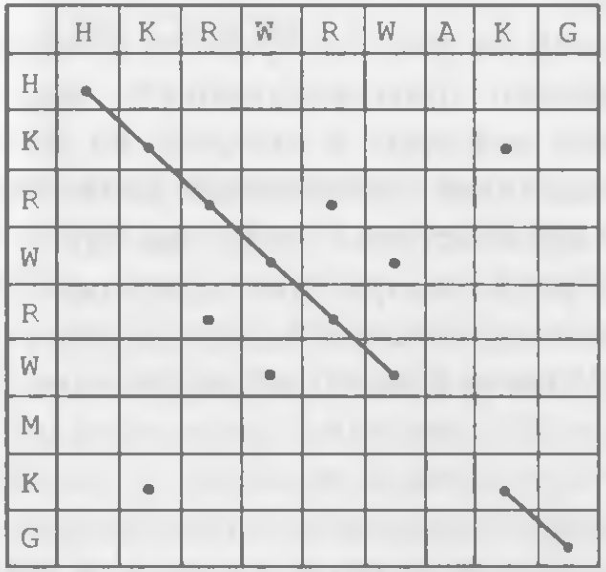
```
1   AGATCCGA.CTCTACG.
    ||*|| |* |* *|||
2   AGGTC.GTTCA.GACGT
```

```
1   AGATCCGACTCT.ACG.
    ||*|| | *||* |||
2   AGGTC.G.TTCAGACGT
```

Алгоритм Нидлмана — Вунша

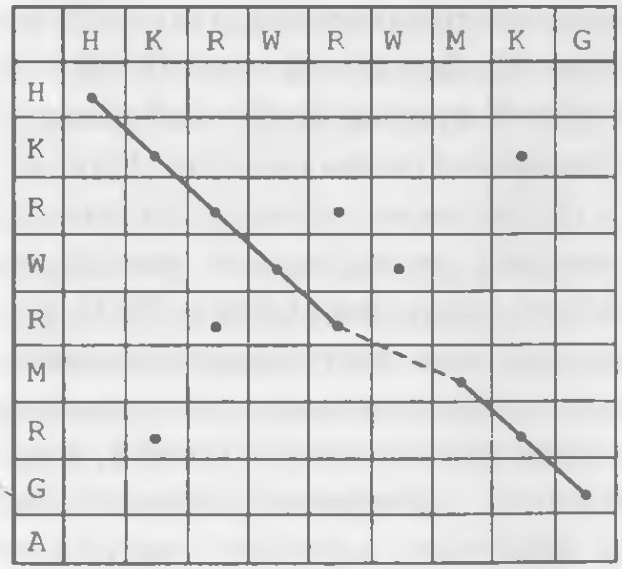
- Был предложен в 1970 году
- является примером динамического программирования, и он оказался первым примером приложения динамического программирования к сравнению биологических последовательностей

Графическое представление. Точечная матрица



последовательности
до и после выравнивания:

```
HKRWRWAKG
||||||*||
HKRWRWMKG
```



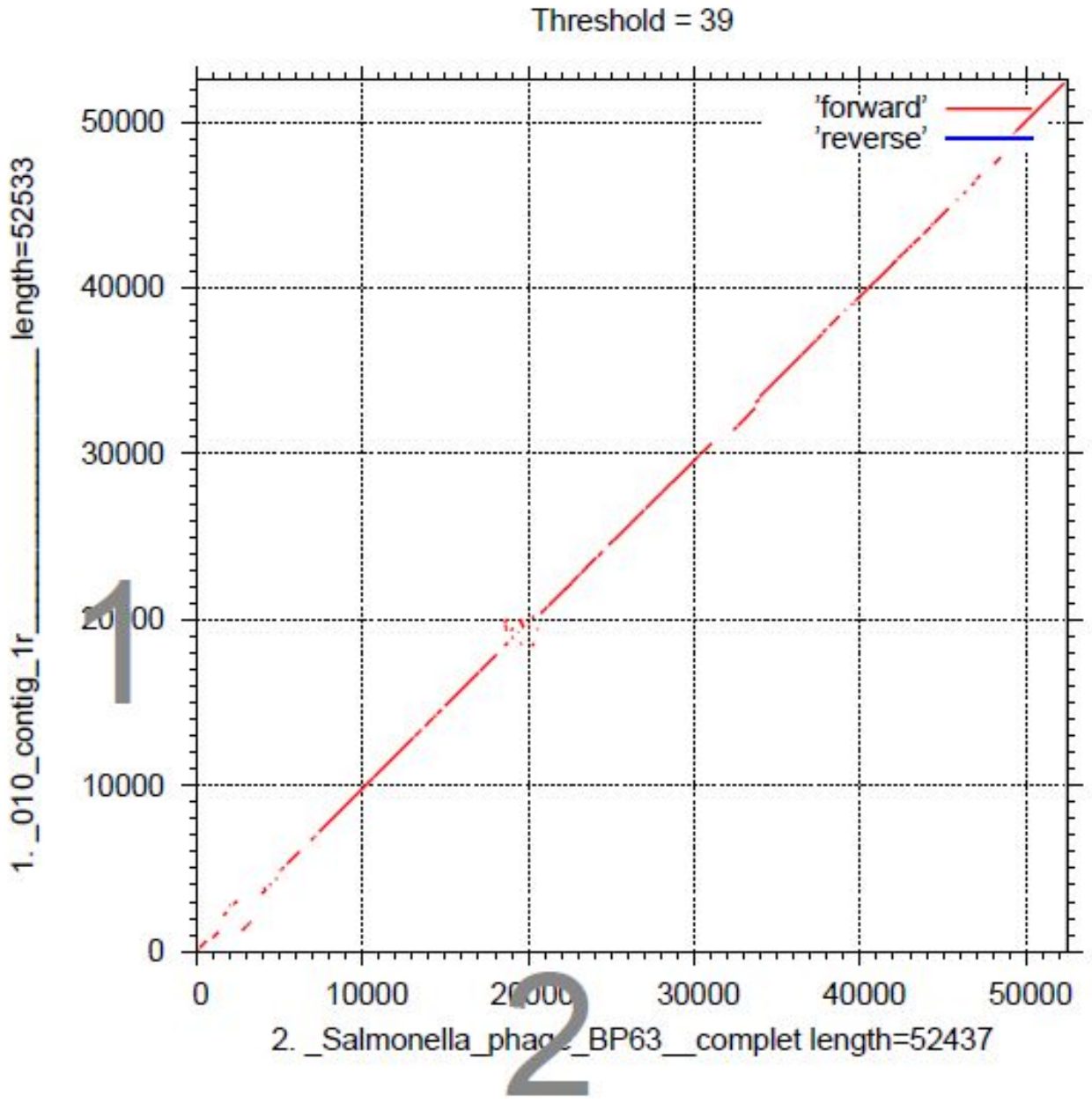
последовательности
до выравнивания:

```
HKRWRWMKG
|||||****
HKRWRMKGA
```

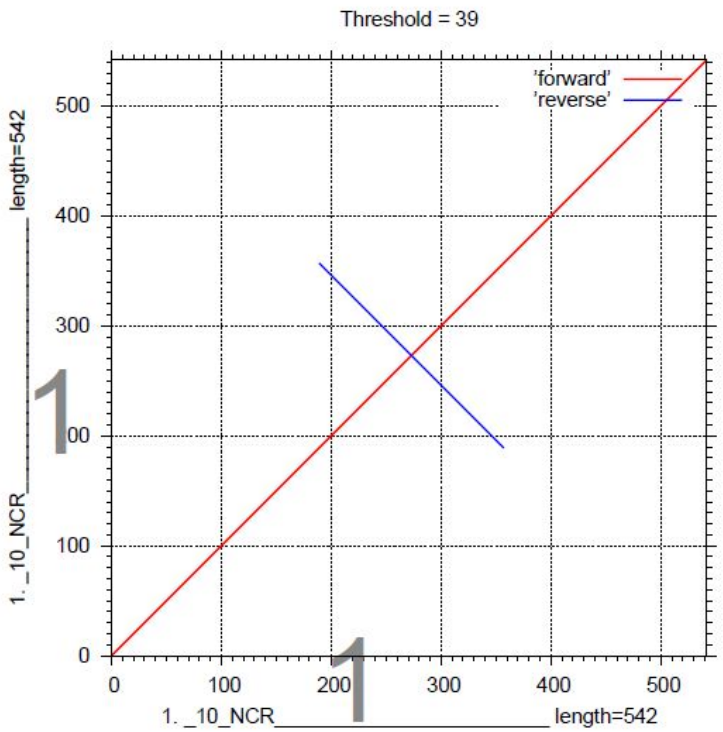
и после выравнивания:

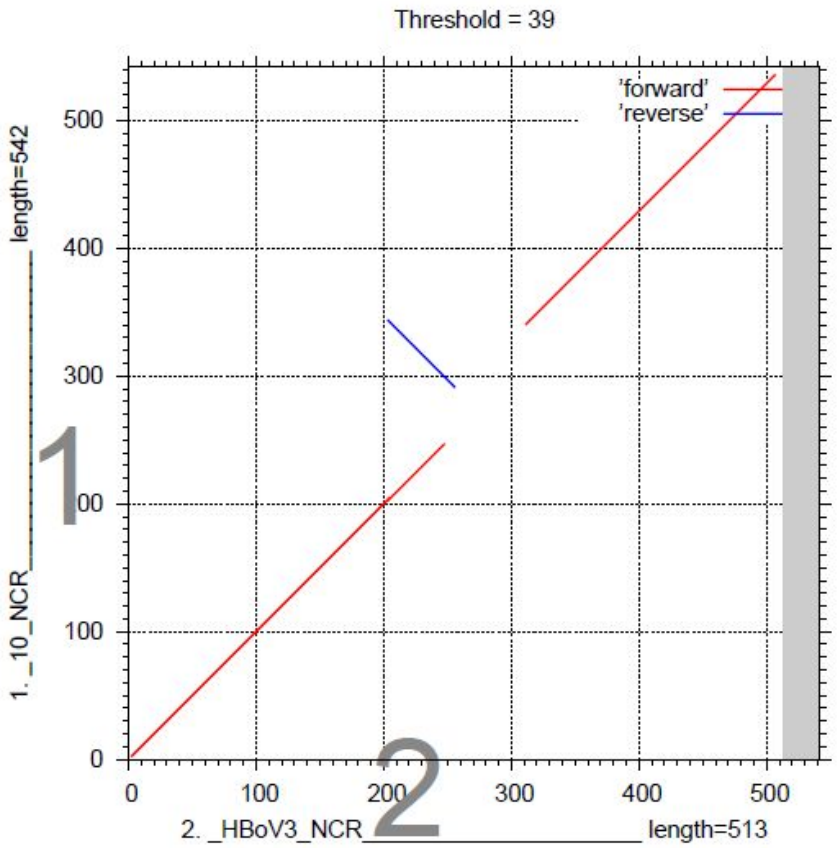
```
HKRWRWMKG.
|||||  |||
HKRWR.MKGA
```

Графическое представление. Точечная

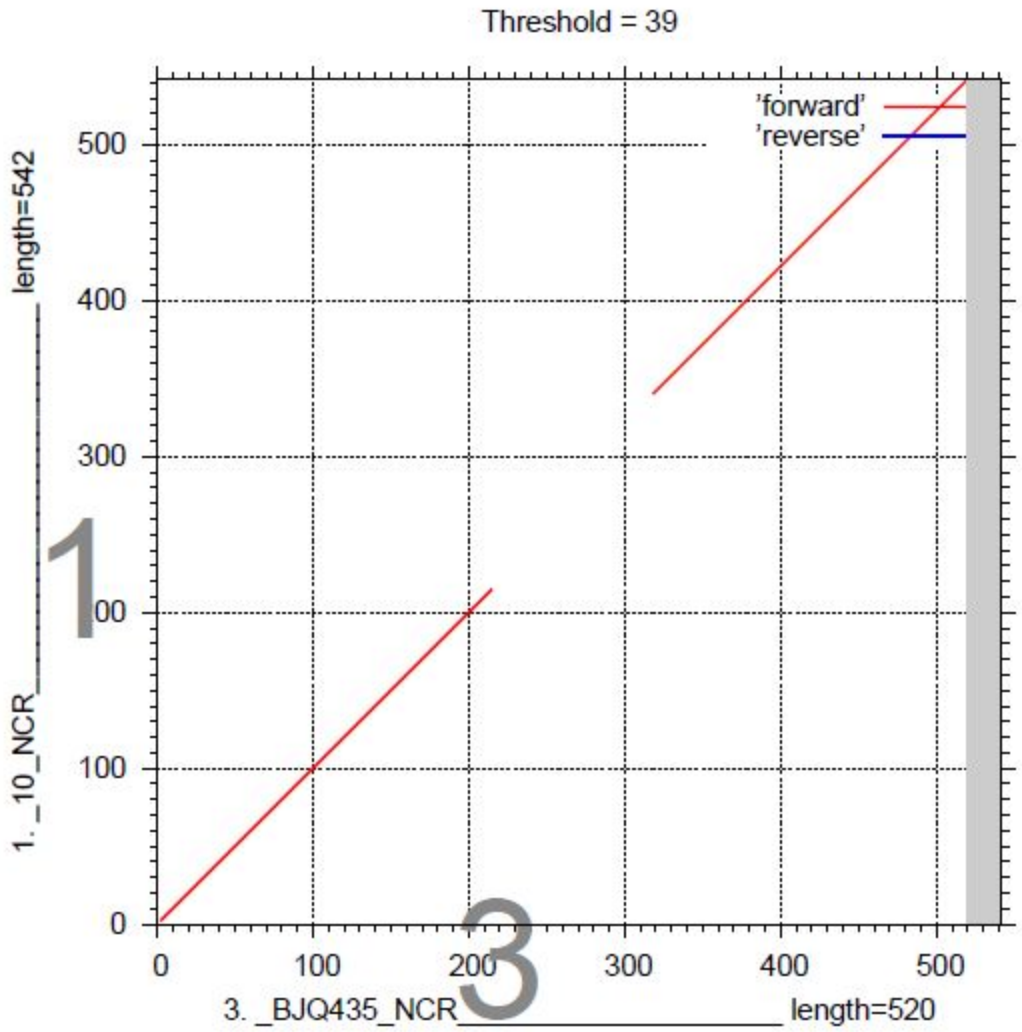


Графическое представление. Точечная ма

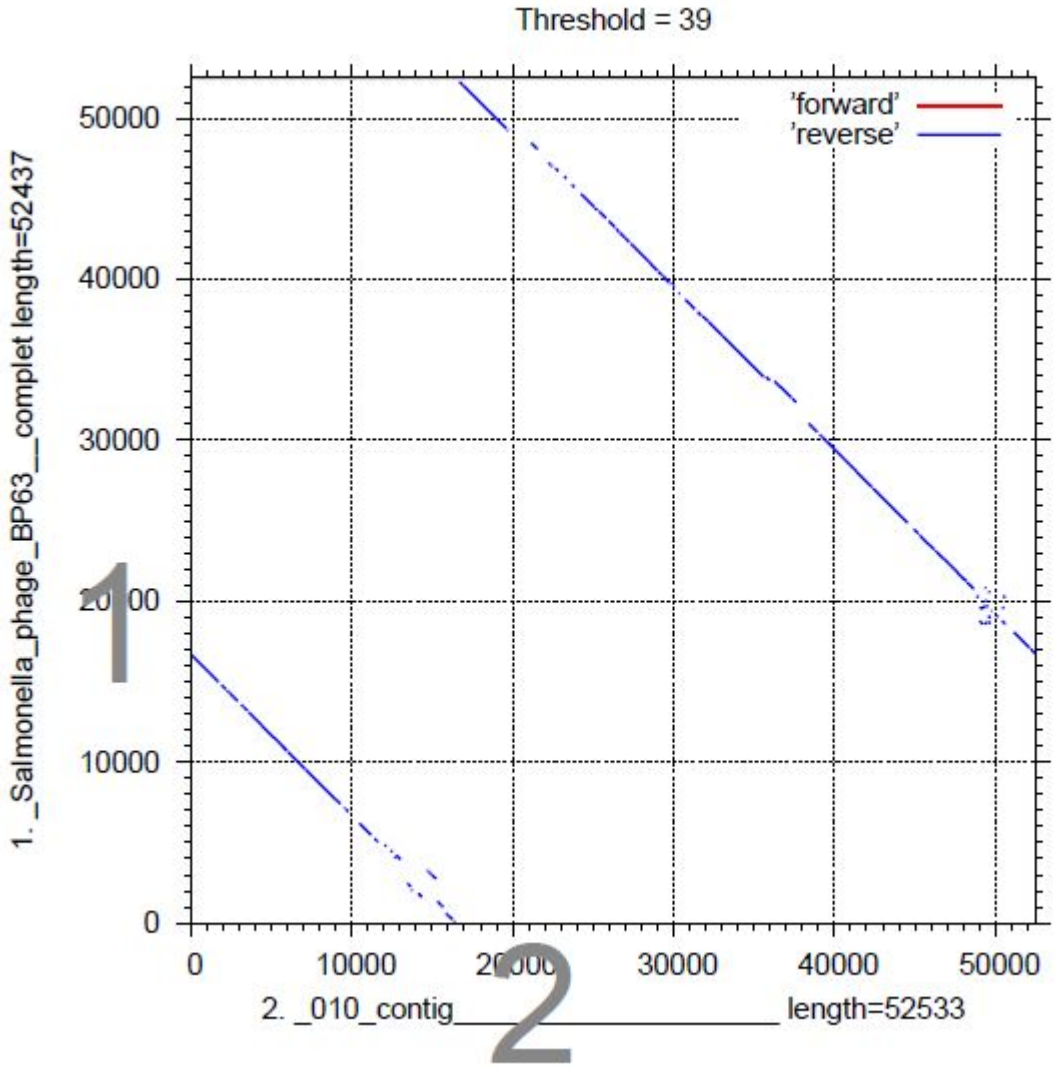




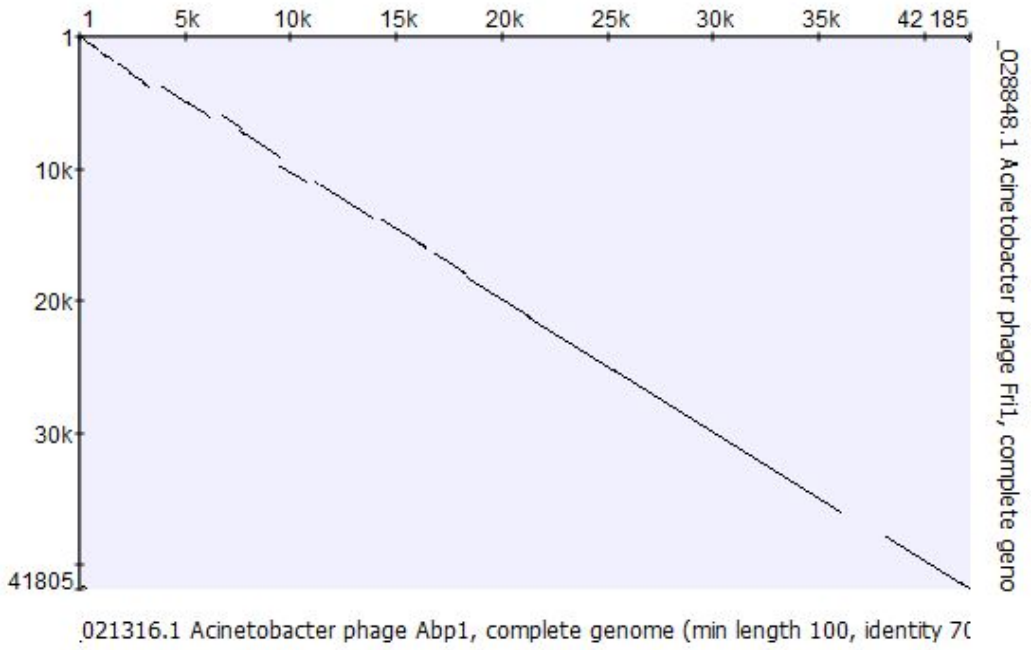
Графическое представление. Точечная

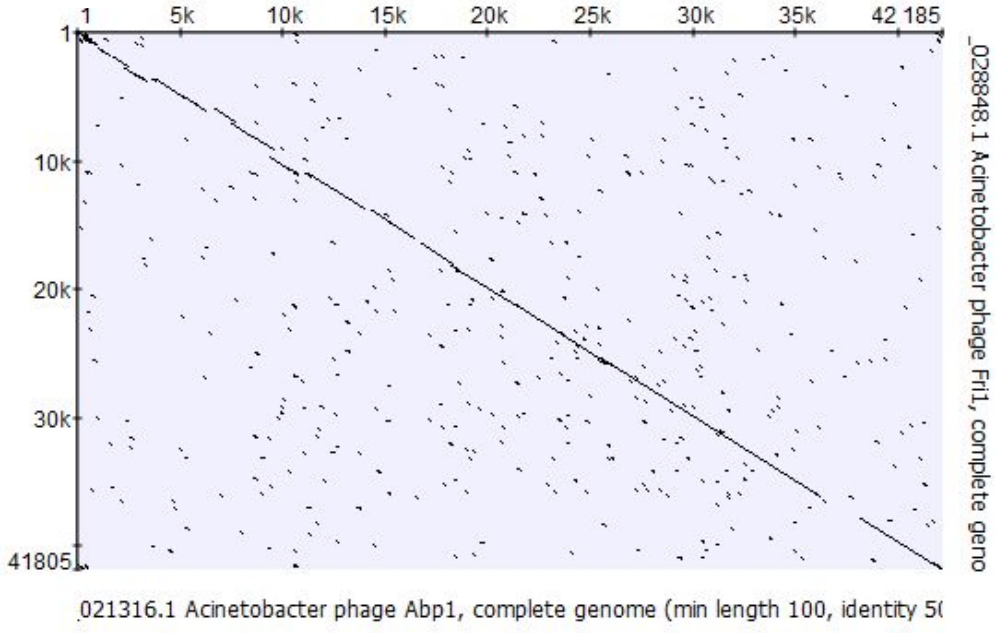


Графическое представление. Точечная матрица

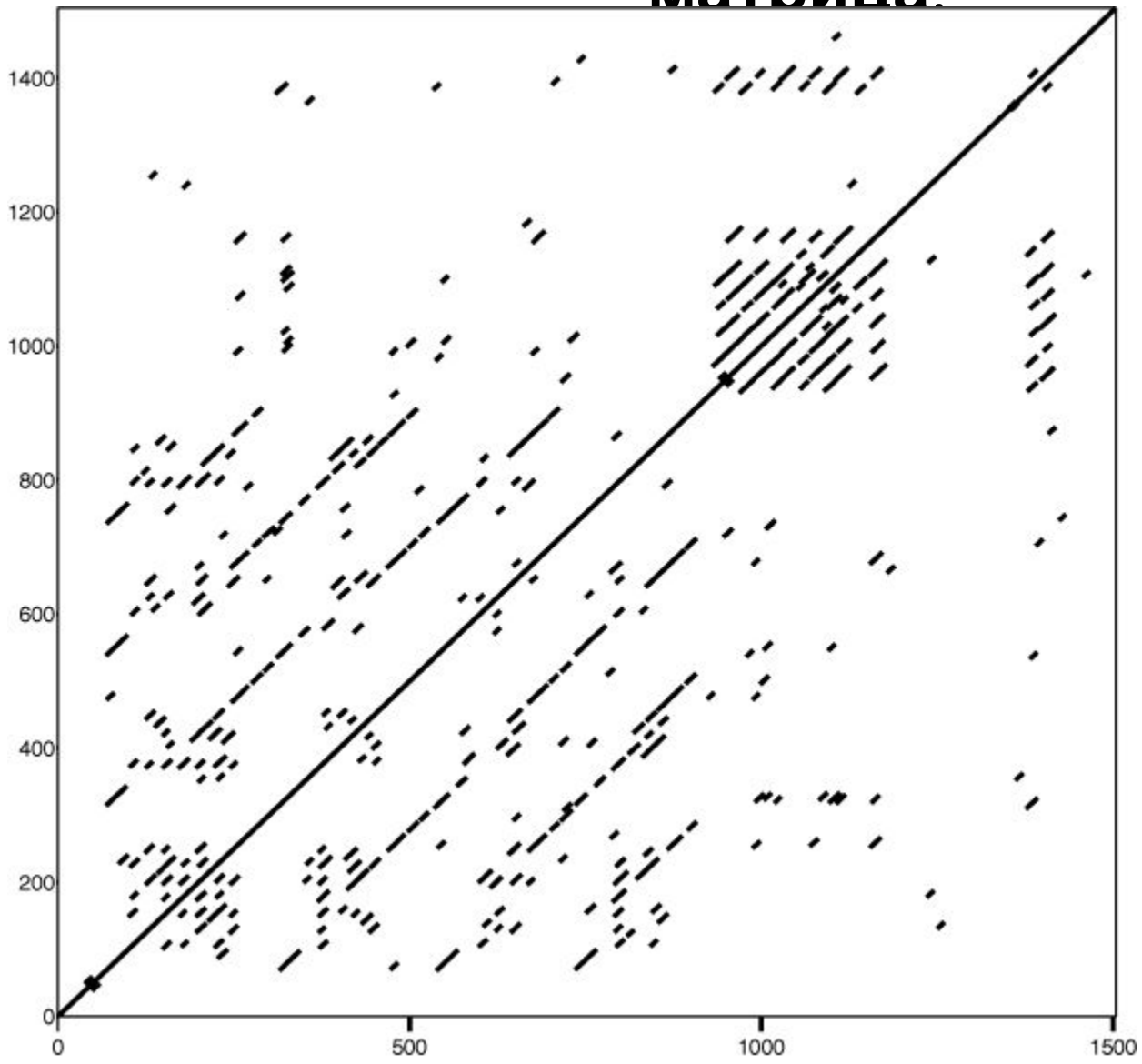


Графическое представление. Точечная ца.



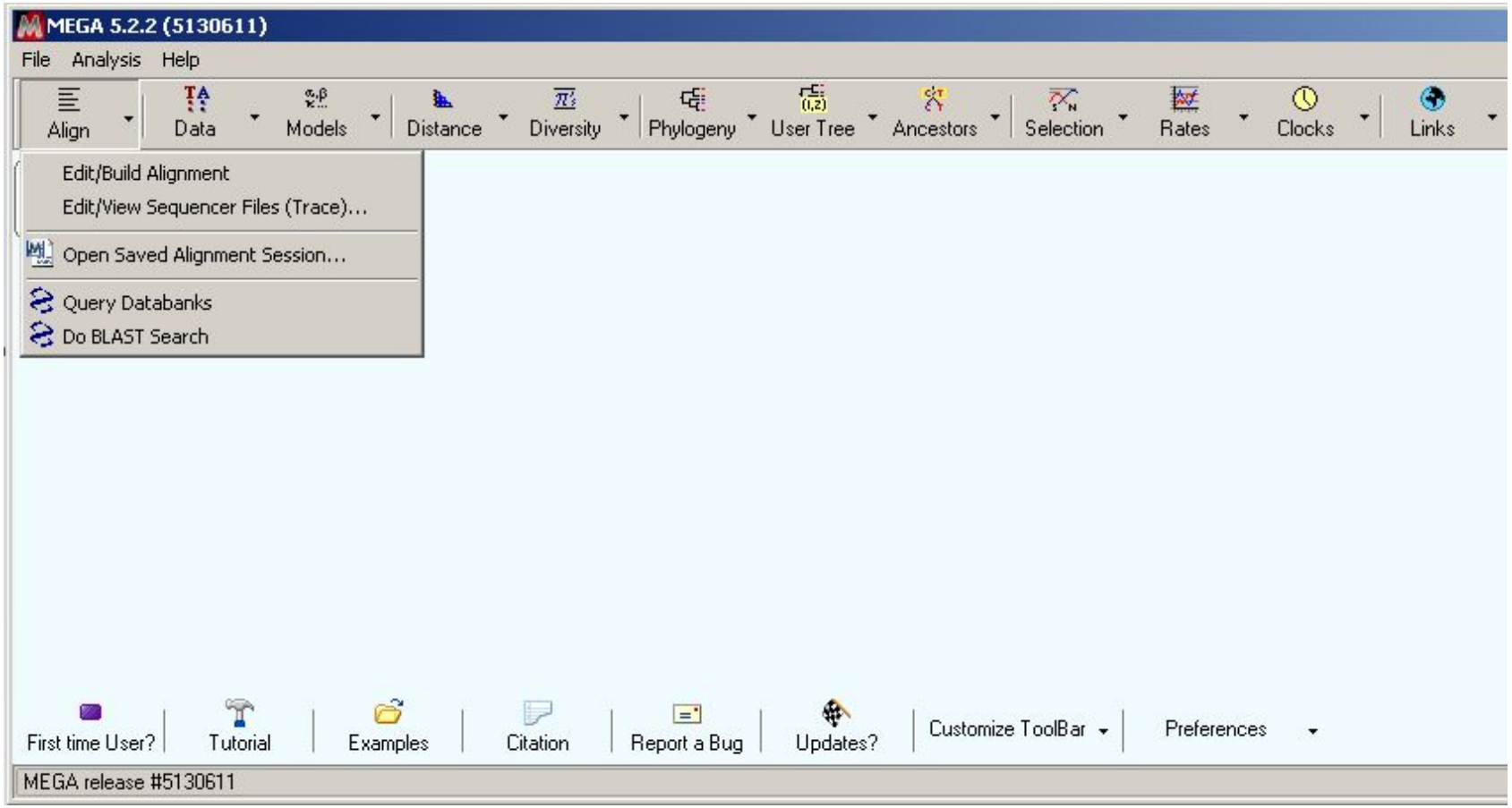


Графическое представление. Точечная матрица.

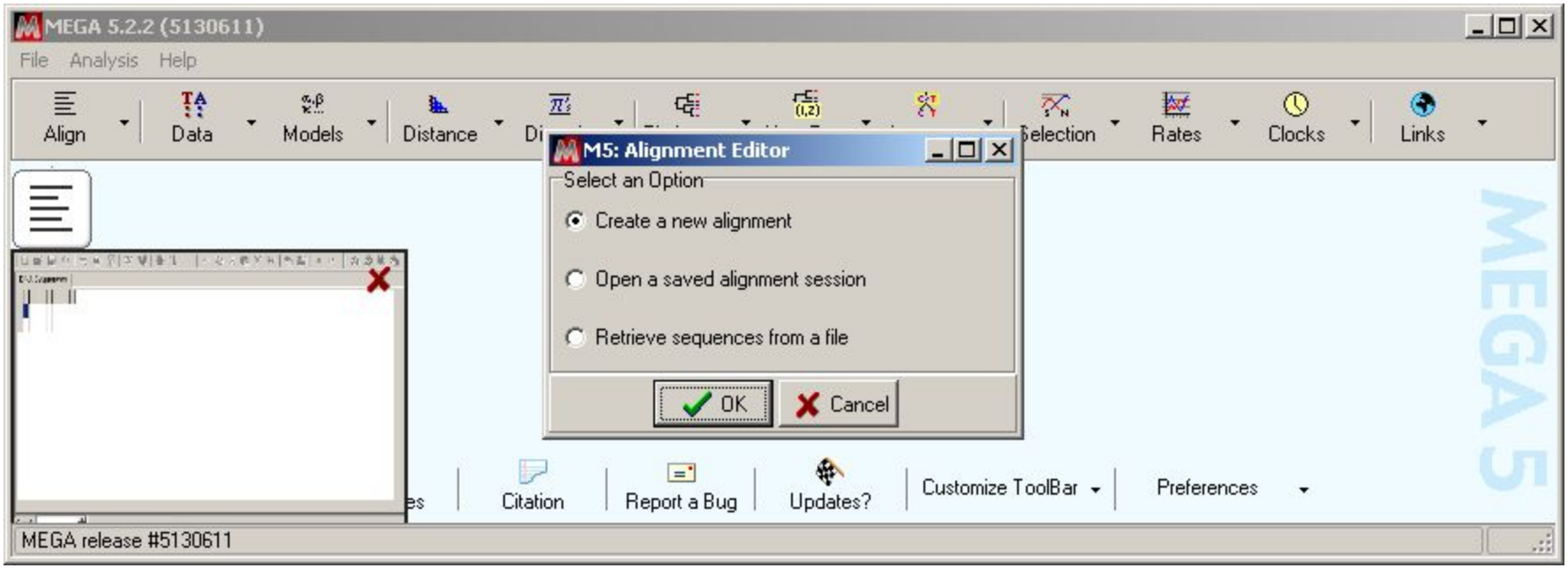


Dot Plot. Drosophila melanogaster SLIT protein aligned against itself.

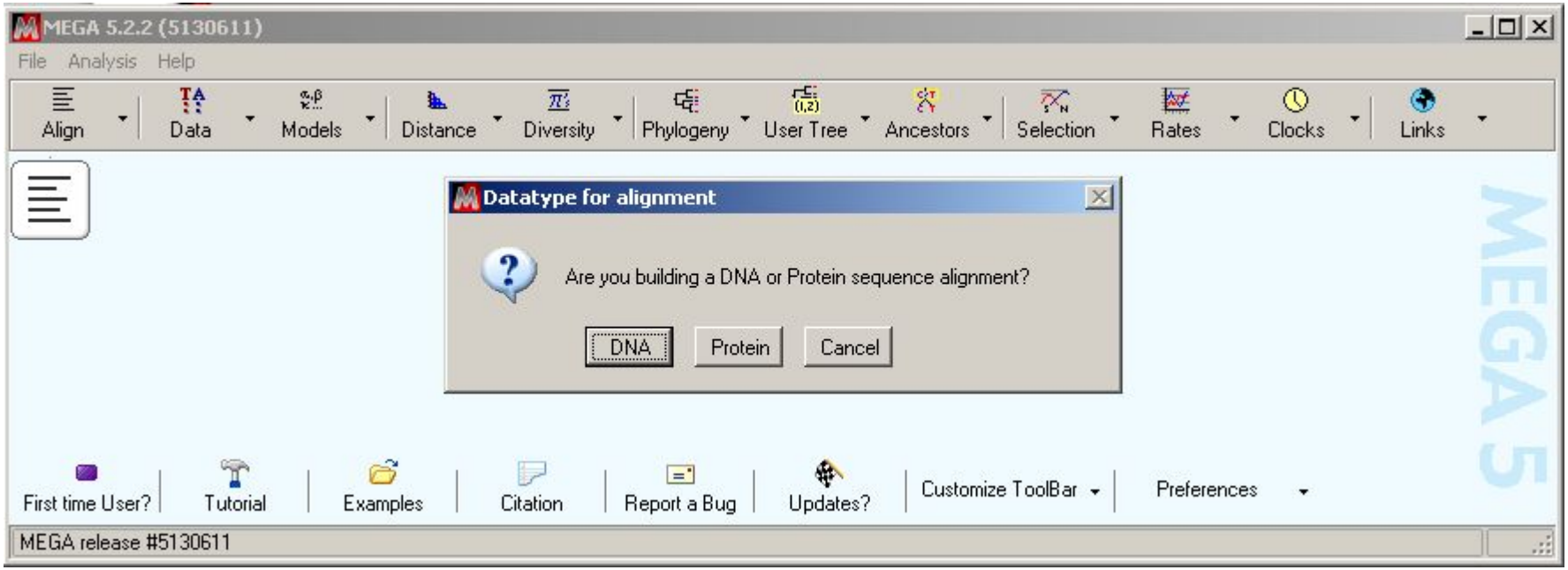
Создание выравнивания (1)



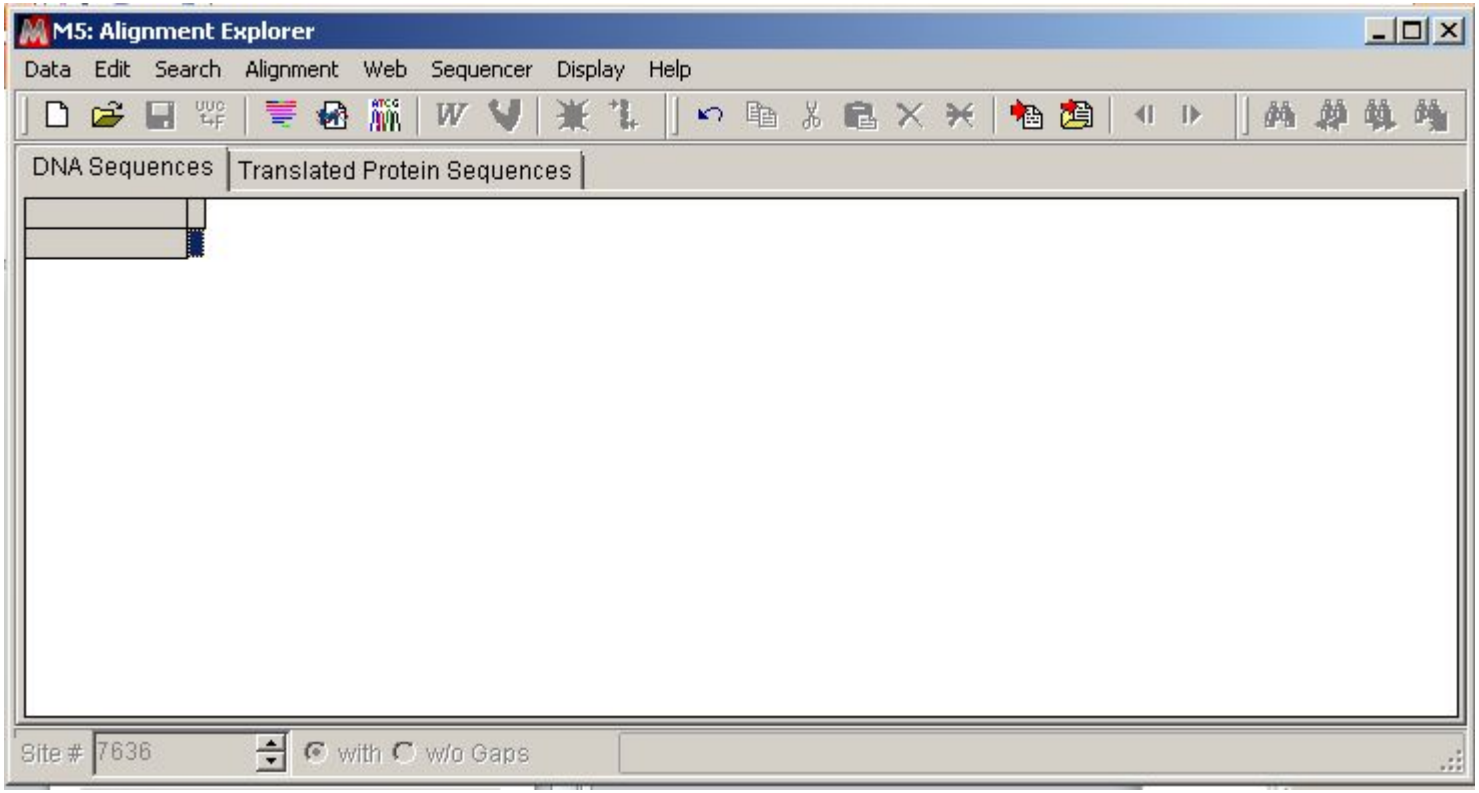
Создание выравнивания (2)



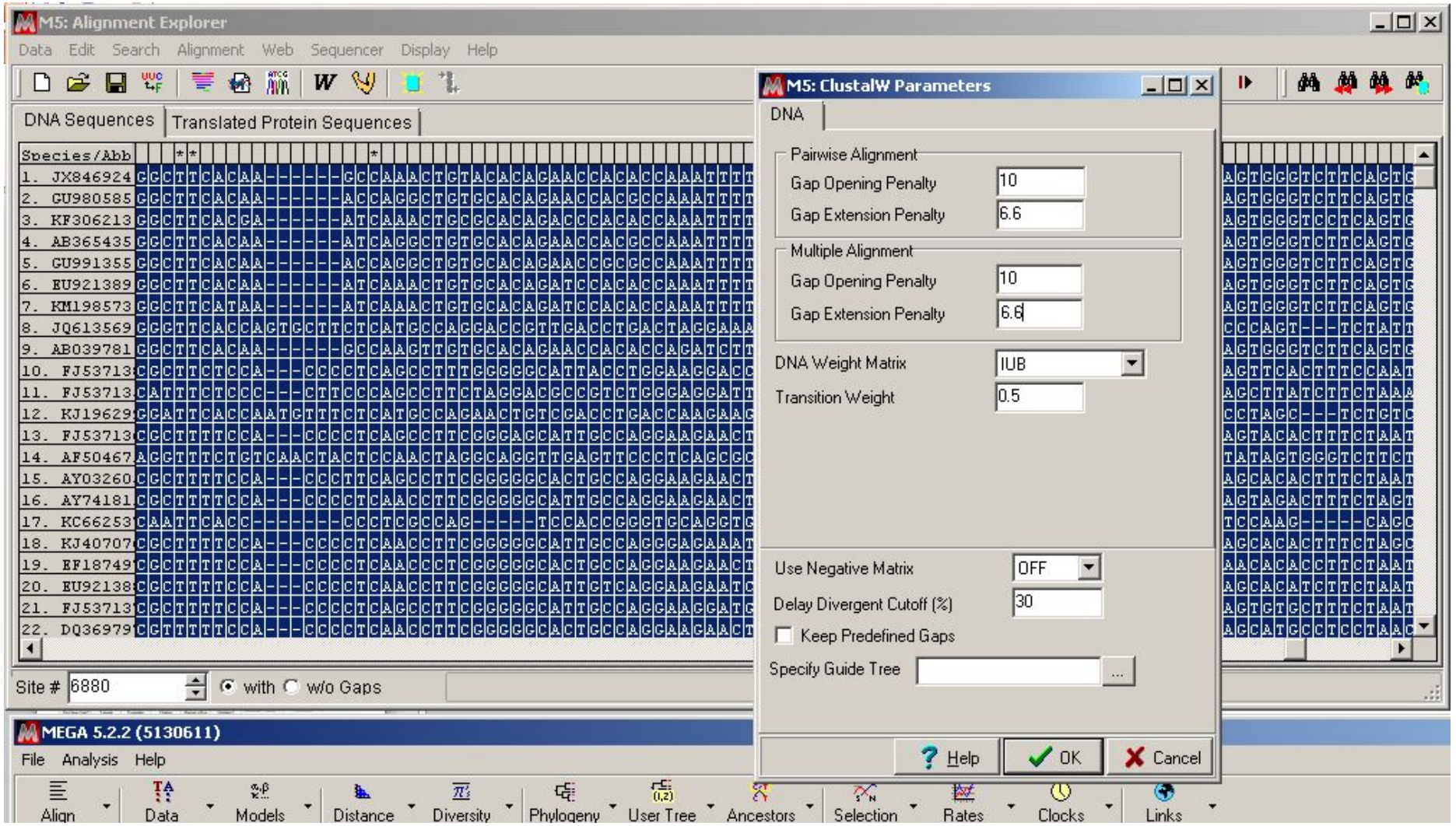
Создание выравнивания (3)



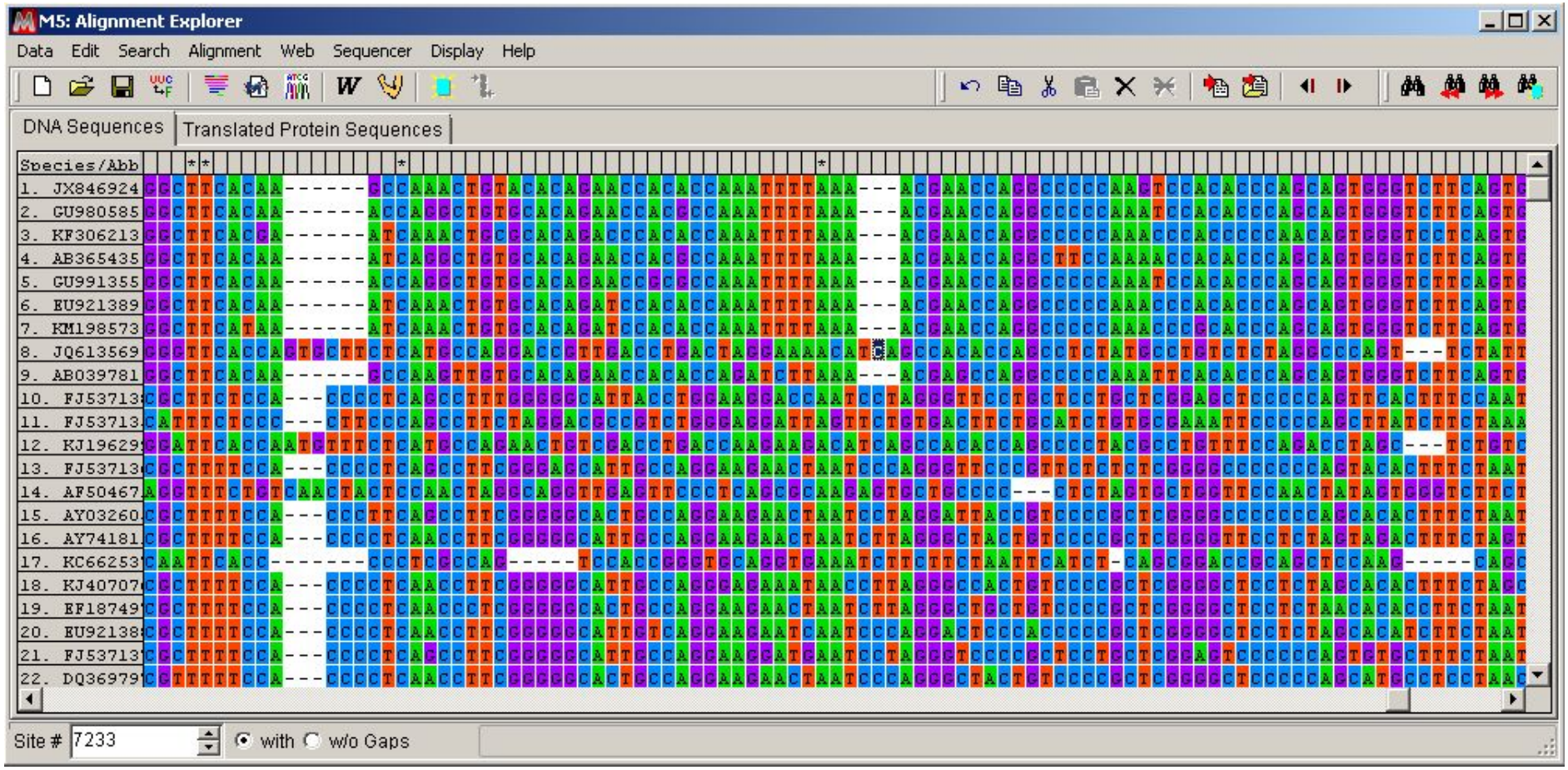
Создание выравнивания (4)



Параметры Clustal



Общий вид


M5: Alignment Explorer
Data Edit Search Alignment Web Sequencer Display Help
DNA Sequences
Translated Protein Sequences
Species/Abb
1. JX846924
2. GU980585
3. KF306213
4. AB365435
5. GU991355
6. EU921389
7. KM198573
8. JQ613569
9. AB039781
10. FJ53713
11. FJ53713
12. KJ19629
13. FJ53713
14. AF50467
15. AY03260
16. AY74181
17. KC66253
18. KJ40707
19. EF18749
20. EU92138
21. FJ53713
22. DQ36979
Site # 7233
with
w/o Gaps

MAFFT

https://mafft.cbrc.jp/

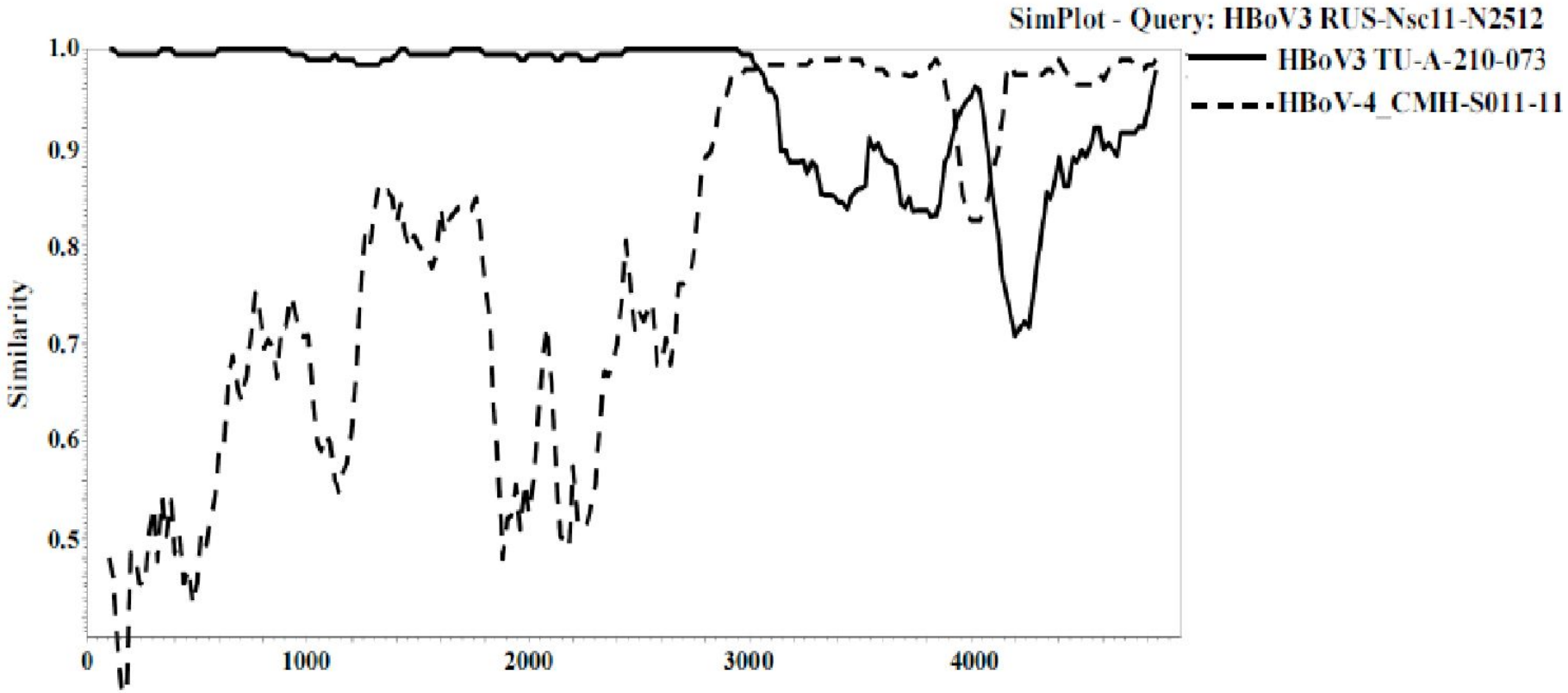
T-Coffee

http://www.tcoffee.org/

Рекомбинационный анализ

Simplot (Lole et al., 1999)

https://sray.med.som.jhmi.edu/SCRoftware/simplot/

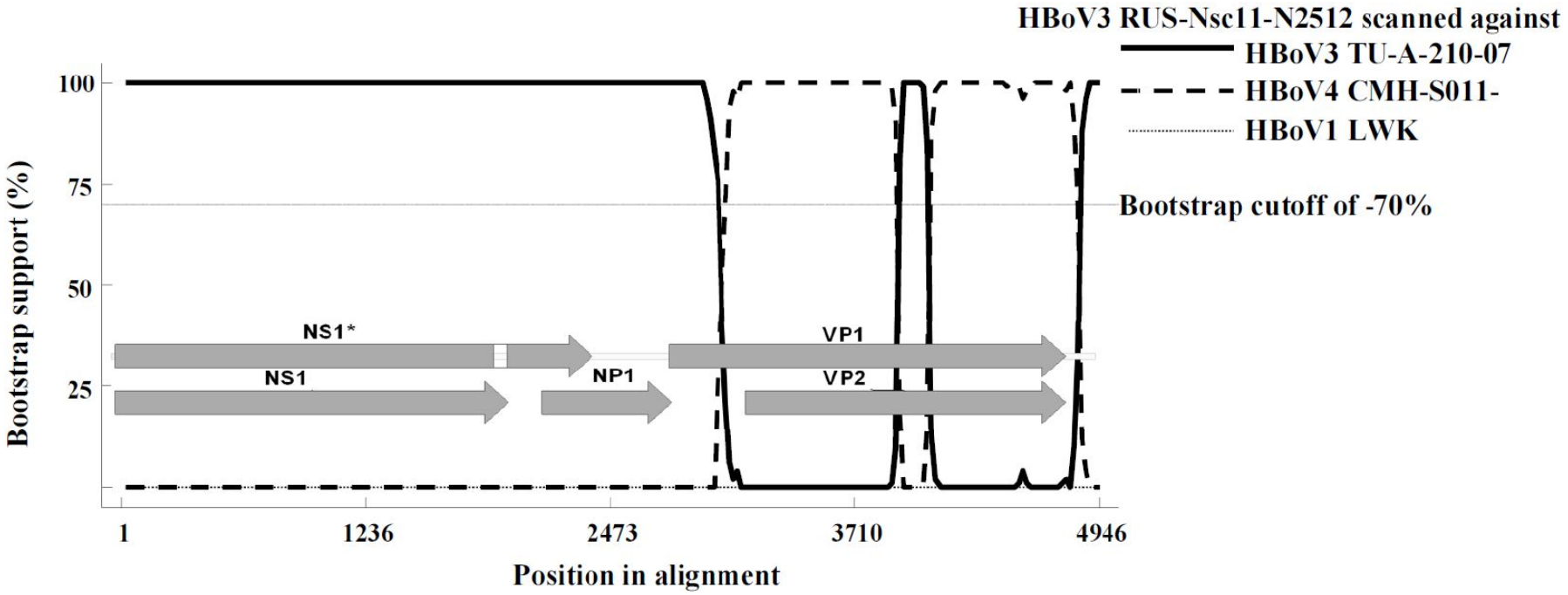


Analysis of HBoV genome sequences using Simplot (A) and Bootscan (B) methods. Gray arrows show HBoV ORFs.

Рекомбинационный анализ

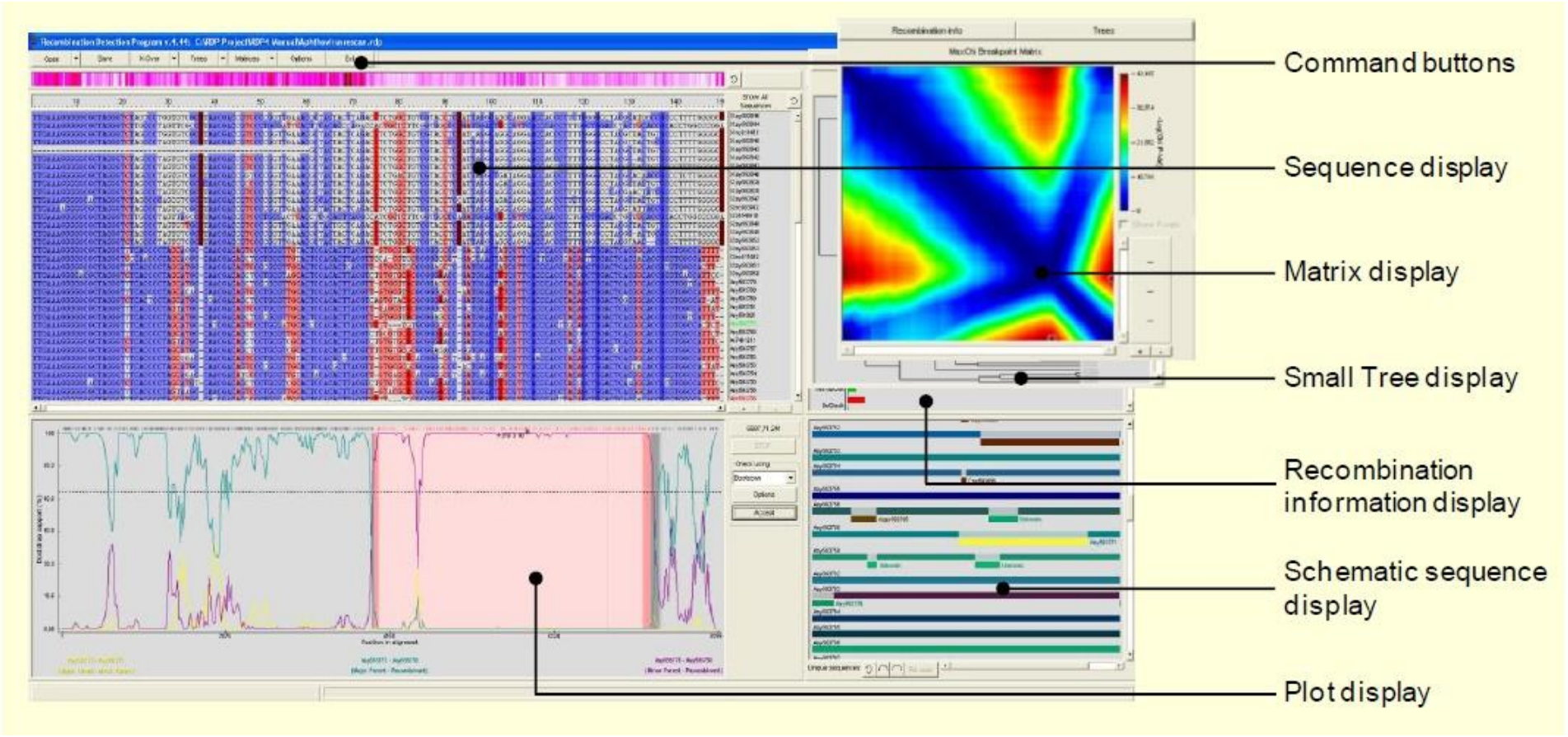
Simplot (Lole et al., 1999)

https://sray.med.som.jhmi.edu/SCRoftware/simplot/



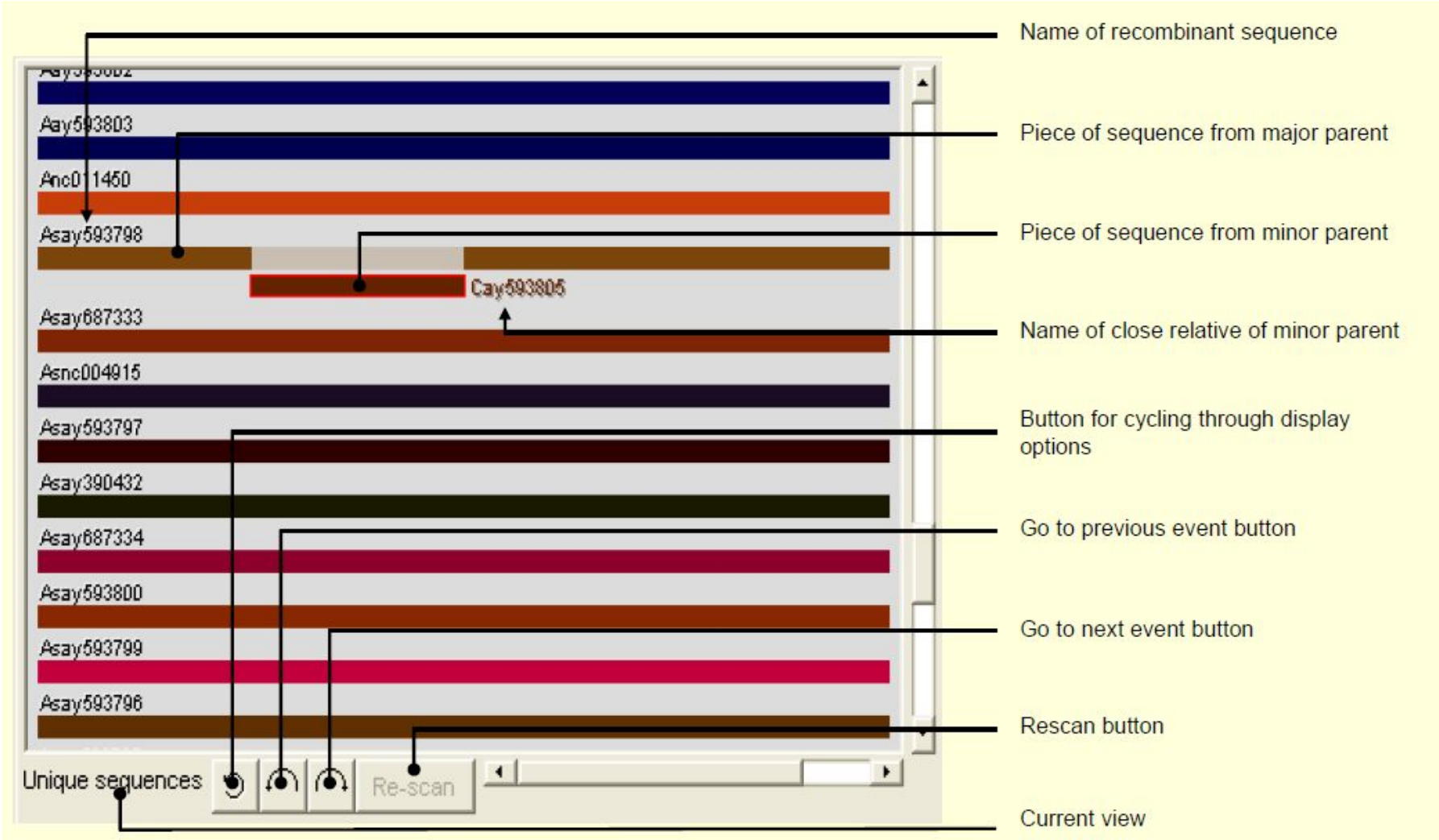
Analysis of HBoV genome sequences using Simplot (A) and Bootscan (B) methods. Gray arrows show HBoV ORFs.

RDP – (http://web.cbio.uct.ac.za)



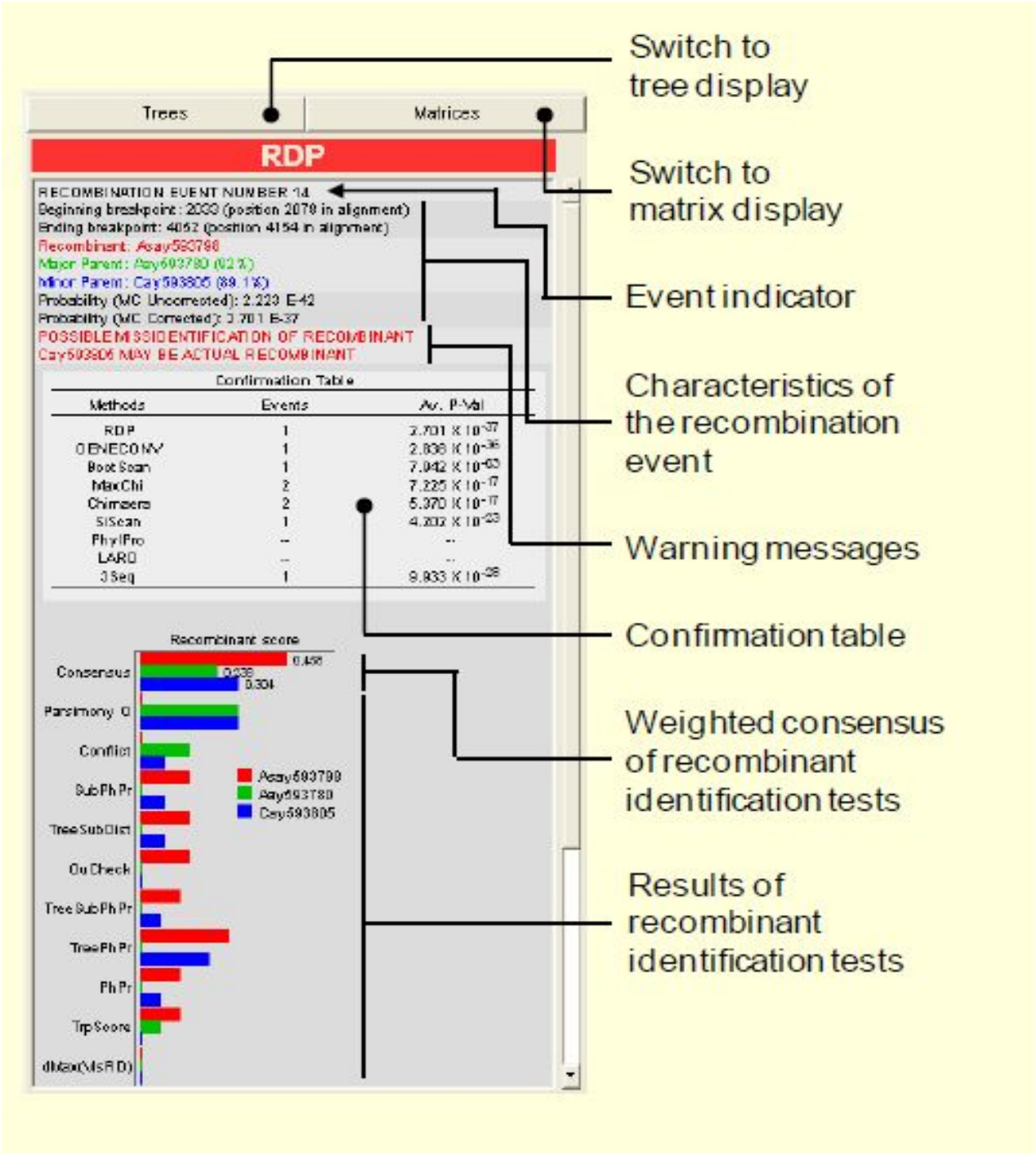
Bootscan, Chimaera, GENECONV, MaxChi, RDP, SisScan и др. методы

RDP – (http://web.cbio.uct.ac.za)

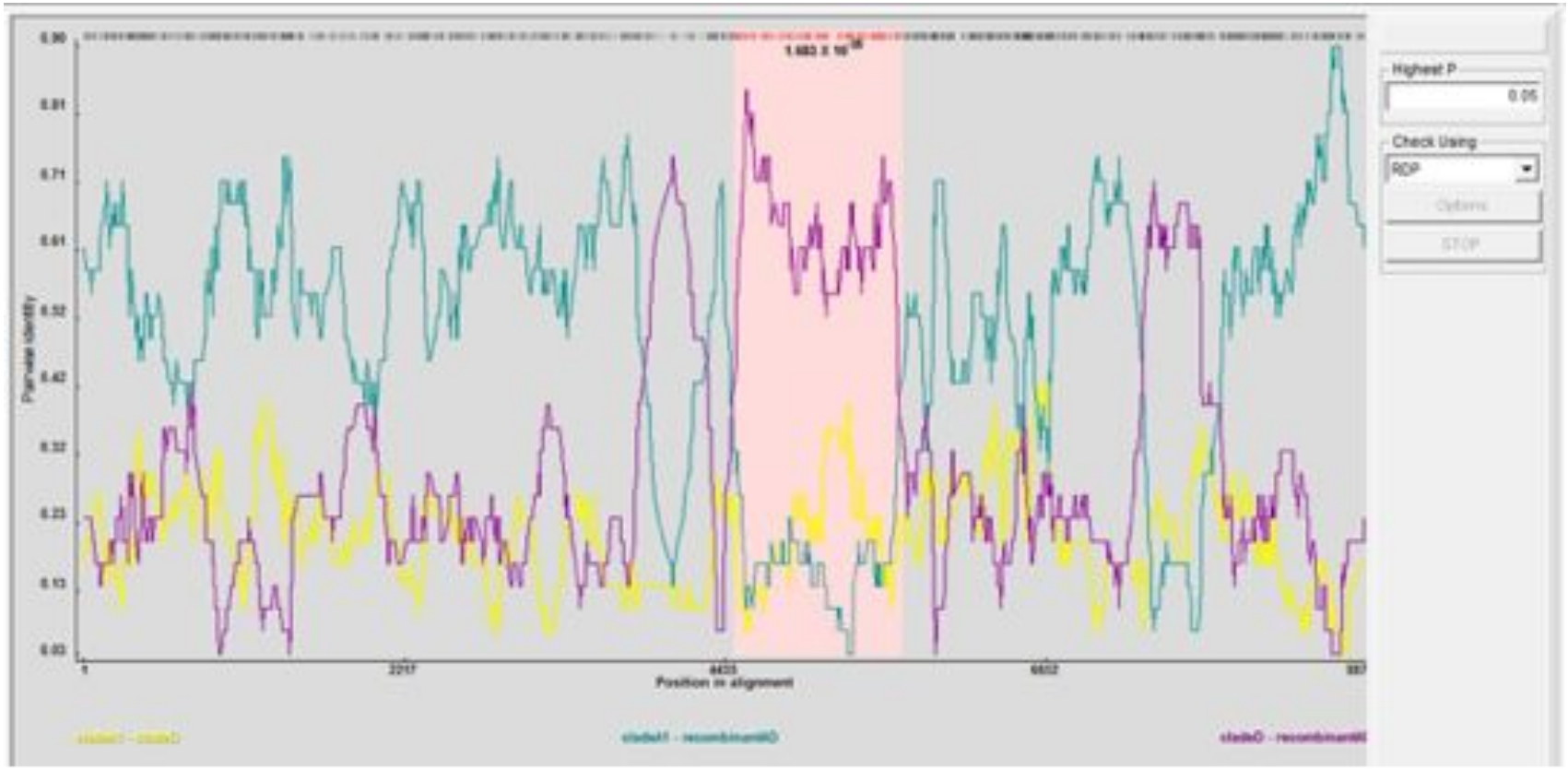


Bootscan, Chimaera, GENECONV, MaxChi, RDP, SisScan и др. методы

RDP
–


Switch to tree display
Switch to matrix display
Event indicator
Characteristics of the recombination event
Warning messages
Confimation table
Weighted consensus of recombinant identification tests
Results of recombinant identification tests
Trees
Matrices
RDP
RECOMBINATION EVENT NUMBER 14
Beginning breakpoint: 2033 (position 2079 in alignment)
Ending breakpoint: 4052 (position 4154 in alignment)
Recombinant: Asay593798
Major Parent: Asy593780 (92%)
Minor Parent: Cay593805 (89.1%)
Probability (MC Uncorrected): 2.223 E-42
Probability (MC Corrected): 3.701 E-37
POSSIBLE MISIDENTIFICATION OF RECOMBINANT
Cay593806 MAY BE ACTUAL RECOMBINANT
Confirmation Table
Methods
Events
Av. P-Val
RDP
1
2.701 X 10-37
GENECONV
1
2.836 X 10-36
Bootscan
1
7.042 X 10-33
MaxChi
2
7.225 X 10-17
Chimaera
2
5.370 X 10-17
SiScan
1
4.202 X 10-23
PhylPro
--
--
LARD
--
--
3Seq
1
9.833 X 10-28
Recombinant score
Consensus
0.458
0.039
0.304
Parsimony
Conflict
SubPhPr
TreeSubDist
OuCheck
TreeSubPhPr
TreePhPr
PhPr
TripScore
dMax(MsRD)
Asay593798
Asay593780
Cay593805

RDP – (можно скачать с сайта http://web.cbio.uct.ac.za)



Preparing your data Examples: Influenza A H5N1 hemagluttinin | HIV-1 pol (recombinant data)

Choose a sequence alignment (data formats): Выберите файл HAstV.fas

Please note that **all** selection analyses require a coding alignment. See Data type for the list of Datamonkey analyses and data types that they can accept. To ensure that no single job takes too long to run, there also are alignment size restrictions based on the analysis type.

Data type Nucleotide (DNA/RNA)

If you suspect that your data may contain recombinant sequences, please run a recombination (SBP or GARD) screen prior to performing selection analyses on Datamonkey. **Recombination can mislead** selection analyses if it is not accounted for!

Click to Upload

RAPID DETECTION OF POSITIVE SELECTION

a Web-Server of the HyPhy Package

SUCCESSFUL FILE UPLOAD

Read **16** sequences and **6858** nucleotide alignment columns and **1** partitions.

Nucleotide composition

A 30.2108%
C 21.5166%
G 22.4216%
T 25.851%

8 sequences were renamed to conform to HyPhy standards. You can look at the renamed alignment in NEXUS format for reference.

```
AF141381(HAstV-3)→AF141381_HASTV_3_
HM237363(HastV-6)→HM237363_HASTV_6_
GQ495608(HAstV-6)→GQ495608_HASTV_6_
FJ755402(HAstV-1)→FJ755402_HASTV_1_
FJ755405(HAstV-1)→FJ755405_HASTV_1_
DQ344027(HAstv-4)→DQ344027_HASTV_4_
DQ028633(HAstV-5)→DQ028633_HASTV_5_
FJ375759(HAstV-1)→FJ375759_HASTV_1_
```

BLAST your sequences? AF141381_HASTV_3_ ▾ BLAST Away

JOB ID:UPLOAD.868814535518859.1 [INFORMATION:OTHER ANALYSES]

Proceed to the analysis menu

Analysis Options

Job ID: upload.868814535518859.1 [Information: Other analyses]

Method: GARD Help

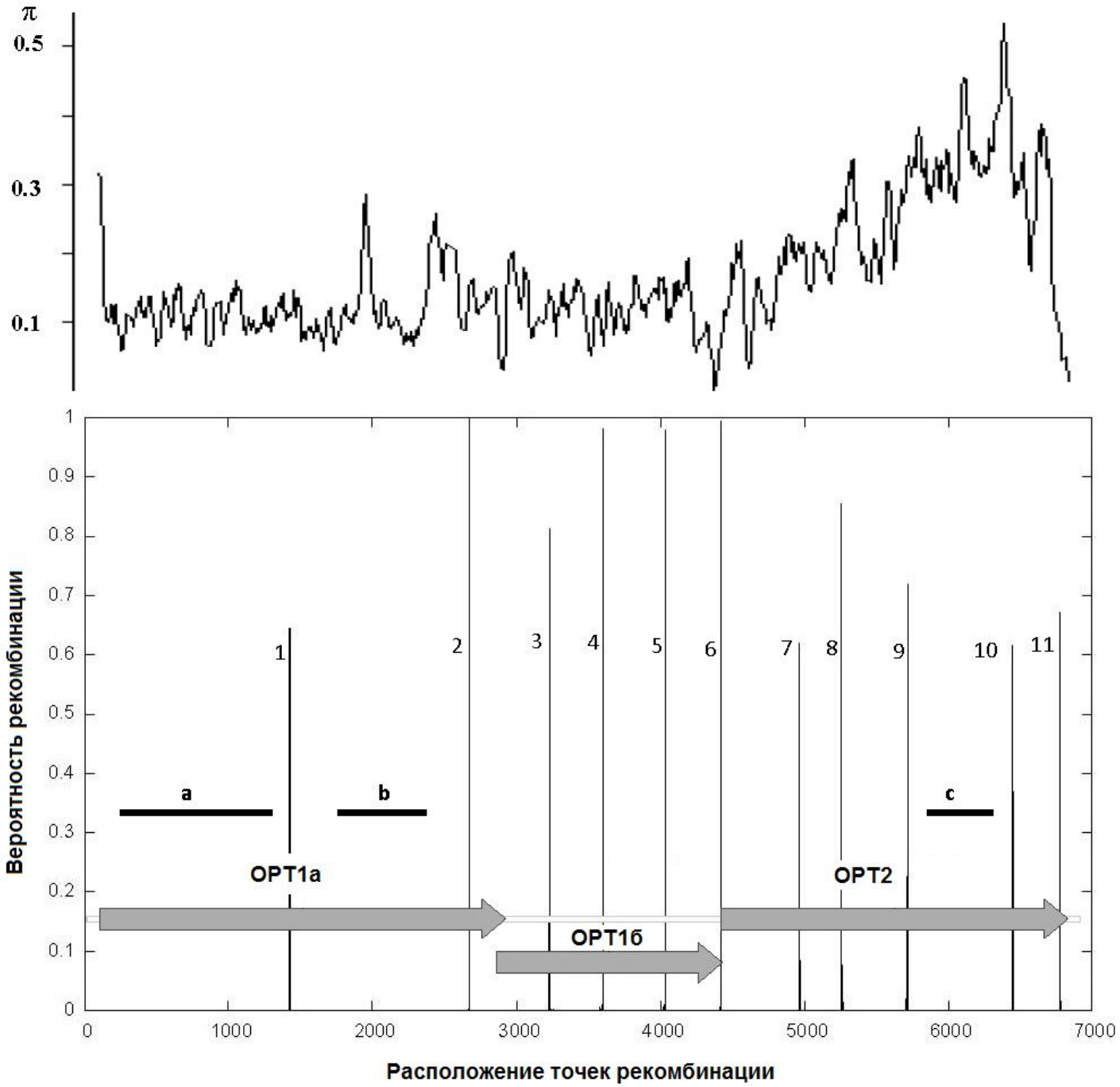
Define a custom (or choose a "named" REV) nucleotide substitution bias model (Help)

To/From	A	C	G	T
A	*	AC	1	AT
C	-	*	CG	CT
G	-	-	*	GT
T	-	-	-	*

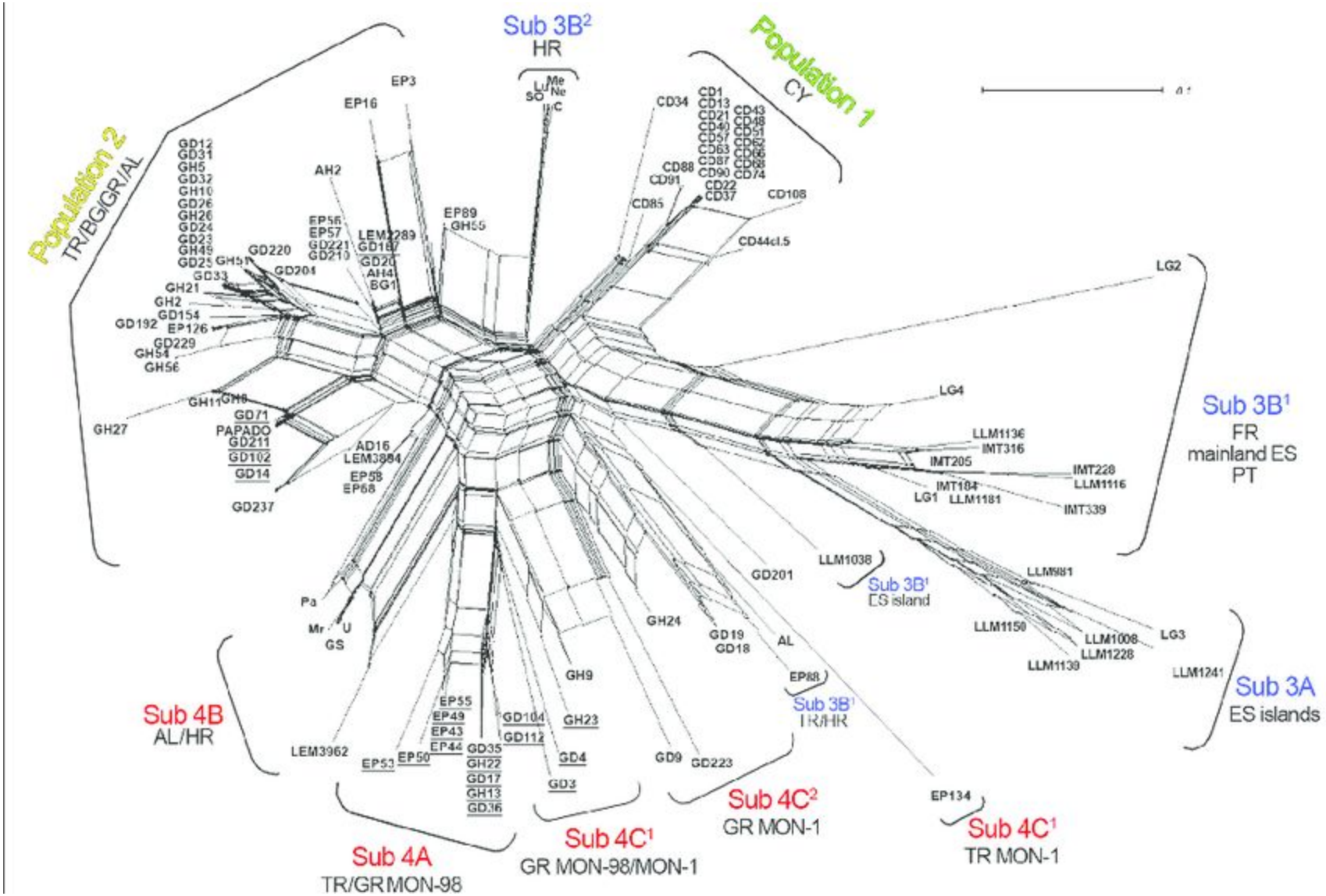
Site-to-site rate variation	None	Help
Rate classes	2	Help

Click to Run the analysis.

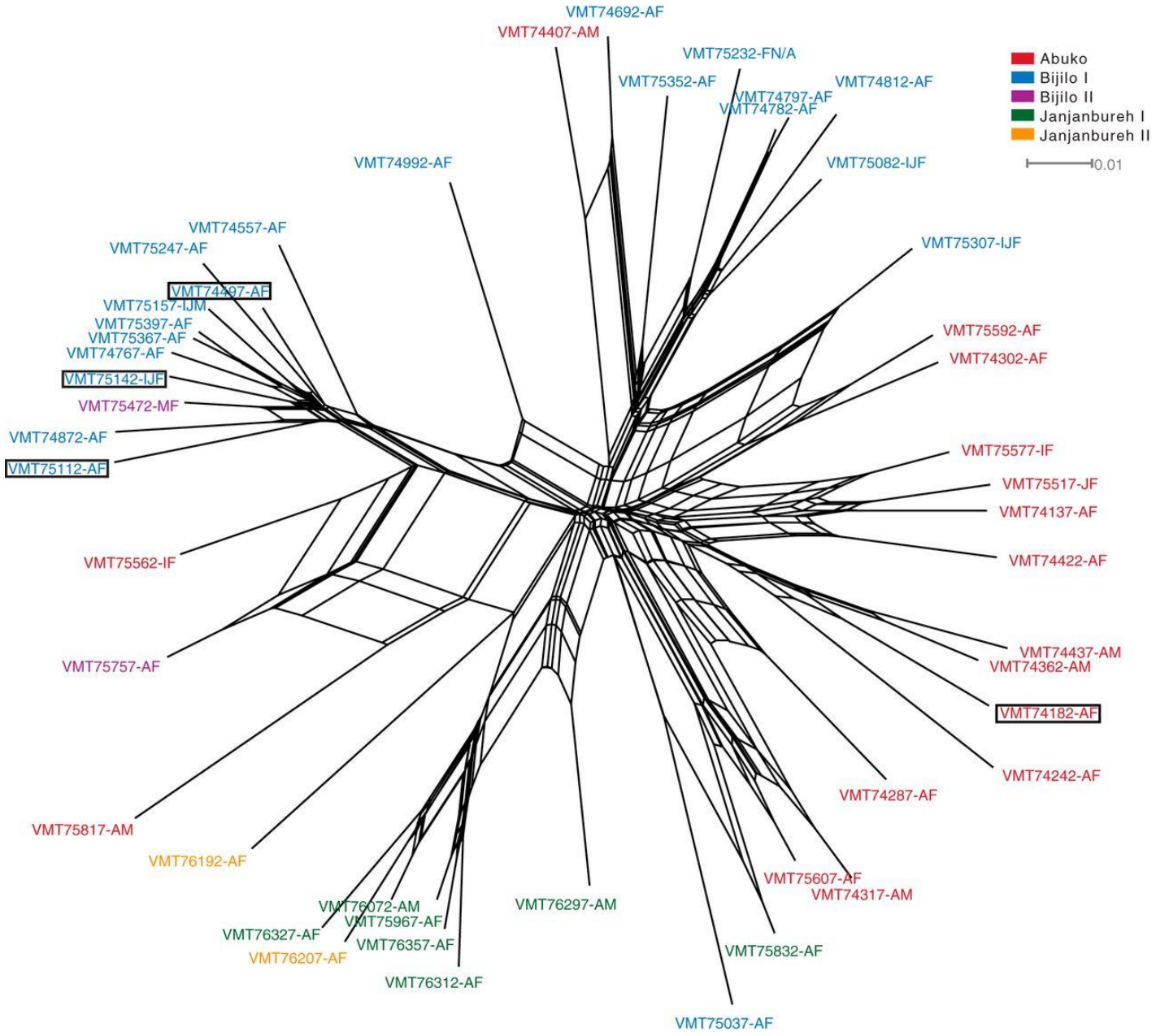
Unsure which substitution model to use? Execute an automatic model selection tool.


π
0.5
0.3
0.1
Вероятность рекомбинации
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
10
11
a
b
c
OPT1a
OPT1б
OPT2
0
1000
2000
3000
4000
5000
6000
7000
Расположение точек рекомбинации

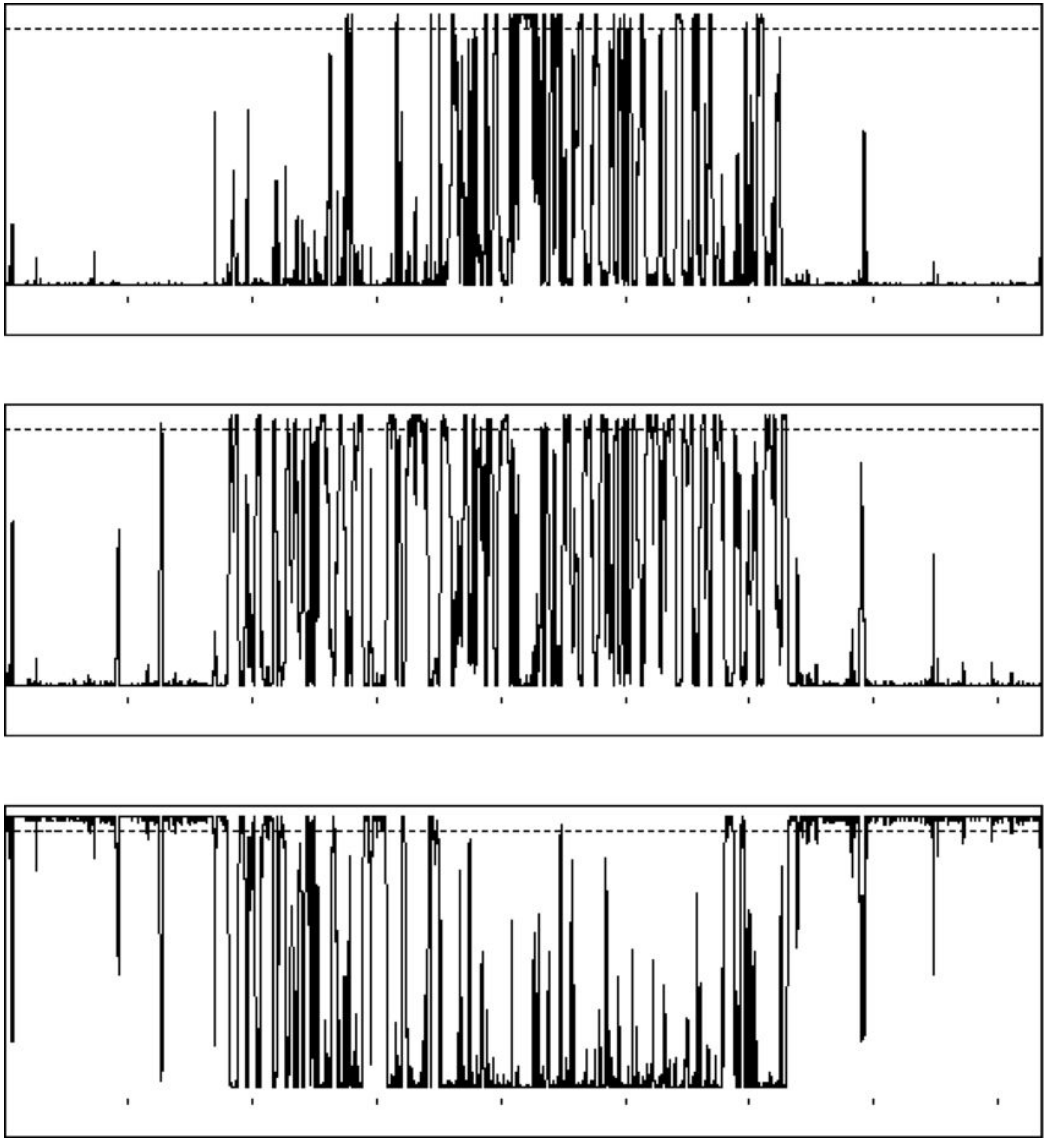
SplitsTree4 Рекомбинационный анализ



SplitsTree4 Рекомбинационный анализ



TOPALi Рекомбинационный анализ



Predicting recombination regions with HMM (Hidden Markov Model) implemented in TOPALi. Default parameter values were used. The horizontal axis represents the site in the alignment, the vertical axis represents the probability for topology change, and the dotted line shows the 95 percentile under the null hypothesis of no recombination. SARS-CoV, IBV, BCoV and HCoV was used, where SARS-CoV-severe acute respiratory syndrome-associated coronavirus, BCoV-bovine coronavirus, HCoV-human coronavirus, and IBV-avian infectious bronchitis virus