Варианты выравнивания

Попарное / множественное

Варианты выравнивания

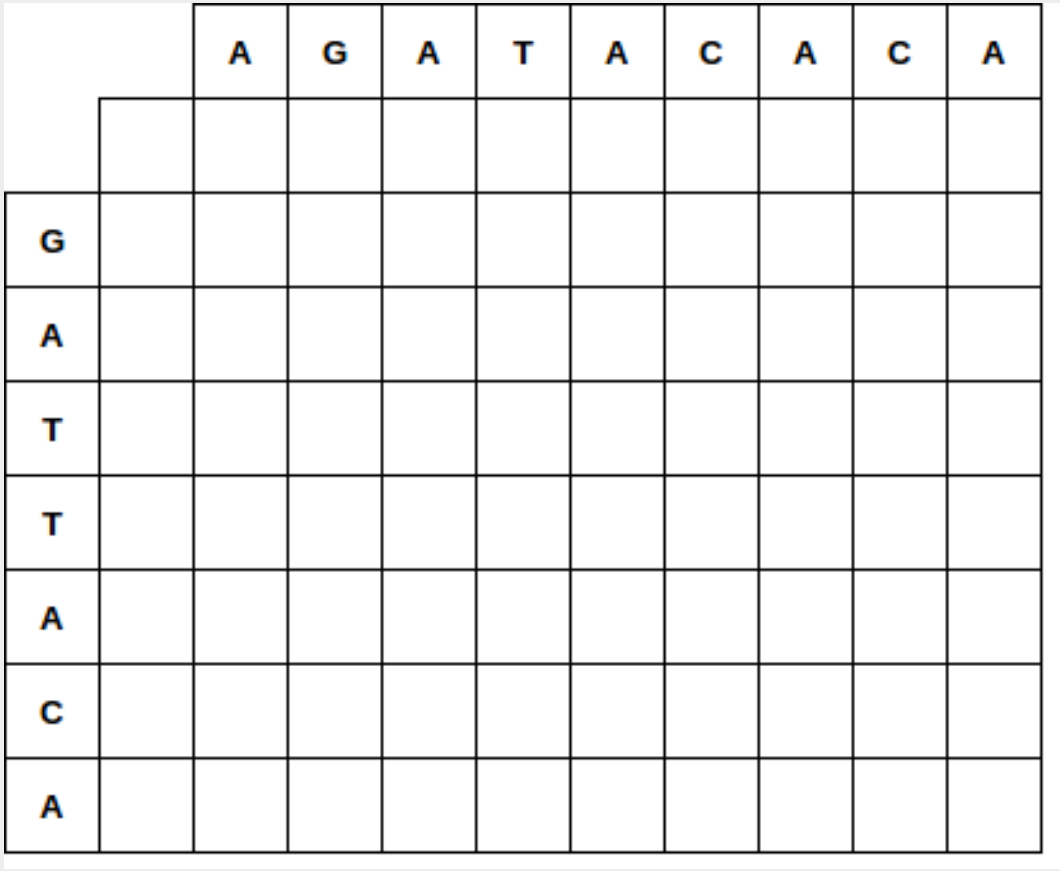
Попарное / множественное

Последовательности:

AGATACACA
GATTACA

Выравнивание:

AGATACACA
-GATT-ACA



Варианты выравнивания

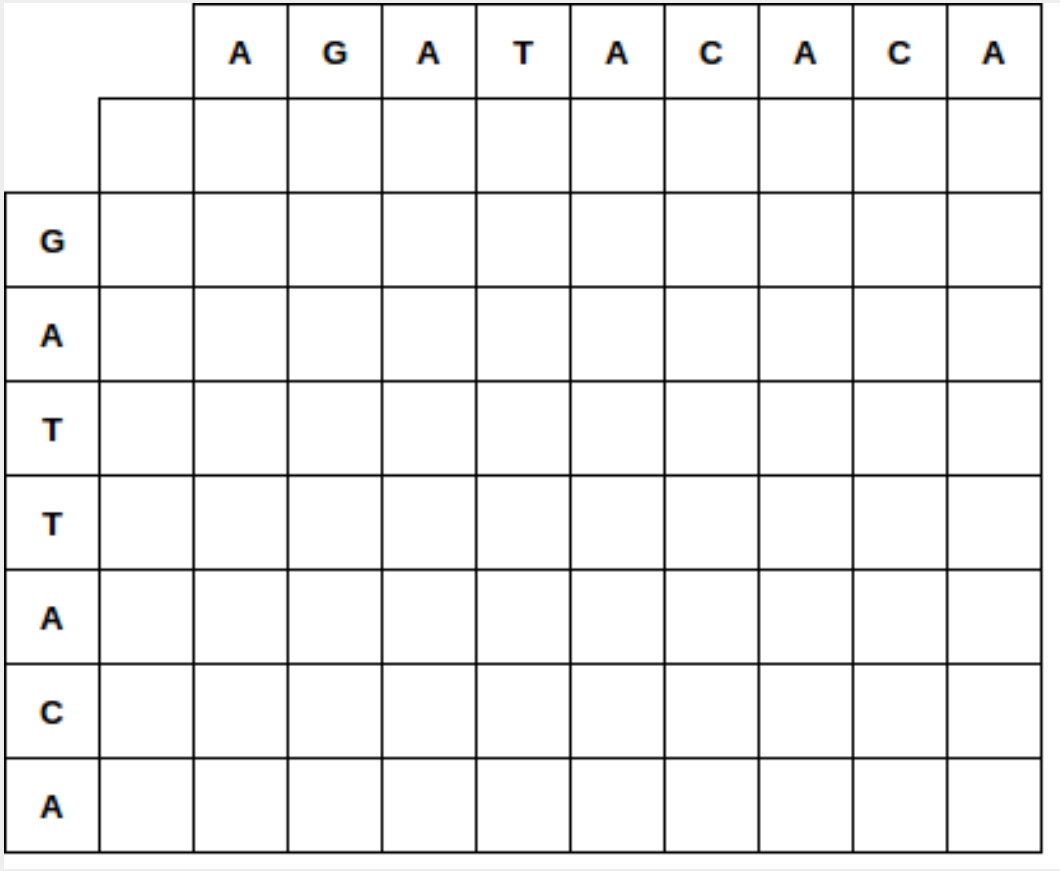
Попарное / множественное

Последовательности:

AGATACACA
GATTACA

Выравнивание:

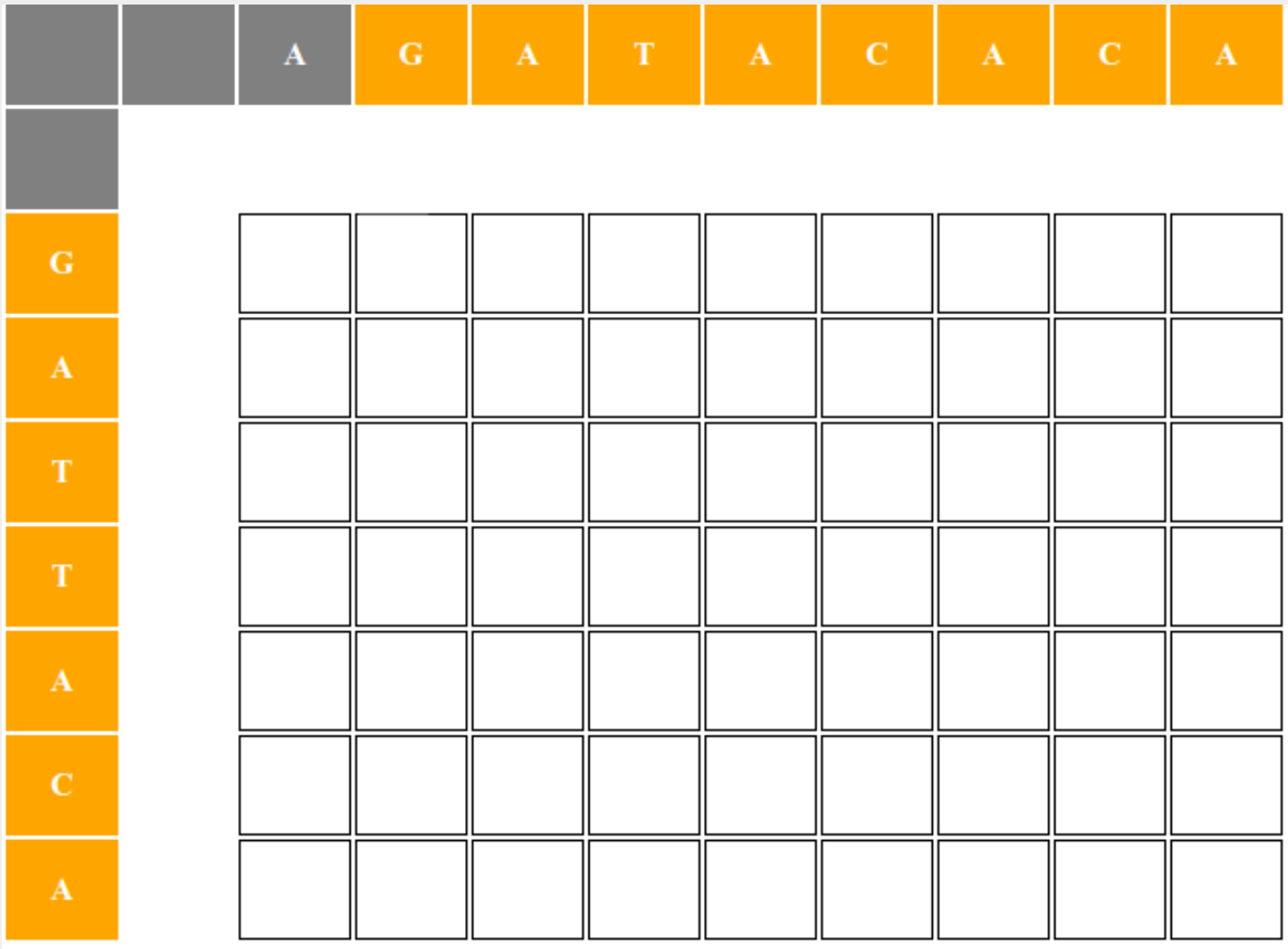
AGATACACA
-GATT-ACA



Глобальное / полуглобальное / локальное
Global / semi-global = glocal / local

Алгоритм Нидлмана-Вунша [Ванча] (Needleman-Wunsch algorithm)

Глобальное выравнивание



Алгоритм Нидлмана-Вунша [Ванча] (Needleman-Wunsch algorithm)

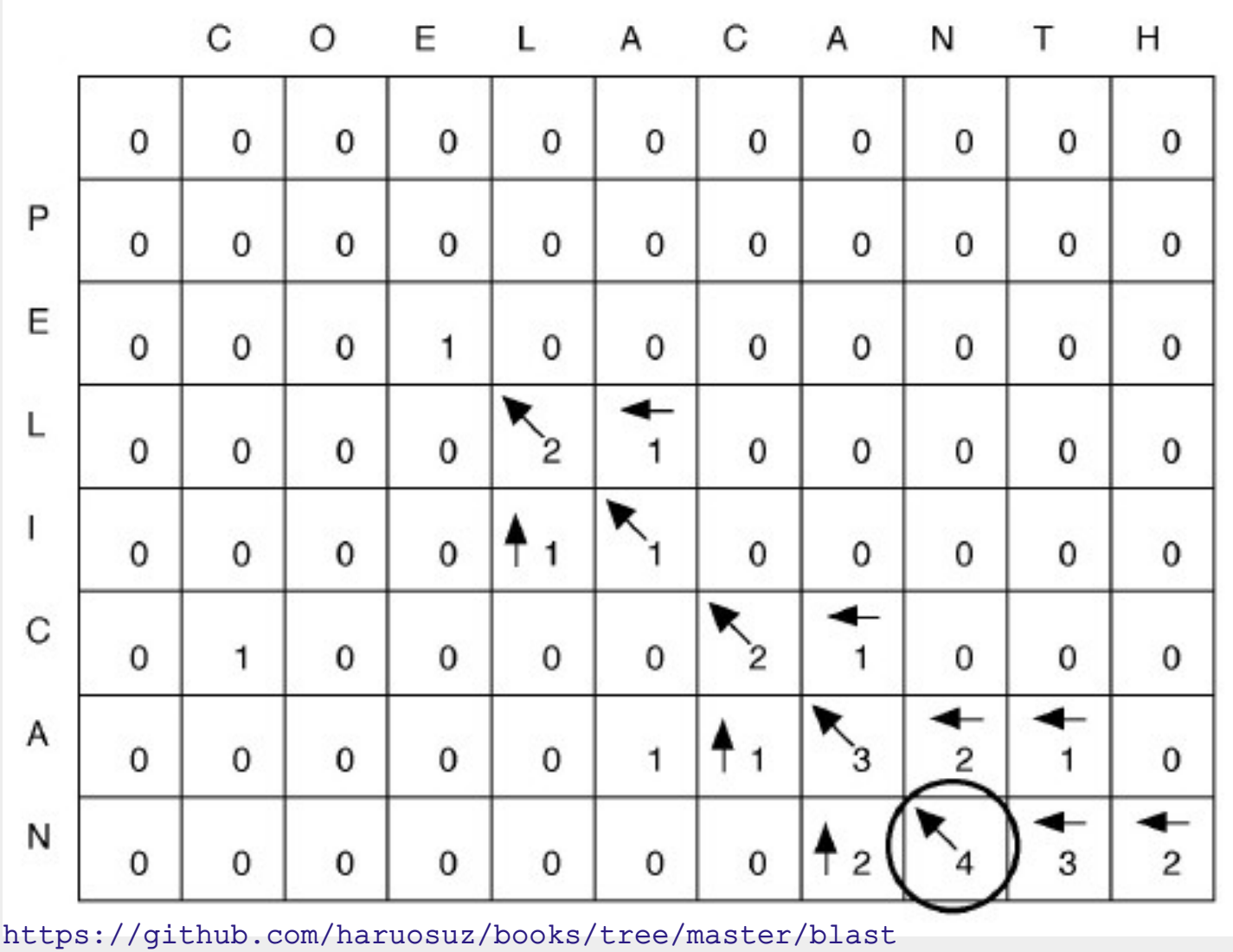
match = 1, mismatch = -1, gap = -1

		A	G	A	T	A	C	A	C	A
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
G	-1	-1	0	-1	-2	-3	-4	-5	-6	-7
A	-2	0	-1	1	0	-1	-2	-3	-4	-5
T	-3	-1	-1	0	2	1	0	-1	-2	-3
T	-4	-2	-2	-1	1	1	0	-1	-2	-3
A	-5	-3	-3	-1	0	2	1	1	0	-1
C	-6	-4	-4	-2	-1	1	3	2	2	1
A	-7	-5	-5	-3	-2	0	2	4	3	3

http://experiments.mostafa.io/public/needleman-wunsch/index.html

Алгоритм Смита-Ватермана (Smith-Waterman algorithm)

Локальное выравнивание



		C	O	E	L	A	C	A	N	T	H
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0	0
L	0	0	0	0	2	1	0	0	0	0	0
I	0	0	0	0	1	1	0	0	0	0	0
C	0	1	0	0	0	0	2	1	0	0	0
A	0	0	0	0	0	1	1	3	2	1	0
N	0	0	0	0	0	0	0	2	4	3	2

https://github.com/haruosuz/books/tree/master/blast

FASTA

FASTA = FASTP + FASTN

FASTA

FASTA = FASTP + FASTN

Точечная матрица сходства (dot matrix)

	A	G	A	T	A	C	A	C	A
G									
A									
T									
T									
A									
C									
A									

FASTA

FASTA = FASTP + FASTN

Точечная матрица сходства (dot matrix)

	A	G	A	T	A	C	A	C	A
G									
A									
T									
T									
A									
C									
A									

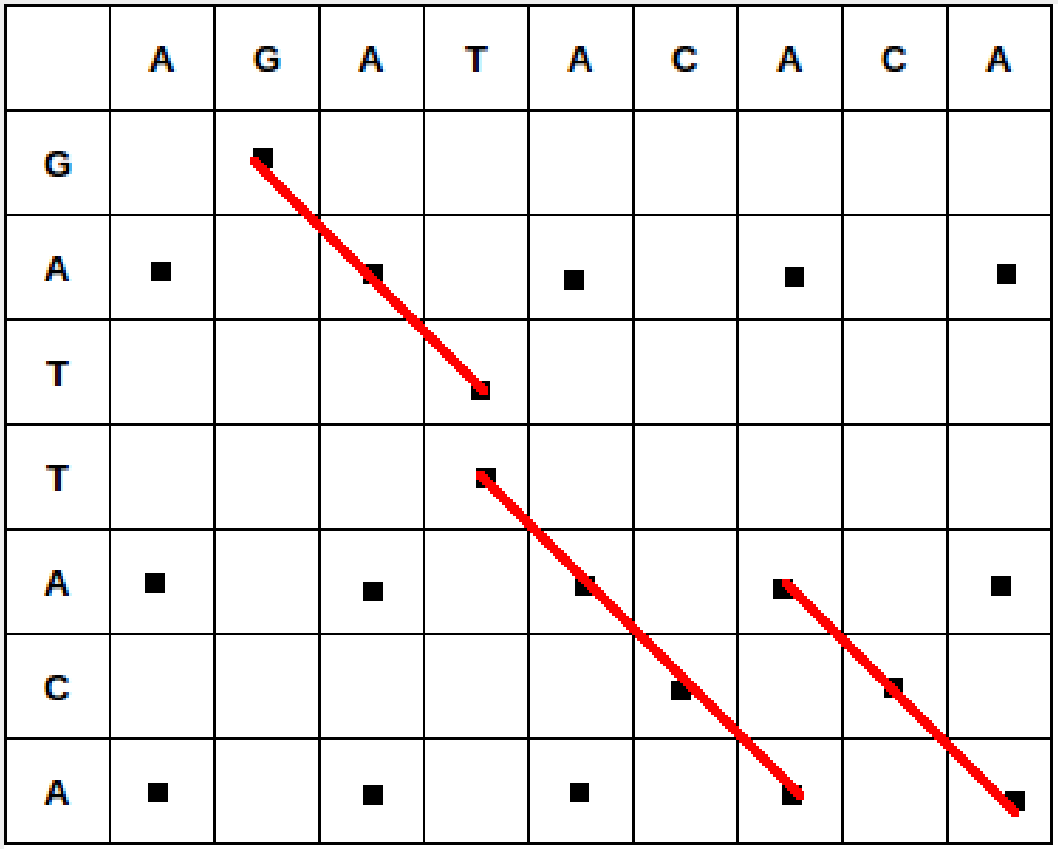
http://www.gen.tcd.ie/molevol/fasta.html

http://www.ebi.ac.uk/Tools/sss/fasta/

Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in enzymology. 1990 Dec 31;183:63-98.

FASTA

Точечная матрица сходства (dot matrix)



http://www.gen.tcd.ie/molevol/fasta.html
http://www.ebi.ac.uk/Tools/sss/fasta/

Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods in enzymology. 1990 Dec 31;183:63-98.

BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Варианты BLAST

Вариант BLAST	**Последовательность в запросе**	**Последовательности в базе**
blastn	Нуклеотидная	Нуклеотидные
blastp	Белковая	Белковые
blastx	Транслированная	Белковые
tblastn	Белковая	Транслированные
tblastx	Транслированная	Транслированные

NCBI Blast / WU-Blast

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990 Oct 5;215(3):403-10.

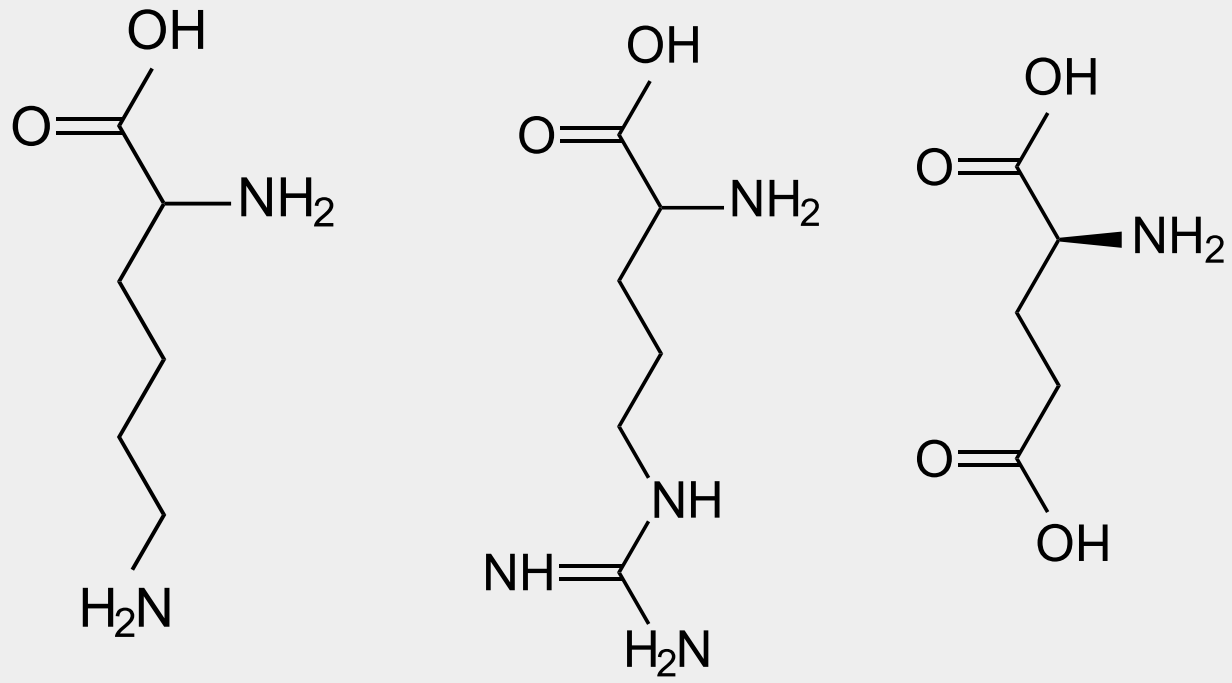
Метрики качества выравнивания

- Bit score — нормализованное (не зависящее от базы) значение score.
  $S' = \frac{\lambda S - \ln K}{\ln 2}$, где K и λ — параметры, позволяющие описать объём пространства поиска и систему оценки качества выравнивания.
- E-value — ожидаемое число случайных находок.
  $E = m \cdot n \cdot 2^{-S'}$, где m и n — длины сравниваемых последовательностей.

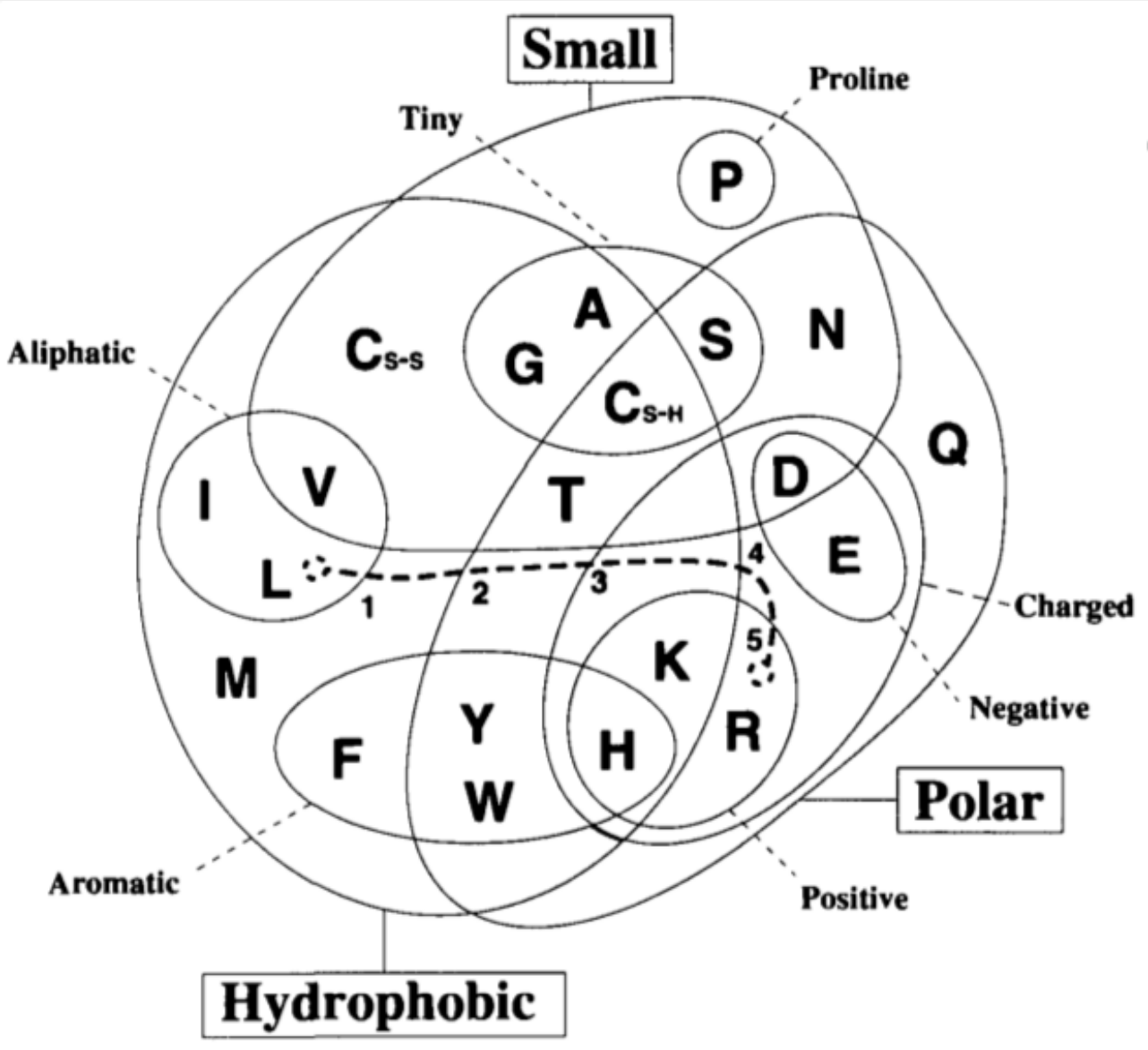
https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html
http://faculty.virginia.edu/wrpearson/fasta/fasta_guide.pdf

Не все аминокислотные замены одинаковы

AAA	AAG	AGA	GAA
Lys	*Lys*	*Arg*	*Glu*



Не все аминокислотные замены одинаковы



Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. CABIOS. 1993;9(6):745-56.

Матрицы сходства аминокислот

PAM = point accepted mutation

Чем **больше** номер, тем **больше** замен.

PAM100–PAM250

Dayhoff MO, Schwartz RM, Orcutt BC. A Model of evolutionary change in proteins. Atlas of protein sequence and structure. 1978; 5:345–52.

BLOSUM = BLOcks SUbstitution Matrix

Чем **меньше** номер, тем **больше** замен.

BLOSUM90–BLOSUM45

BLOSUM**62** по умолчанию в BLAST.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. PNAS U.S.A. 1992 Nov 15;89(22):10915-9.

Неклассические варианты BLASTP

PSI-BLAST

PSI = position specific iterative

https://www.ncbi.nlm.nih.gov/books/NBK2590/

PHI-BLAST

PHI = pattern hit initiated

https://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html

Программы для множественного выравнивания (MSA)

1. Семейство Clustal =
CLUSTer analysis of pairwise alignments

http://www.clustal.org/

Этапы ClustalW (прогрессивный алгоритм):

1. Попарные выравнивания;
2. NJ-дерево;
3. Выравнивание от листьев к корню.

Главная проблема: если пропуск появился, от него уже не избавиться.

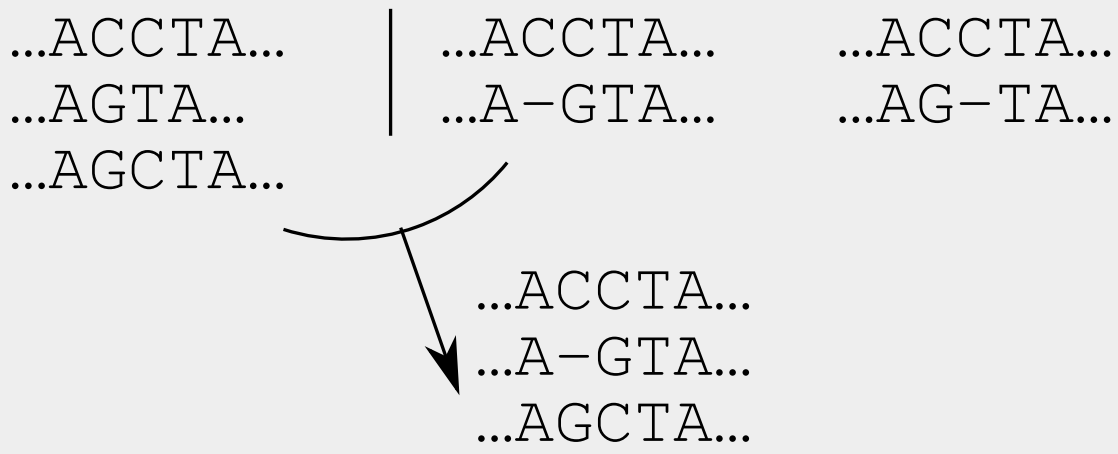
Программы для множественного выравнивания (MSA)

1. Семейство Clustal

Этапы ClustalW (прогрессивный алгоритм):

1. Попарные выравнивания;
2. NJ-дерево;
3. Выравнивание от листьев к корню.

Главная проблема: зависимость конечного результата от первого выравнивания.



Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.

Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.
2. T-Coffee = Tree-based Consistency Objective Function for alignment Evaluation: точнее ClustalW.

Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.
2. T-Coffee = Tree-based Consistency Objective Function for alignment Evaluation: точнее ClustalW.
3. MAFFT = multiple alignment with fast Fourier transform: быстрее.

Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.
2. T-Coffee = Tree-based Consistency Objective Function for alignment Evaluation: точнее ClustalW.
3. MAFFT = multiple alignment with fast Fourier transform: быстрее.
4. MUSCLE = MUltiple Sequence Comparison by Log Expectation: итеративный, точнее ClustalW, быстрый.

Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.
2. T-Coffee = Tree-based Consistency Objective Function for alignment Evaluation: точнее ClustalW.
3. MAFFT = multiple alignment with fast Fourier transform: быстрее.
4. MUSCLE = MUltiple Sequence Comparison by Log Expectation: итеративный, точнее ClustalW, быстрый.
5. DIALIGN: собирает по фрагментам; много настроек.

Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.
2. T-Coffee = Tree-based Consistency Objective Function for alignment Evaluation: точнее ClustalW.
3. MAFFT = multiple alignment with fast Fourier transform: быстрее.
4. MUSCLE = MUltiple Sequence Comparison by Log Expectation: итеративный, точнее ClustalW, быстрый.
5. DIALIGN: собирает по фрагментам; много настроек.
6. Kalign: довольно быстрый, необычный алгоритм сравнения строк.

Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.
2. T-Coffee = Tree-based Consistency Objective Function for alignment Evaluation: точнее ClustalW.
3. MAFFT = multiple alignment with fast Fourier transform: быстрее.
4. MUSCLE = MUltiple Sequence Comparison by Log Expectation: итеративный, точнее ClustalW, быстрый.
5. DIALIGN: собирает по фрагментам; много настроек.
6. Kalign: довольно быстрый, необычный алгоритм сравнения строк.
7. Probalign и ProbCons: используют HMM.

Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.
2. T-Coffee = Tree-based Consistency Objective Function for alignment Evaluation: точнее ClustalW.
3. MAFFT = multiple alignment with fast Fourier transform: быстрее.
4. MUSCLE = MUltiple Sequence Comparison by Log Expectation: итеративный, точнее ClustalW, быстрый.
5. DIALIGN: собирает по фрагментам; много настроек.
6. Kalign: довольно быстрый, необычный алгоритм сравнения строк.
7. Probalign и ProbCons: используют HMM.
8. prank: немного знает филогенетику.

Программы для MSA

1. Семейство Clustal = CLUSTer analysis of pairwise alignments: см. выше.
2. T-Coffee = Tree-based Consistency Objective Function for alignment Evaluation: точнее ClustalW.
3. MAFFT = multiple alignment with fast Fourier transform: быстрее.
4. MUSCLE = MUltiple Sequence Comparison by Log Expectation: итеративный, точнее ClustalW, быстрый.
5. DIALIGN: собирает по фрагментам; много настроек.
6. Kalign: довольно быстрый, необычный алгоритм сравнения строк.
7. Probalign и ProbCons: используют HMM.
8. prank: немного знает филогенетику.

Где взять программы?

1. Версия для командной строки;
2. онлайн-сервер;
3. пакеты для R (напр., ape, phyloch, ips);
4. UGENE;
5. MEGA;
6. Geneious.

Форматы файлов, хранящие MSA

1. Fasta

```
>d1a1x__ b.63.1.1 (-) p13-MTCP1 {Human (Homo sapiens)}
PPDHLWVHQEGIYRDEYQRTWVAVVEE--E--T--SF---------LR----------AR
>gi|6678257|ref|NP_033363.1|:(7-103) [Mus musculus]
HPNRLWIWEKHVYLDEFRRSWLPVVIK--S--N--EK---------FQ----------VI
>gi|7305557|ref|NP_038800.1|:(8-103) [Mus musculus]
PPRFLVCTRDDIYEDENGRQWVVAKVE--T--S--RSpygsrietcIT----------VH
```

1a. A2M:

```
>d1a1x__ b.63.1.1 (-) p13-MTCP1 {Human (Homo sapiens)}
PPDHLWVHQEGIYRDEYQRTWVAVVEE..E..T..SF.........LR..........AR
>gi|6678257|ref|NP_033363.1|:(7-103) [Mus musculus]
HPNRLWIWEKHVYLDEFRRSWLPVVIK..S..N..EK.........FQ..........VI
>gi|7305557|ref|NP_038800.1|:(8-103) [Mus musculus]
PPRFLVCTRDDIYEDENGRQWVVAKVE..T..S..RSpygsrietcIT..........VH
```

1b. A3M:

```
>d1a1x__ b.63.1.1 (-) p13-MTCP1 {Human (Homo sapiens)}
PPDHLWVHQEGIYRDEYQRTWVAVVEEETSFLRAR
>gi|6678257|ref|NP_033363.1|:(7-103) [Mus musculus]
HPNRLWIWEKHVYLDEFRRSWLPVVIKSNEKFQVI
>gi|7305557|ref|NP_038800.1|:(8-103) [Mus musculus]
PPRFLVCTRDDIYEDENGRQWVVAKVETSRSpygsrietcITVH
```

https://toolkit.tuebingen.mpg.de/reformat/help_params

Форматы файлов, хранящие MSA

2. CLUSTAL

```
CLUSTAL X (1.83) multiple sequence alignment

d1a1x__       GIYRDEYQRTWVAVVEE--E--T--SF---------LR-----AR
6678257       HVYLDEFRRSWLPVVIK--S--N--EK---------FQ-----VI
7305557       DIYEDENGRQWVVAKVE--T--S--RSPYGSRIETCIT-----VH
11415028      ---LDEKQHAWLPLTIEIKD--R--LQ---------LR-----VL
7305561       GIYEDEHHRVWIAVNVE--T--S--HS---------SHGNRT-VH
7305553       GIYEDEHHRVWIVANVE--TSHS--SH---------GN-----RR
27668591      -VYLDEFRRSWLPIVIK--S--N--GK---------FQ-----VI
27668589      GIYEDEHHRLWVVLDLQ--A--SHLSF---------SN-----RL
7305559       DIYEDEHGRQWVAAKVE--T--S--SH---------SPYCSVTVH
7305555       --YEDEHHRLWMVAKLE--T--C--SH---------SPYCNVTVH
```

https://toolkit.tuebingen.mpg.de/reformat/help_params

Форматы файлов, хранящие MSA

3. MEGA

```
#mega
 TITLE: Noninterleaved sequence data

 #mouse      AATTTTTACCCCGGGGGG
             AGGGGGGACCCCGGGGGG
 #human      AACCCTTACCCCGGGGGG
             AGGGGGGACCCCGGGGGG
 #cat        AATTTTTACA " This is a comment " AAGGGGGG
             AGGGGGGACCCCGGGGGG

 #mega
 TITLE: Interleaved sequence data

 #mouse      AATTTTTACCCCGGGGGG
 #human      AACCCTTACCCCGGGGGG
 #cat        AATTTTTACAAAGGGGGG

 #mouse      AGGGGGGACCCCGG
 #human      AGGGG" This is a comment " GGACCCCGG
 #cat        AGGGGGGACCCCGG
```

https://toolkit.tuebingen.mpg.de/reformat/help_params

Форматы файлов, хранящие MSA

4. MAF

```
##maf version=1 scoring=tba.v8
# tba.v8 (((human chimp) baboon) (mouse rat))
# multiz.v7
# maf_project.v5 _tba_right.maf3 mouse _tba_C
# single_cov2.v4 single_cov2 /dev/stdin

a score=23262.0
s hg16.chr7    27578828 9 + 158545518 AAA-GGGAAT
s panTro1.chr6 28741140 9 + 161576975 AAA-GGGAAT
s baboon         116834 9 +   4622798 AAA-GGGAAT
s mm4.chr6     53215344 9 + 151104725 -AATGGGAAT
s rn3.chr4     81344243 8 + 187371129 -AA-GGGGAT

a score=5062.0
 s hg16.chr7    27699739 6 + 158545518 TAAAGA
 s panTro1.chr6 28862317 6 + 161576975 TAAAGA
 s baboon         241163 6 +   4622798 TAAAGA
 s mm4.chr6     53303881 6 + 151104725 TAAAGA
 s rn3.chr4     81444246 6 + 187371129 taagga
```

https://cgwb.nci.nih.gov/FAQ/FAQformat.html#format5

Форматы файлов, хранящие MSA

5. MSF

```
!!AA_MULTIPLE_ALIGNMENT 1.0

  stdout MSF:  98 Type: P 16/01/02 CompCheck: 3003 ..

  Name: IXI_234 Len: 131  Check: 6808 Weight: 1.00
  Name: IXI_235 Len: 131  Check: 4032 Weight: 1.00
  Name: IXI_236 Len: 131  Check: 2744 Weight: 1.00
//

           1                                                  50
IXI_234    ~~TSPASIRPPAGPSSRPAMVSSRRTRPSPPGPRRPTGRPCCSAAPRRPQ
IXI_235    ~~TSPASIRPPAGPSSR.........RPSPPGPRRPTGRPCCSAAPRRPQ
IXI_236    ~~TSPASIRPPAGPSSRPAMVSSR..RPSPPPPRRPPGRPCCSAAPPRPQ

           51                                                 98
IXI_234    TAGGWKTCSGTCTTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACS
IXI_235    TAGGWKTCSGTCTTSTSTRHRGRSGW..........RASRKSMRAACS
IXI_236    TAGGWKTCSGTCTTSTSTRHRGRSGWSARTTTAACLRASRKSMRAAC~
```

http://emboss.sourceforge.net/docs/themes/AlignFormats.html

https://cgwb.nci.nih.gov/FAQ/FAQformat.html#format5

Форматы файлов, хранящие MSA

6. PHYLIP

```
3 95
IXI_234   TSPASIRPPA GPSSRPAMVS SRRTRPSPPG PRRPTGRPCC SAAPRRPQAT
          GGWKTCSGTC TTSTSTRHRG RSGWSARTTT AACLRASRKS MRAAC

IXI_235   TSPASIRPPA GPSSR----- ----RPSPPG PRRPTGRPCC SAAPRRPQAT
          GGWKTCSGTC TTSTSTRHRG RSGW------ ----RASRKS MRAAC

IXI_236   TSPASIRPPA GPSSRPAMVS SR--RPSPPP PRRPPGRPCC SAAPPRPQAT
          GGWKTCSGTC TTSTSTRHRG RSGWSARTTT AACLRASRKS MRAAC


3 95
IXI_234   TSPASIRPPA GPSSRPAMVS SRRTRPSPPG PRRPTGRPCC SAAPRRPQAT
IXI_235   TSPASIRPPA GPSSR----- ----RPSPPG PRRPTGRPCC SAAPRRPQAT
IXI_236   TSPASIRPPA GPSSRPAMVS SR--RPSPPP PRRPPGRPCC SAAPPRPQAT

          GGWKTCSGTC TTSTSTRHRG RSGWSARTTT AACLRASRKS MRAAC
          GGWKTCSGTC TTSTSTRHRG RSGW------ ----RASRKS MRAAC
          GGWKTCSGTC TTSTSTRHRG RSGWSARTTT AACLRASRKS MRAAC
```

https://toolkit.tuebingen.mpg.de/reformat/help_params

Форматы файлов, хранящие MSA

7. PAUP NEXUS

```
begin data;
dimensions ntax=3 nchar=4;
matrix
char1 ACGT
char2 AGCT
char3 CGAT
;
endblock;

begin data;
dimensions ntax=3 nchar=20;
format datatype=dna gap=- interleave;
matrix
   One   ATGCTATCCG [1-10]
   Two   ATCCTAGCCG
   Three CTGCTAGCCG
   One   TCATGACCTA [11-20]
   Two   T--AGACGGA
   Three TGGAGTCCTA
;
endblock;
```

https://toolkit.tuebingen.mpg.de/reformat/help_params

Форматы файлов, хранящие MSA

1. FASTA
2. CLUSTAL
3. MEGA
4. MAF
5. MSF
6. PHYLIP
7. NEXUS
8. VCF
9. SAM
10. ...

Базы данных

Нуклеотидные:

1. EMBL http://www.ebi.ac.uk/ena
2. NCBI (GENBANK) — в т. ч. RefSeq
   https://www.ncbi.nlm.nih.gov/genbank/
3. DDBJ http://www.ddbj.nig.ac.jp/

Базы данных

Нуклеотидные:

1. EMBL http://www.ebi.ac.uk/ena
2. NCBI (GENBANK) — в т. ч. RefSeq
   https://www.ncbi.nlm.nih.gov/genbank/
3. DDBJ http://www.ddbj.nig.ac.jp/

Белковые:

- UniProt = Swiss-Prot + TrEMBL
  http://www.uniprot.org/

Базы данных

Нуклеотидные:

1. EMBL http://www.ebi.ac.uk/ena
2. NCBI (GENBANK) — в т. ч. RefSeq
   https://www.ncbi.nlm.nih.gov/genbank/
3. DDBJ http://www.ddbj.nig.ac.jp/

Белковые:

- UniProt = Swiss-Prot + TrEMBL
  http://www.uniprot.org/

Базы для конкретных организмов, типов последовательностей *etc*:

- *Drosophila* http://flybase.org/
- rRNA https://www.arb-silva.de/
- SGD http://yeastgenome.org
- 1KITE http://www.1kite.org/
- Ensembl http://www.ensembl.org/index.html

Базы данных

Нуклеотидные:

1. EMBL http://www.ebi.ac.uk/ena
2. NCBI (GENBANK) <= **E-utilities**
   https://www.ncbi.nlm.nih.gov/genbank/
3. DDBJ http://www.ddbj.nig.ac.jp/

Белковые:

- UniProt = Swiss-Prot + TrEMBL
  http://www.uniprot.org/

Базы для конкретных организмов, типов последовательностей *etc*:

- *Drosophila* http://flybase.org/
- rRNA https://www.arb-silva.de/
- SGD http://yeastgenome.org
- 1KITE http://www.1kite.org/
- Ensembl http://www.ensembl.org/index.html