

BLAST: Basic Local Alignment Search Tool

Ю.Пеков, А.Алексеевский, С.Спирин

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.

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

Длина найденного белка

Length=129 Number of matches=1

Вес в битах

Вес

E-value

Score = 78.6 bits (192), Expect = 9e-15, Method: Compositional matrix adjust.
Identities = 34/73 (47%), Positives = 50/73 (68%), Gaps = 0/73 (0%)

Query 17 YRLEEVQKHNNSSQSTWIIIVHHRIYDITKFLDEHPGGEEVLREQAGGDATENFEDVGHSTD 76
 Y EEV +H W+I++ ++Y+I+ ++DEHPGGEEV+ + AG DATE F+D+GHS +
Sbjct 11 YTHEEVAQHHTTHDDLWVILNGKVYNISNYIDHPGGEEVILDCAGTDATEAFDDIGHSDE 70

Query 77 ARALSETFIIGEL 89
 A + E IG L
Sbjct 71 AHEILEKLYIGNL 83

Число сходных
"букв"

Число символов гэпа

Число совпадений

Длина выравнивания

E-value – ожидаемое количество **случайных** находок с таким же и лучшим весом (в той же базе данных, с 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.

Как посчитать E-value

Прямой способ — вычислительный эксперимент: перемешать банк (или запрос) очень много раз, каждый раз запуская BLAST, и посмотреть, сколько в среднем найдётся находок с весом выше данного.

Такой способ, естественно, не применяется :)

Как посчитать E-value

Имеется замечательная теорема (С.Карлина):

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

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

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

m – длина исходной последовательности

n – размер базы данных

(констант K и λ теперь нет, они “загнаны внутрь B ”)

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

Здесь 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

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastp suite

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear Query subrange

From To

Or, upload file Файл не выбран

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism Optional

Enter organism name or id--completions will be suggested

Exclude Optional

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

☐ Show results in a new window

[Algorithm parameters](#)

ВВОДИМ ПОСЛЕДОВАТЕЛЬНОСТЬ

банк

организм (если надо ограничить)

дополнительные параметры

nr	Non-redundant
refseq	Reference Sequences
swissprot	SWISS-PROT
pat	Patents
pdb	Protein Data Bank
env_nr	Environmental samples

Дополнительные параметры

максимальный
размер выдачи

▼ Algorithm parameters

General Parameters

Max target sequences Select the maximum number of aligned sequences to display ⓘ

Short queries ☒ Automatically adjust parameters for short input sequences ⓘ

Expect threshold ⓘ

Word size ⓘ

Max matches in a query range ⓘ

Scoring Parameters

Matrix ⓘ

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 ⓘ

порог на E-value

параметры
выравнивания

борьба с «участками
малой сложности»

Участок малой сложности

Ищем: белок P02929

если отключить “Compositional adjustment” и фильтр, то одной из находок (18-ой от начала) будет следующее:

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   ERPQAVQPPPEPVVEPEPEPEPIPEP-PKEAPVVIEKPKPKPKPKPKPKFVKKVQEQQPKRDV   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 среднее по всем белкам распределение частот аминокислотных остатков
- ✓ Обычно ведет к ложным предсказаниям гомологии (false positives)
- ✓ Лучше использовать «Compositional adjustment» (по умолчанию включен)

Переход к текстовому виду

Чтобы увидеть выдачу самой программы (а не его обработку интерфейсом), можно поступить так:

выбираем formatting options

① Your search parameters were adjusted to search for a short input sequence.

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

Formatting options

Reformat

Show	Alignment	as	Plain text	<input checked="" type="checkbox"/> Advanced View	<input type="checkbox"/> Use old BLAST report format	Reset form to defaults
Alignment View	Pairwise					
Display	<input checked="" type="checkbox"/> Graphical Overview	<input checked="" type="checkbox"/> Linkout	<input checked="" type="checkbox"/> Sequence Retrieval	<input type="checkbox"/> NCBI-gi		
Masking	Character: Lower Case	Color: Grey				
Limit results	Descriptions: 100	Graphical overview: 100	Alignments: 100			
Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.						
Enter organism name or id—completions will be suggested <input type="checkbox"/> Exclude <input data-bbox="1052 792 1072 806" type="button" value="+"/>						
Entrez query: <input type="text"/>						
Expect Min: <input type="text"/> Expect Max: <input type="text"/>						
Percent Identity Min: <input type="text"/> Percent Identity Max: <input type="text"/>						
Format for	<input type="checkbox"/> PSI-BLAST with inclusion threshold: <input type="text"/>					

подтверждаем выбор

Выравнивание, выданное BLAST

Длина найденного белка

Length=129 Number of matches=1

Вес в битах

Вес

E-value

Score = 78.6 bits (192), Expect = 9e-15, Method: Compositional matrix adjust.
Identities = 34/73 (47%), Positives = 50/73 (68%), Gaps = 0/73 (0%)

Query 17 YRLEEVQKHNNSSQSTWIIIVHHRIYDITKFLDEHPGGEEVLREQAGGDATENFEDVGHSTD 76
 Y EEV +H W+I++ ++Y+I+ ++DEHPGGEEV+ + AG DATE F+D+GHS +
Sbjct 11 YTHEEVAQHHTTHDDLWVILNGKVYNISNYIDEHPGGEEVILDCAGTDATEAFDDIGHSDE 70

Query 77 ARALSETFIIGEL 89
 A + E IG L
Sbjct 71 AHEILEKLYIGNL 83

Число сходных
"букв"

Число символов гэпа

Число совпадений

Длина выравнивания

Якорь (в примере — длины 3)

Как работает BLAST?

Query: GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVED

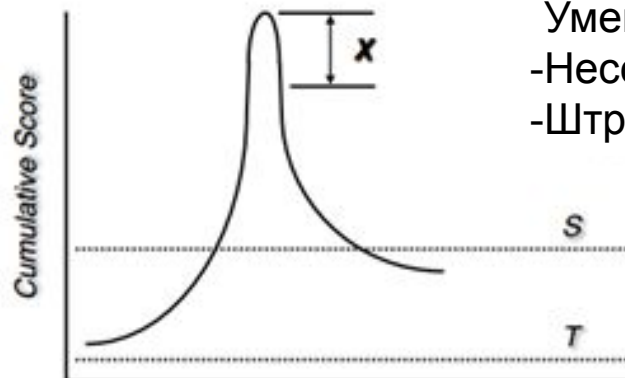
Сходные
слова

PQG	18	=7+5+6
PEG	15	
PRG	14	
PKG	14	
PNG	13	
PDG	13	
PHG	13	
PMG	13	
PSG	13	
PQA	12	
PQN	12	
etc.		

Порог на
score

Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L TP+G R++ +W+ +P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330

Поиск коротких сходных
слов (якорей)



Уменьшение за счёт:
-Несовпадений
-Штрафов за гэпы

Расширение

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.



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.


Параметры сервиса


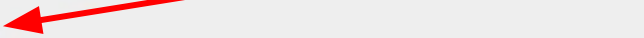
▼ **Algorithm parameters**


General Parameters

Max target sequences 
Select the maximum number of aligned sequences to display 


Short queries ☒ Automatically adjust parameters for short input sequences 


Expect threshold 


Word size  

Max matches in a query range 


Scoring Parameters



Matrix 

Gap Costs Existence: 11 Extension: 1 

Compositional adjustments 

Filters and Masking

Filter ☐ Low complexity regions 

Mask ☐ Mask for lookup table only 
☐ Mask lower case letters 

Длина слова

Длина слова

Одним из параметров BLAST является длина слова (word size). Это начальная длина якоря для поиска (см. слайд 19, где длина слова равна 3).

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

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