BLAST:

Basic Local Alignment Search Tool

С.А. Спирин, 21 апреля 2020

BLAST – алгоритм для нахождения участков локального сходства между последовательностями

Алгоритм сравнивает входную последовательность с последовательностями в базе данных, ищет сходные последовательности в базе данных и оценивает статистическую значимость находок.

Напоминание: сходство и гомология

Гомология — общность происхождения

- У гомологичных белков можно говорить о парах гомологичных остатков
- В эволюционно правильном выравнивании все остатки в одной колонке гомологичны друг другу

Признак гомологии — сходство последовательностей

- Для выявления сходства последовательности надо выровнять
- Подбирают оптимальное выравнивание, то есть имеющее наибольший вес
- Оптимальное выравнивание существует для любых последовательностей, в том числе негомологичных
- Для двух последовательностей можно рассматривать или глобальное, или локальное выравнивание

Идея поиска гомологов в банке последовательностей

На входе — последовательность, для которой хочется найти гомологичные («запрос»), и банк

Выровняем запрос с каждой последовательностью банка, посчитаем веса этих парных выравниваний

Отберём те последовательности банка («находки»), для которых вес **существенно выше, чем мог бы быть по случайным причинам.**

Почему локальное выравнивание?

Глобальное выравнивание следует применять только в случае заранее известной гомологии последовательностей по всей длине.

Часто у последовательностей гомологичны только отдельные части (примеры: гомеобелки, полипротеины, …)

Если про белки заранее ничего не известно, то более информативным будет локальное выравнивание. Поэтому именно оно применяется при поиске в банках данных.

Protein BLAST: поиск гомологов данного белка в банке аминокислотных последовательностей

Алгоритмы

- blastp
- psi-blast
- phi-blast

Можно использовать:

– из командной строки
– через веб-интерфейс

Что подаётся на вход программе BLAST?

- Последовательность запроса
- Банк последовательностей
- Параметры:
  - параметры выравнивания: матрица аминокислотных замен, штрафы за гэпы;
  - параметры поиска: длина слова и другие (см. далее);
  - параметры выдачи: максимальное число находок, пороги на качество выравнивания, форма выдачи (обычная, табличная, формат ASN, …)

Что выдает BLAST?

Выдача самой программы состоит из четырёх частей:

- заголовок с описанием программы, банка, запроса (query);
- список находок;
- выравнивания запроса с находками;
- несколько строк со статистическими показателями.

Веб-интерфейсы тем или иным способом перерабатывают выдачу программы. Раздел со статистикой обычно не показывается. Часто вставляется графическое изображение находок.

Выравнивание, выданное BLAST

```
Sequence ID: Q51368.2 Length: 342
Range 1: 234 to 338

Score:80.9 bits(198), Expect:1e-16,
Method:Compositional matrix adjust.,
Identities:46/115(40%), Positives:63/115(54%), Gaps:15/115(13%)

Query  123  SPFENTAPARLTSSTATAATSKPVTSVASGPRALSRNQPQYPARAQALRIEGQVKVKFDV  182
            +P  +  PA L S +   +  KP          L  + P YP  AQA  IEG+VKV F +
Sbjct  234  APSGSQGPAGLPSGSLNDSDIKP----------LRMDPPVYPRMAQARGIEGRVKVLFTI  283

Query  183  TPDGRVDNVQILSAKPANMFEREVKNAMRRWRYEPGKPGSGIVVN-----ILFKI  232
            T DGR+D++Q+L + P+ MF+REV+ AM +WR+EP   G  IV         FKI
Sbjct  284  TSDGRIDDIQVLESVPSRMFDREVRQAMAKWRFEPRVSGGKIVARQATKMFFFKI  338
```

Участок найденного белка, попавший в выравнивание

Отображение консервативности: между одинаковыми буквами ставится эта же буква, между сходными (positive) — знак +

Выравнивание, выданное BLAST

Длина найденного белка

```
Sequence ID: Q51368.2 Length: 342
Range 1: 234 to 338

Score:80.9 bits(198), Expect:1e-16,
Method:Compositional matrix adjust.,
Identities:46/115(40%), Positives:63/115(54%), Gaps:15/115(13%)

Query  123  SPFENTAPARLTSSTATAATSKPVTSVASGPRALSRNQPQYPARAQALRIEGQVKVKFDV  182
            +P  +  PA L S +   +  KP          L  + P YP  AQA  IEG+VKV F +
Sbjct  234  APSGSQGPAGLPSGSLNDSDIKP----------LRMDPPVYPRMAQARGIEGRVKVLFTI  283

Query  183  TPDGRVDNVQILSAKPANMFEREVKNAMRRWRYEPGKPGSGIVVN-----ILFKI  232
            T DGR+D++Q+L + P+ MF+REV+ AM +WR+EP   G  IV         FKI
Sbjct  284  TSDGRIDDIQVLESVPSRMFDREVRQAMAKWRFEPRVSGGKIVARQATKMFFFKI  338
```

Вес в битах

Вес

E-value

Число совпадений

Длина выравнивания

Число сходных букв

Число символов гэпа

Словарик BLAST

Identities — совпадения
Positives — сходные буквы, то есть те, для которых значение матрицы положительно
Gaps — знаки гэпа "–" (не индели!)

Для всех трёх приводится их число в виде числителя со знаменателем из длины выравнивания (не длины находки!) и процент от длины выравнивания

Score — вес выравнивания. Приводится в двух видах: сначала в битах (см. далее), затем в скобках обычный = сумма значений матрицы по сопоставлениям минус штраф за гэпы

Expect — E-value, то есть ожидаемое число выравниваний с тем же или большим весом. Запись вида 9e-15 означает $9 \cdot 10^{-15}$.

E-value – ожидаемое количество случайных находок с таким же и лучшим весом выравнивания, при поиске в той же базе данных, со случайным запросом той же длины и состава, с теми же параметрами на вычисление веса выравнивания.

В выдаче BLAST E-value называется “Expect”

Чем **меньше** E-value, тем **выше** значимость находки.

E-value зависит от:
– веса выравнивания (чем больше вес, тем **меньше** E-value);
– размера банка (чем больше банк, тем больше E-value);
– длины запроса (чем длиннее запрос, тем больше E-value);
– параметров, используемых для вычисления веса.

E-value – **ожидаемое** количество случайных находок с таким же и лучшим весом выравнивания, при поиске в той же базе данных, со случайным запросом той же длины и состава, с теми же параметрами на вычисление веса выравнивания.

Что означает слово «ожидаемое»?

E-value – **ожидаемое** количество **случайных** находок с таким же и лучшим весом выравнивания, при поиске в той же базе данных, со случайным запросом той же длины и состава, с теми же параметрами на вычисление веса выравнивания.

Что означает слово «ожидаемое»?

Формально это то, что называется «математическое ожидание случайной величины». Случайной величиной в данном случае является **число находок** *(NB! Просьба запомнить!)*

На практике ожидаемое вычисляется как **среднее** по достаточно большому количеству испытаний.

Другое ключевое слово — «случайных». Нам нужно понять, сколько можно ожидать именно случайных, то есть бессмысленных, негомологичных находок, чтобы оценить, насколько надёжно утверждение, что данная находка — действительно гомолог.

Как посчитать E-value

Прямой способ — вычислительный эксперимент:
перемешать буквы в запросе очень много раз, каждый раз запуская BLAST, и посмотреть, сколько в среднем при одном запуске бывает находок с весом выше данного.

Такой способ, естественно, не применяется :)

Стоит подумать: от чего и как может зависеть число случайных находок

Как посчитать E-value

Имеется замечательная теорема (С.Карлина):

$$\text{E-value} = Kmn \cdot e^{-\lambda S}$$

S – Score (вес)
m – длина исходной последовательности
n – размер базы данных (суммарная длина всех последовательностей)
K и λ – две константы

Коэффициенты K и λ зависят от параметров вычисления веса, то есть матрицы и штрафов за гэпы.

BLAST хранит значения K и λ для нескольких наборов параметров вычисления веса (их раз и навсегда нашли посредством вычислительного эксперимента).

Вес в битах

Вес в битах B зависит от обычного веса S и параметров вычисления веса. Эта зависимость подобрана так, чтобы

$$\text{E-value} = mn \cdot 2^{-B}$$

m – длина исходной последовательности
n – размер базы данных
(констант K и λ теперь нет, они “загнаны внутрь B”)

Нетрудно подсчитать, что $B = (\lambda S - \ln K)/\ln 2$

Далее описан интерфейс, установленный на «родине» BLAST: National Center for Biotechnology Information (NCBI) в США, http://blast.ncbi.nlm.nih.gov/

http://blast.ncbi.nlm.nih.gov/ → Protein BLAST

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

Reset page | Bookmark

Enter Query Sequence

BLASTP programs search protein databases using a protein query. more...

Enter accession number(s), gi(s), or FASTA sequence(s) | Clear | Query subrange | From | To

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

New

Or, upload file | Выберите файл | Файл не выбран

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

вводим последовательность

Choose Search Set

Database | Non-redundant protein sequences (nr)

банк

Non-redundant protein sequences (nr)
Reference proteins (refseq_protein)
Model Organisms (landmark)
UniProtKB/Swiss-Prot(swissprot)
Patented protein sequences(pataa)
Protein Data Bank proteins(pdb)
Metagenomic proteins(env_nr)
Transcriptome Shotgun Assembly proteins (tsa_nr)

Organism Optional | Enter organism name or id--completions will be suggested | exclude | +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional | Models (XM/XP) | Non-redundant RefSeq proteins (WP) | Uncultured/environmental sar

организм (если надо ограничить)

Program Selection

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST | Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

дополнительные параметры

Algorithm parameters

Дополнительные параметры

максимальный размер выдачи

порог на E-value

параметры выравнивания

борьба с «участками малой сложности»

Algorithm parameters

General Parameters

Max target sequences: 100

Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: ☐ Low complexity regions

Mask: ☐ Mask for lookup table only

☐ Mask lower case letters

Участок малой сложности

Ищем по белку P02929
Если отключить “Compositional adjustment” и фильтр, то среди прочих выдаётся следующее выравнивание:

```
Query: P02929 TONB_ECOLI; Subject: Q95P09 TSEP_GLOMM
 Score = 63.5 bits (153),  Expect = 1e-09
 Identities = 32/76 (42%), Positives = 47/76 (62%), Gaps = 6/76 (8%)

Query  56   EPPQAVQPPPEPVVEPEPEPEPIPEP-PKEAPVVIEKPKPKPKPKPKPVKKVQEQPKRDV  114
            EP    +P PEP  EPEPEPEP PEP P+  P    +P+P+P+P+P+P  + + +P+ +
Sbjct  243  EPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEP  302

Query  115  KP-----VESRPASPF  125
            +P      ES+P S F
Sbjct  303  EPEPQPEPESKPNSLF  318
```

в исходном белке имеется участок, содержащий очень много пролина и глутаминовой кислоты

Данное выравнивание **не свидетельствует о гомологии**, несмотря на хорошее значение E-value (10^{-9})

Участок малой сложности

Определяется как участок с смещенным составом (biased composition)

- Гомополимерные участки
- Короткие повторы
- Перепредставленность отдельных остатков

✓ Может мешать анализу последовательностей
✓ Вычисление E-value (параметры *K* и λ) опирается на средние по всем белкам частоты аминокислотных остатков, поэтому на участках малой сложности оно становится некорректным
✓ Обычно ведет к ложным предсказаниям гомологии (false positives)
✓ Лучше использовать «Compositional adjustment» (по умолчанию включен)

Выдача BLAST в интерфейсе NCBI

BLAST® » blastp suite » results for RID-9X5D3ACN014

Home Recent Results Saved Strategies Help

‹ Edit Search | Save Search | Search Summary

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	P02929:RecName: Full=Protein TonB
RID	9X5D3ACN014 *Search expires on 04-22 14:41 pm* Download All
Program	BLASTP Citation
Database	swissprot See details
Query ID	P02929.2
Description	RecName: Full=Protein TonB [Escherichia coli K-12]
Molecule type	amino acid
Query Length	239
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to **E value** to **Query Coverage** to

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *10 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
RecName: Full=Protein TonB [Escherichia coli K-12]	471	471	100%	4e-170	100.00%	P02929.2
RecName: Full=Protein TonB [Salmonella enterica subsp. enterica serovar Typhimurium str. LT2]	313	313	100%	5e-108	83.54%	P25945.2
RecName: Full=Protein TonB [Klebsiella pneumoniae]	270	270	97%	1e-90	67.08%	P45610.1
RecName: Full=Protein TonB [Serratia marcescens]	125	125	52%	5e-34	54.69%	P26185.1
RecName: Full=Protein TonB [Klebsiella aerogenes KCTC 2190]	116	116	25%	1e-30	87.10%	P46383.2
RecName: Full=Protein TonB [Yersinia enterocolitica]	110	110	48%	4e-28	47.06%	Q05740.1

Переход к текстовому виду

Чтобы скачать выдачу самой программы (а не её обработку интерфейсом), можно поступить так:

Edit Search | Save Search | Search Summary

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	P02929:RecName: Full=Protein TonB
RID	9X5D3ACN014 *Search expires on 04-22 14:41 pm* Download All
Program	BLASTP Citation
Database	swissprot See details
Query ID	P02929.2
Description	RecName: Full=Protein TonB [Escherichia coli K-12]
Molecule type	amino acid
Query Length	239
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity [ ] to [ ] E value [ ] to [ ] Query Coverage [ ] to [ ]

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments

Download | Manage Columns | Show 100

select all *10 sequences selected*

ance tree of results Multiple alignment

Description	Total Score	Query Cover	E value	Per. Ident	Accession
RecName: Full=Protein TonB [Escherichia coli K-12]	471	100%	4e-170	100.00%	P02929.2
RecName: Full=Protein TonB [Salmonella enterica subsp. enterica serovar Typhimurium str. LT2]	313	100%	5e-108	83.54%	P25945.2
RecName: Full=Protein TonB [Klebsiella pneumoniae]	270	97%	1e-90	67.08%	P45610.1
RecName: Full=Protein TonB [Serratia marcescens]	125	52%	5e-34	54.69%	P26185.1
RecName: Full=Protein TonB [Klebsiella aerogenes KCTC 2190]	116	25%	1e-30	87.10%	P46383.2
RecName: Full=Protein TonB [Yersinia enterocolitica]	110	48%	4e-28	47.06%	Q05740.1
RecName: Full=Protein TonB [Pseudomonas aeruginosa PAO1]	0.9	46%	1e-16	40	

FASTA (complete sequence)
FASTA (aligned sequences)
GenBank (complete sequence)
Hit Table (text)
Hit Table (CSV)
Text
XML
ASN.1

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

Feedback

Текстовая выдача BLAST

```
RID: 9X5D3ACN014
Job Title:P02929:RecName: Full=Protein TonB
Program: BLASTP
Query: RecName: Full=Protein TonB [Escherichia coli K-12] ID: P02929.2(amino acid) Length: 239
Database: swissprot Non-redundant UniProtKB/SwissProt sequences

Sequences producing significant alignments:
                                                                    Max    Total Query    E    Per.
Description                                                         Score  Score cover Value  Ident  Accession
RecName: Full=Protein TonB [Escherichia coli K-12]                  471    471   100%  4e-170 100.00 P02929.2
RecName: Full=Protein TonB [Salmonella enterica subsp. enteric...   313    313   100%  5e-108 83.54  P25945.2
RecName: Full=Protein TonB [Klebsiella pneumoniae]                  270    270   97%   1e-90  67.08  P45610.1
RecName: Full=Protein TonB [Serratia marcescens]                    125    125   52%   5e-34  54.69  P26185.1
RecName: Full=Protein TonB [Klebsiella aerogenes KCTC 2190]         116    116   25%   1e-30  87.10  P46383.2
RecName: Full=Protein TonB [Yersinia enterocolitica]                110    110   48%   4e-28  47.06  Q05740.1
RecName: Full=Protein TonB [Pseudomonas aeruginosa PAO1]            80.9   80.9  46%   1e-16  40.00  Q51368.2
RecName: Full=Protein TonB [Vibrio cholerae O1 biovar El Tor...     43.1   43.1  92%   7e-04  27.50  O52042.2
RecName: Full=Protein TonB [[Haemophilus] ducreyi 35000HP]          33.5   33.5  17%   1.2    36.36  O51810.1
RecName: Full=Translation initiation factor IF-2 [Laribacter...     31.6   31.6  13%   6.8    53.12  C1D8X2.1

Alignments:
>RecName: Full=Protein TonB [Escherichia coli K-12]
Sequence ID: P02929.2 Length: 239
Range 1: 1 to 239

Score:471 bits(1211), Expect:4e-170,
Method:Compositional matrix adjust.,
Identities:239/239(100%), Positives:239/239(100%), Gaps:0/239(0%)

Query  1    MTLDLPRRFPWPTLLSVCIHGAVVAGLLYTSVHQVIELPAPAQPISVTMVTPADLEPPQA  60
            MTLDLPRRFPWPTLLSVCIHGAVVAGLLYTSVHQVIELPAPAQPISVTMVTPADLEPPQA
Sbjct  1    MTLDLPRRFPWPTLLSVCIHGAVVAGLLYTSVHQVIELPAPAQPISVTMVTPADLEPPQA  60

Query  61   VQPPPEPVVEPEPEPEPIPEPPKEAPVVIEKPKPKPKPKPKPVKKVQEQPKRDVKPVESR  120
            VQPPPEPVVEPEPEPEPIPEPPKEAPVVIEKPKPKPKPKPKPVKKVQEQPKRDVKPVESR
Sbjct  61   VQPPPE[illegible]VEPEPEPEPIPEPPKEAPVVIEKPKPKPKPKPKPVKKVQEQPKRDVKPVESR  120

Query  121  PASPFENTAPARLTSSTA[illegible]KPVTSVASGPRALSRNQPQYPARAQALRIEGQVKVKF  180
            PASPFENTAPARLTSSTATAATSK[illegible]GPRALSRNQPQYPARAQALRIEGQVKVKF
Sbjct  121  PASPFENTAPARLTSSTATAATSKPVTSV[illegible]QPQYPARAQALRIEGQVKVKF  180
```

Текстовая выдача BLAST

```
RID: 9X6N7P7G016
Job Title:ORF1ab
Program: BLASTP
Query: ORF1ab ID: lcl|Query_23045(amino acid) Length: 7050
Database: swissprot Non-redundant UniProtKB/SwissProt sequences

Sequences producing significant alignments:
                                                                  Max    Total Query    E    Per.
Description                                                       Score  Score cover Value  Ident  Accession
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 12928  12928 100%  0.0    85.90  P0C6X7.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 12882  12882 100%  0.0    85.76  P0C6W2.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 12867  12867 100%  0.0    85.52  P0C6V9.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 12861  12861 100%  0.0    85.50  P0C6W6.1
RecName: Full=Replicase polyprotein 1a; Short=pp1a; AltName:...   7476   7476  61%   0.0    80.46  P0C6U8.1
RecName: Full=Replicase polyprotein 1a; Short=pp1a; AltName:...   7461   7461  61%   0.0    80.20  P0C6F8.1
RecName: Full=Replicase polyprotein 1a; Short=pp1a; AltName:...   7436   7436  61%   0.0    79.80  P0C6F5.1
RecName: Full=Replicase polyprotein 1a; Short=pp1a; AltName:...   7431   7431  61%   0.0    79.71  P0C6T7.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 6323   6323  95%   0.0    48.41  P0C6W5.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 6135   6135  99%   0.0    45.73  P0C6W4.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5990   6347  99%   0.0    50.39  K9N7C7.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5608   6235  92%   0.0    55.76  P0C6W1.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5599   6237  93%   0.0    55.66  P0C6W3.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5435   5608  93%   0.0    49.37  P0C6W8.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5434   5606  93%   0.0    49.39  P0C6W7.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5418   5554  83%   0.0    48.88  P0C6X8.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5404   5574  93%   0.0    49.26  P0C6X6.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5397   5555  87%   0.0    48.81  P0C6X9.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5395   5565  93%   0.0    49.23  P0C6W9.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5346   5484  93%   0.0    48.37  P0C6Y0.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5305   5483  91%   0.0    48.52  P0C6X3.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5301   5477  91%   0.0    48.52  P0C6X4.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5299   5474  91%   0.0    48.53  P0C6X2.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 5267   5435  87%   0.0    48.86  P0C6X0.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 4319   4397  72%   0.0    46.73  P0C6W0.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 4282   4358  70%   0.0    46.53  P0C6Y4.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 4266   4344  70%   0.0    46.72  P0C6X5.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 4227   4303  73%   0.0    45.46  P0C6X1.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 4179   4252  72%   0.0    45.66  P0C6Y5.1
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 4154   4217  71%   0.0    45.69  Q98VG9.2
RecName: Full=Replicase polyprotein 1ab; Short=pp1ab; AltName:... 4151   4216  70%   0.0    45.61  P0C6Y3.1
```

Словарик (таблица находок BLAST)

Max Score: самый большой из весов (в битах) выравниваний запроса с данной находкой

Total Score: суммарный вес (в битах) всех выравниваний запроса с данной находкой

Query cover: процент длины запроса, покрытого выравниваниями

E Value: в таблице находок это E-value, посчитанное по особой формуле на основе **всех** выравниваний запроса с данной находкой

Per. Ident: процент идентичных букв в лучшем (по весу) из выраваниваний запроса с данной находкой

BLAST — эвристический алгоритм

Алгоритмы биоинформатики можно разделить на точные и эвристические.
Точные алгоритмы решают какую-либо точно сформулированную формализованную задачу. Пример: алгоритм Нидельмана – Вунша, который для данных последовательностей находит выравнивание с максимальным весом (реализован в программе needle).
Эвристические алгоритмы — те, для которых формальную задачу сформулировать нельзя.

BLAST **не гарантирует** нахождение оптимального локального выравнивания. За счёт этого достигается высокая скорость работы. Но теоретически возможно, что BLAST не найдёт в банке вполне достоверный (судя по выравниванию) гомолог.

Дополнительные параметры

Algorithm parameters

General Parameters

Max target sequences	100 Select the maximum number of aligned sequences to display
Short queries	☑ Automatically adjust parameters for short input sequences
Expect threshold	10
Word size	3
Max matches in a query range	0

Длина слова

Scoring Parameters

Matrix	BLOSUM62
Gap Costs	Existence: 11 Extension: 1
Compositional adjustments	Conditional compositional score matrix adjustment

Filters and Masking

Filter	☐ Low complexity regions
Mask	☐ Mask for lookup table only ☐ Mask lower case letters

Длина слова

Одним из параметров BLAST является длина слова (word size).

Чем больше длина слова, тем быстрее работает BLAST, но тем меньше его **чувствительность**. Это означает, что вероятность пропустить гомологи возрастает.

Сейчас на сайте NCBI значение длины слова по умолчанию равно 6, доступны значения 2 и 3.