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, …)

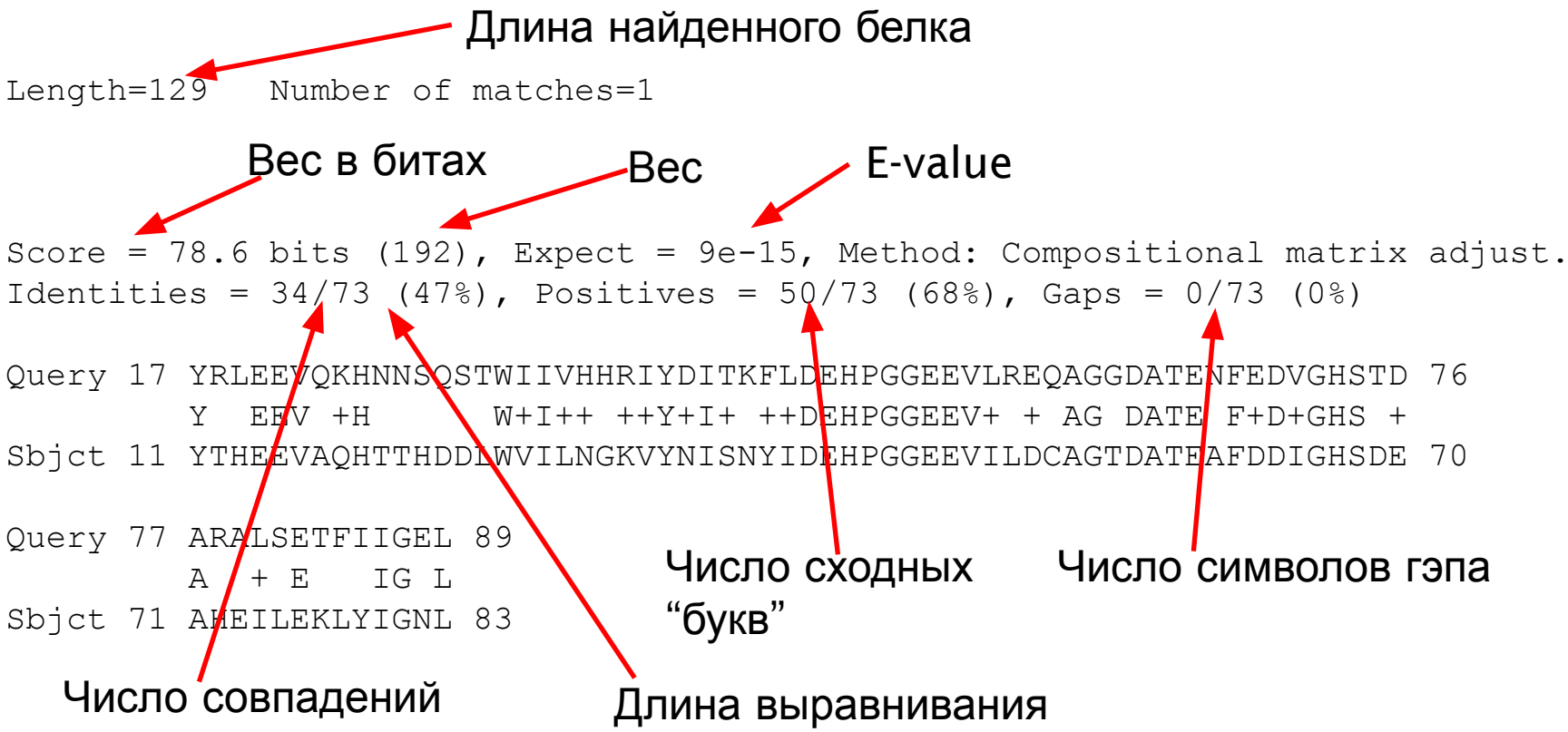
Что выдает BLAST?

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



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):

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

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

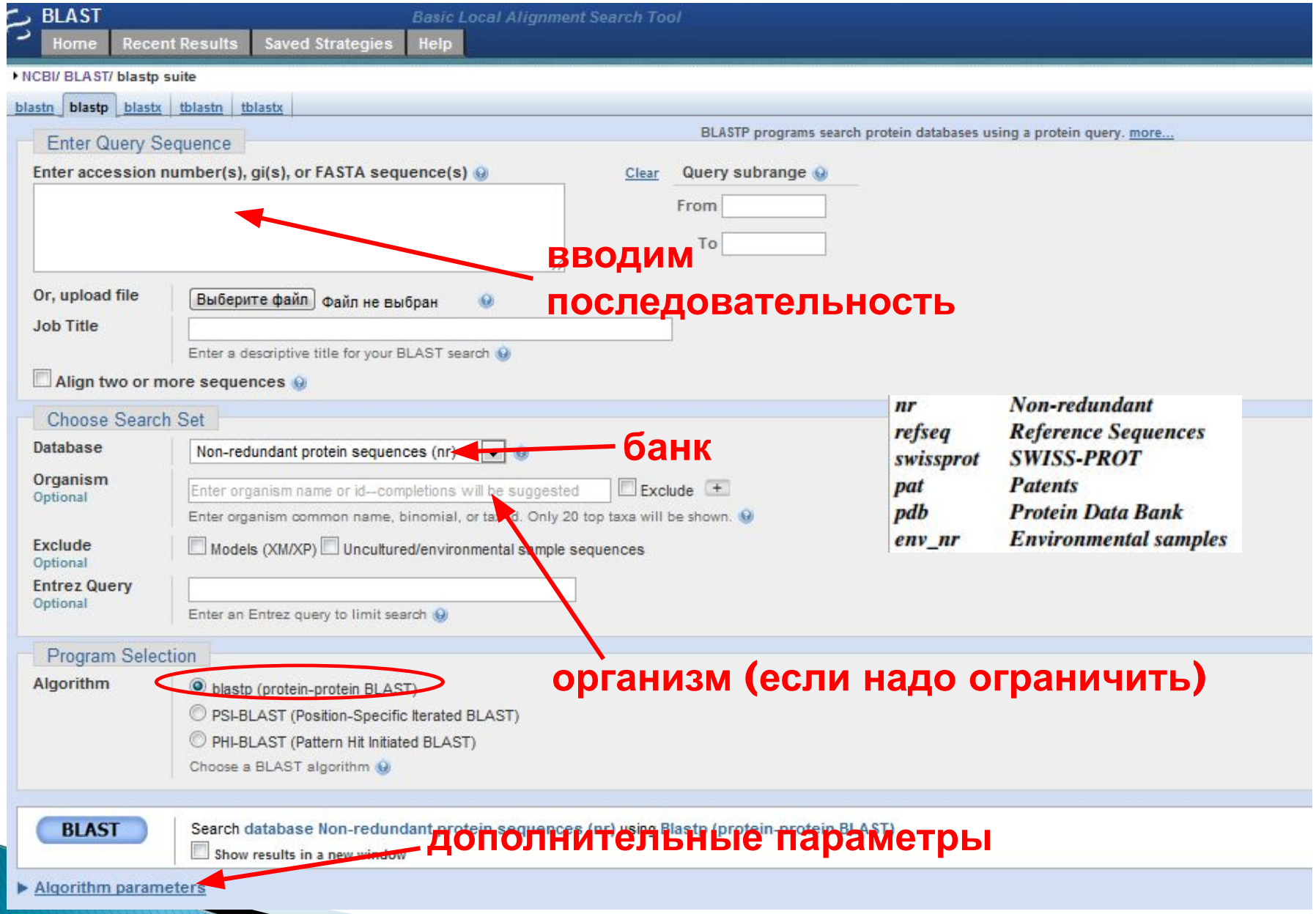
n – размер базы данных

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

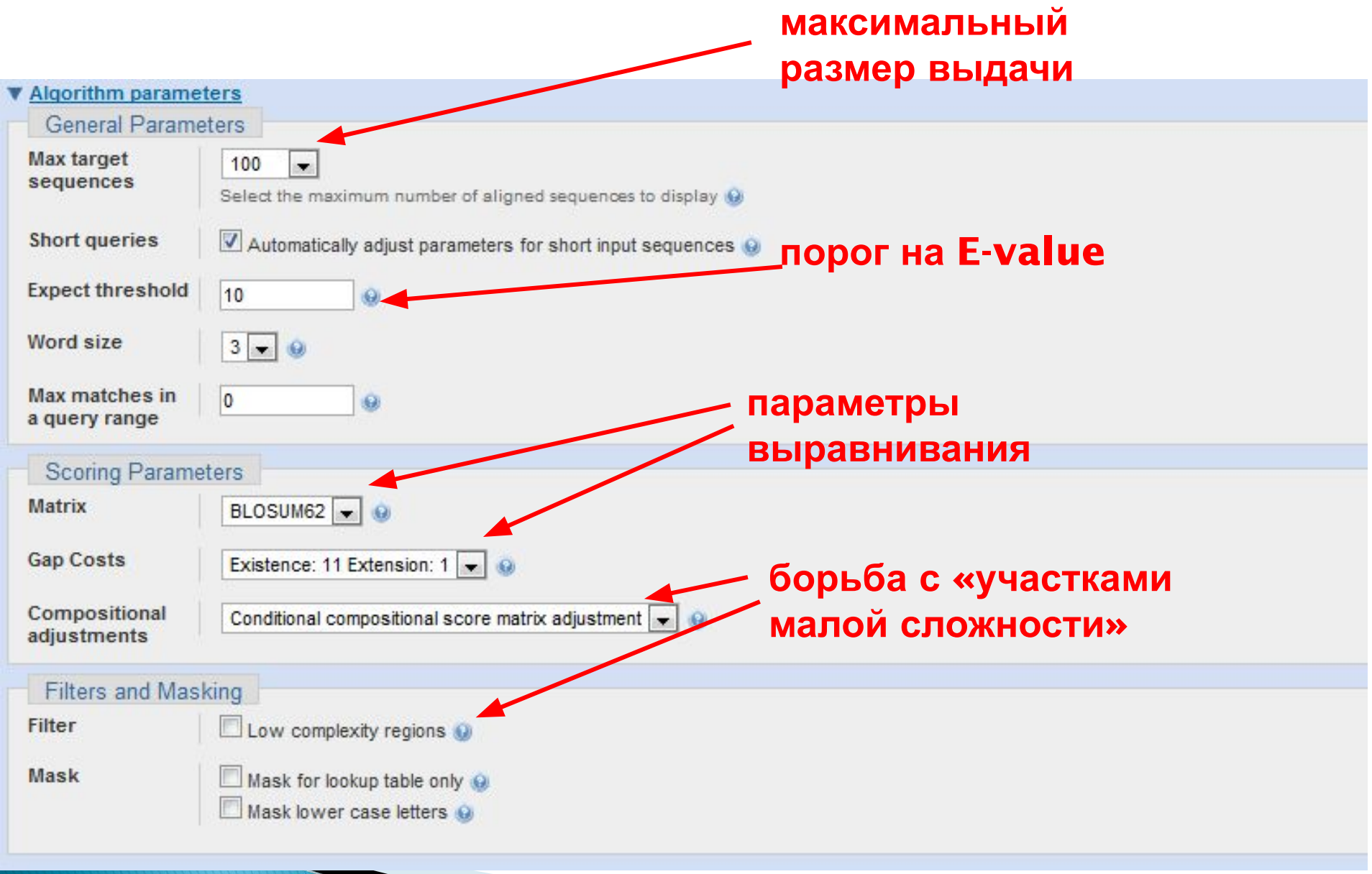
Нетрудно подсчитать, что $B = (\lambda S - \ln K)/\ln 2$

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



Дополнительные параметры



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   EPPQAVQPPPEPVVEPEPEPEPIPEP-PKEAPVVIEKPKPKPKPKPKPVKKVQEQPKRDV  114
            EP    +P PEP  EPEPEPEP PEP P+  P    +P+P+P+P+P+P  + + +P+ +
Sbjct  243  EPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEPEP  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

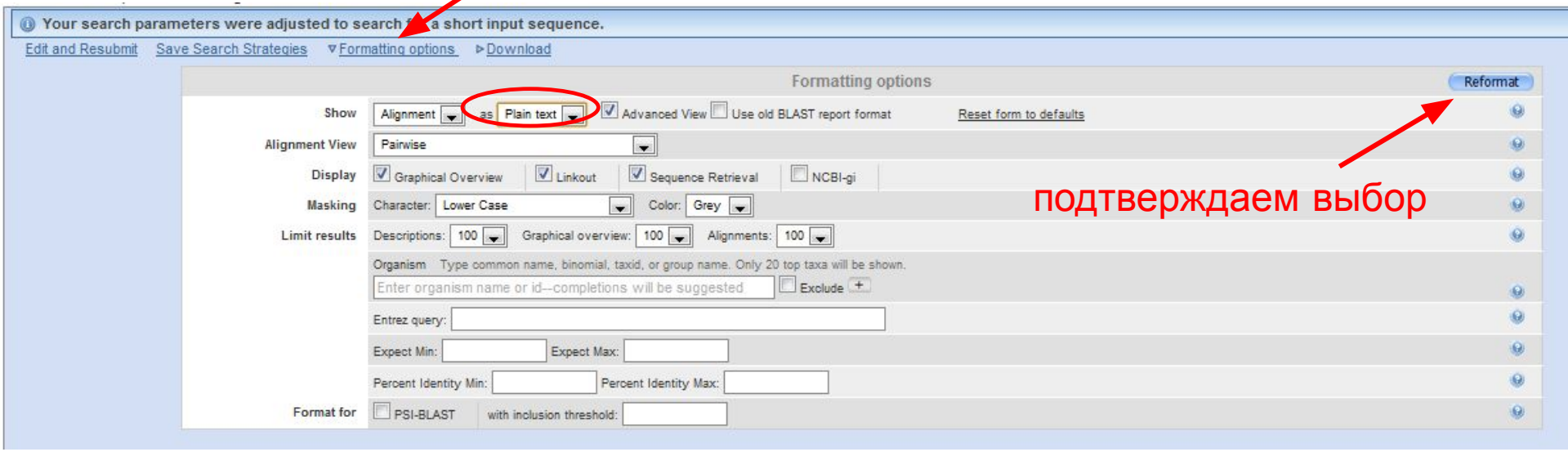
Определяется как участок с смещенным составом (biased composition)

- Гомополимерные участки
- Короткие повторы
- Перепредставленность отдельных остатков

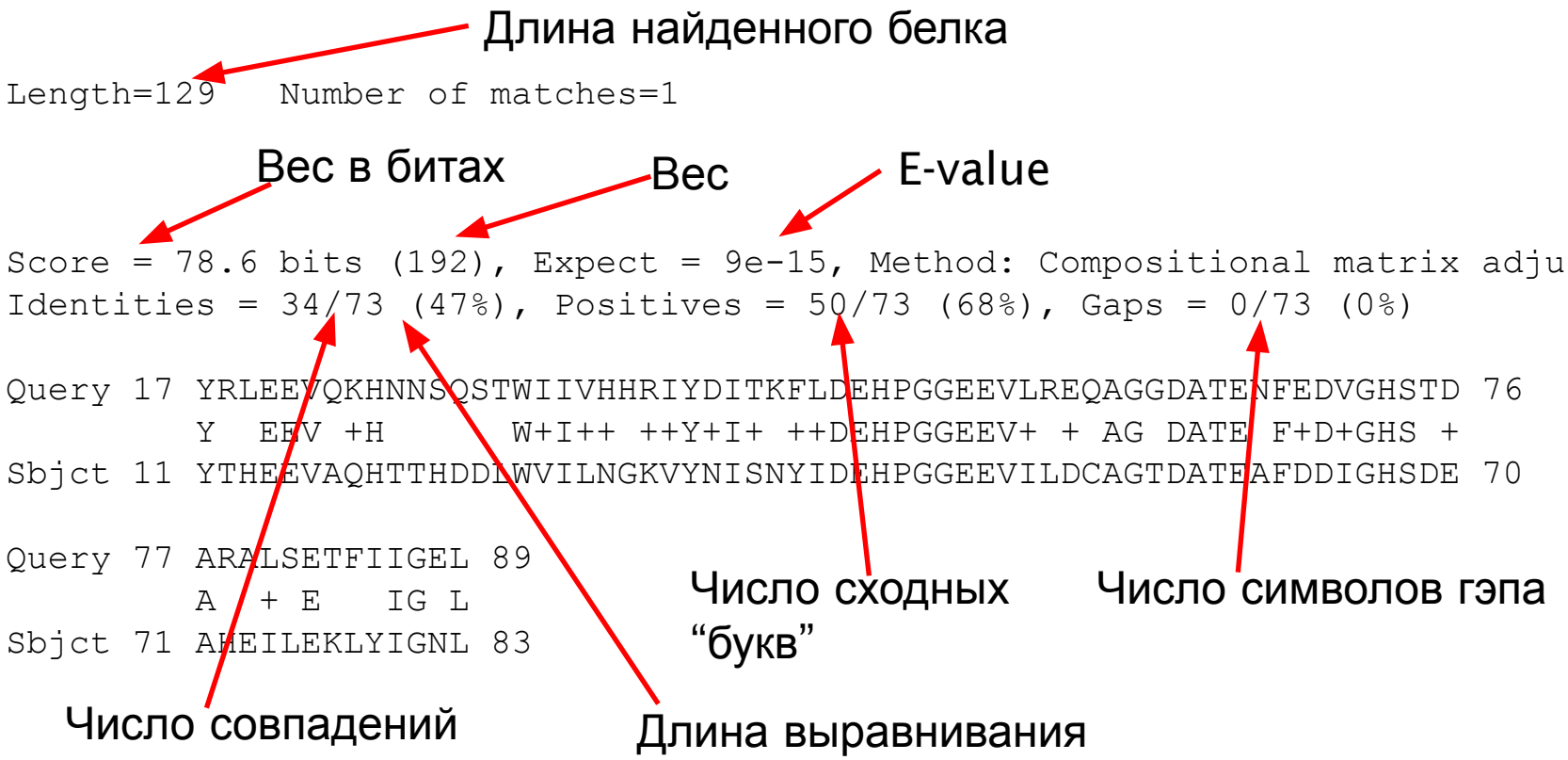
✔Может мешать анализу последовательностей
✔Вычисление E-value (параметры *K* и λ) опирается на среднее по всем белкам распределение частот аминокислотных остатков
✔Обычно ведет к ложным предсказаниям гомологии (false positives)
✔Лучше использовать «Compositional adjustment» (по умолчанию включен)

Переход к текстовому виду

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



Выравнивание, выданное BLAST



Как работает BLAST?

Якорь (в примере —длины 3)

```
Query:   GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVED
```

Сходные слова

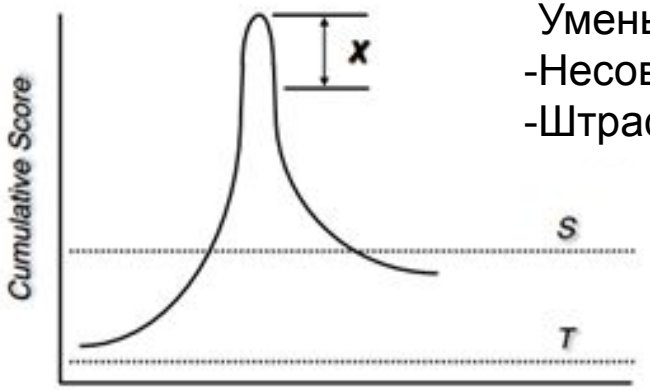
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

=7+5+6

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

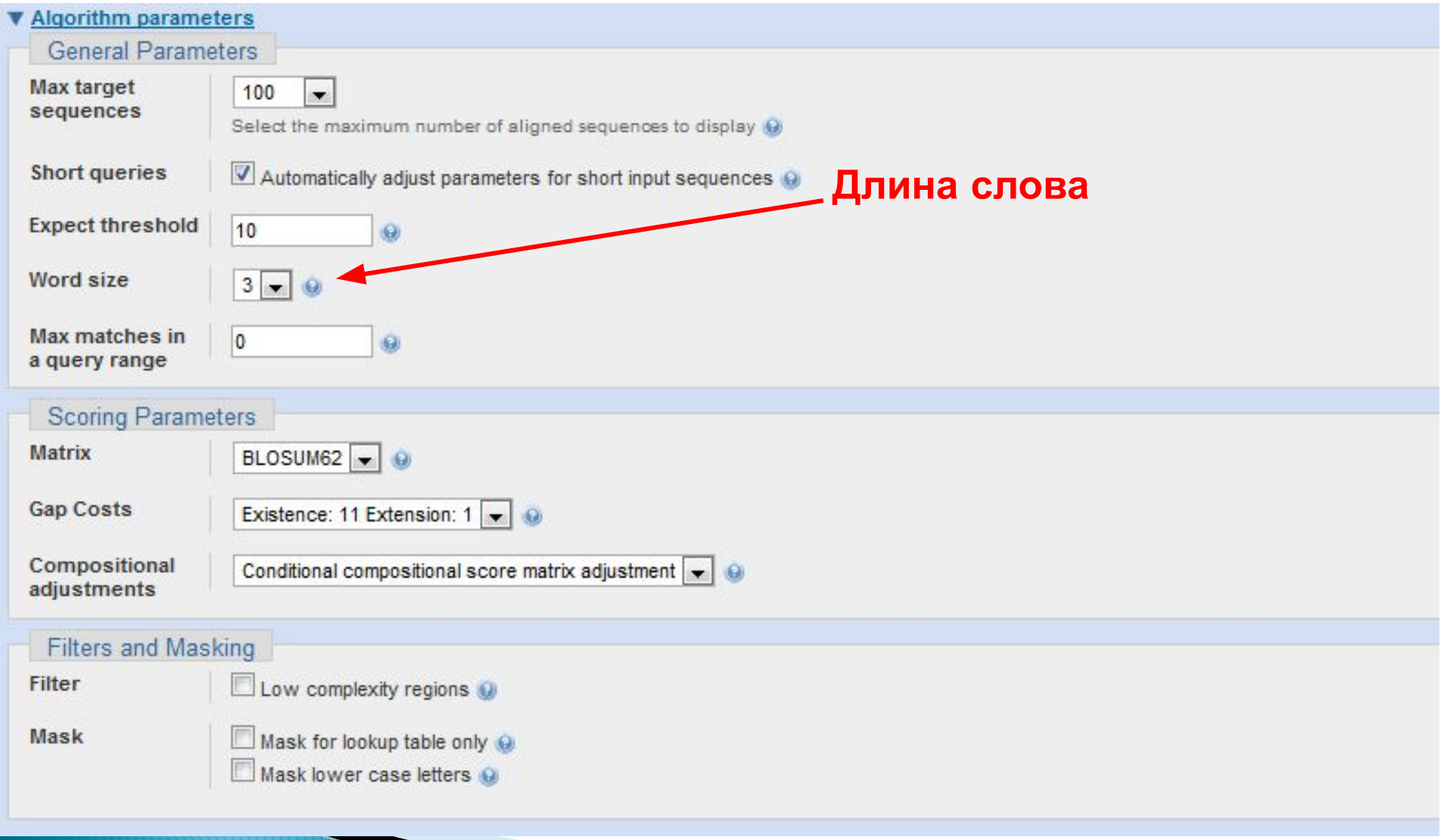
Расширение

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



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.