

BLAST

Школа биоинформатики,
осень 2016

Basic Local Alignment Search Tool

На входе:

- биологическая последовательность (запрос, “query”);
- банк биологических последовательностей.

На выходе:

- лучшие выравнивания запроса с последовательностями из банка.

Важно:

- выравнивания локальные (не глобальные);
- выравнивания не обязательно оптимальны по весу. Быстрый алгоритм позволяет за разумное время искать в большом банке, но может пропустить хорошее выравнивание.

Два воплощения

Online

The screenshot shows the NCBI BLAST web interface. At the top, there's a blue header with the BLAST logo and navigation links: Home, Recent Results, Saved Strategies, and Help. Below this, it says 'NCBI/BLAST/blastn suite' and 'Standard Nucleotide BLAST'. The main section is titled 'Enter Query Sequence' and contains a large text box for 'Enter accession number(s), gi(s), or FASTA sequence(s)'. To the right of this box are links for 'Clear' and 'Query subrange' with 'From' and 'To' input fields. Below the text box is a 'Browse...' button and the text 'No file selected.'. There's also a 'Job Title' field with the prompt 'Enter a descriptive title for your BLAST search'. A checkbox labeled 'Align two or more sequences' is present. The 'Choose Search Set' section includes radio buttons for 'Human genomic + transcript', 'Mouse genomic + transcript', and 'Others (nr etc.):'. A dropdown menu shows 'Nucleotide collection (nr/nt)'. There's an 'Organism' field with the prompt 'Enter organism name or id--completions will be suggested' and an 'Exclude' checkbox. At the bottom, it says 'Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.'

Локальный
(standalone)

```
yury@lmde> blastn
BLAST query/options error: Either a BLAST database or subject sequence(s) must be specified
Please refer to the BLAST+ user manual.
[1] yury@lmde> 
[0] 0: bash 1: python 2: zsh 3: zsh 4: zsh- 5: zsh* 6: zsh 7: zsh 8: zsh 9: zsh
```

Пять видов BLAST

Программа	Запрос	Банк
BLASTN	НК	НК
BLASTP	Белок	Белки
BLASTX	НК	Белки
TBLASTN	Белок	НК
TBLASTX	НК	НК

Standalone BLAST

(точнее, BLAST+)

Имеется файл с запросом (или многими запросами) и файл с «банком»

Банком может быть, например, чей-либо геном или протеом или даже «сырые» риды...

Оба файла – в формате Fasta!

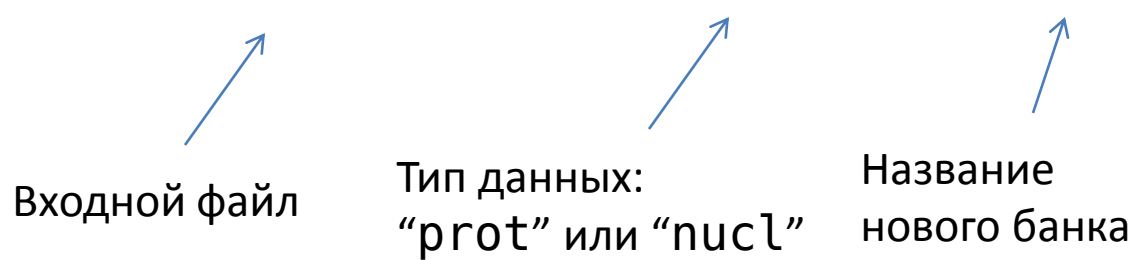
Этапы работы:

- форматирование банка (программа makeblastdb)
- собственно поиск (blastp или blastn или blastx или tblastn)

Подробное описание: <https://www.ncbi.nlm.nih.gov/books/NBK279690/>

Форматирование банка

```
> makeblastdb -in db.fasta -dbtype prot -out mydb
```



Входной файл

Тип данных:
“prot” или “nuc”

Название
нового банка

В результате работы образуются три файла с расширениями

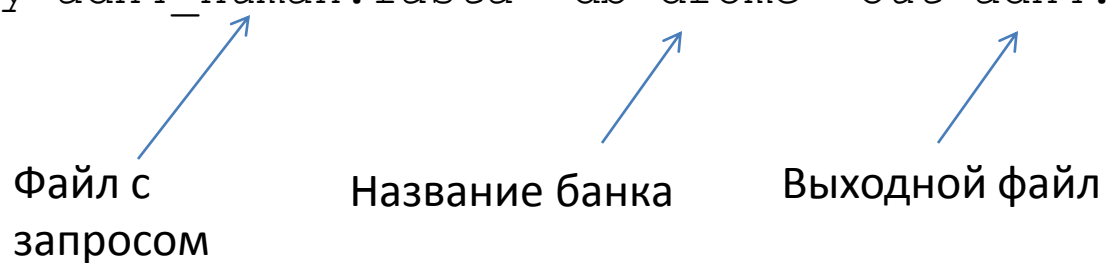
.phr, .pin, .psq для типа prot

.nhr, .nin, .nsq для типа nuc

BLASTP

```
> blastp -query adh4_human.fasta -db drome -out adh4.blastp
```

Файл с
запросом



Название банка

Выходной файл

Выходной файл состоит из четырёх частей:

- справочный заголовок (версия программы, на что ссылаться, размер банка, длина запроса);
- список «хитов» (то есть находок, последовательностей банка, в которых нашлись лучшие выравнивания);
- выравнивания;
- использованные параметры.

Список находок

Sequences producing significant alignments:				Score (Bits)	E Value
ADHX_DROME	P46415	Alcohol dehydrogenase class-3 (1.1.1.1) (Alco...		430	1e-149
MECR_DROME	Q9V6U9	Probable trans-2-enoyl-CoA reductase, mitocho...		33.5	0.032
SMBT_DROME	Q9VK33	Polycomb protein Sfmbt (Scm-like with four MB...		30.0	0.52
PDI_DROME	P54399	Protein disulfide-isomerase (PDI) (dPDI) (5.3....		28.9	0.95
DHE3_DROME	P54385	Glutamate dehydrogenase, mitochondrial (GDH) ...		28.5	1.3
YEMA_DROME	P25992	Yemanuclein-alpha		26.6	5.3
RM39_DROME	Q9VUJ0	39S ribosomal protein L39, mitochondrial (MRP...		25.8	7.6

Связь E-value и веса в битах:

$$\text{E-value} = mn2^{-B} = mnK \cdot \exp(-\lambda S)$$

где S – обычный вес, m – длина последовательности, n – суммарная длина банка, K и λ зависят от матрицы замен и штрафов за гэпы.

Упражнение: выразить B через S , K , λ .

Смысл: хочется, чтобы B имел примерно один смысл для разных матриц замен и штрафов за гэпы.

Выравнивание

```
> MECR_DROME Q9V6U9 Probable trans-2-enoyl-CoA reductase, mitochondrial  
(1.3.1.38) (Precursor)  
Length=357
```

```
Score = 33.5 bits (75), Expect = 0.032, Method: Compositional matrix adjust.  
Identities = 24/83 (29%), Positives = 40/83 (48%), Gaps = 1/83 (1%)
```

```
Query 17 EAGKPLCIEEVEVAPPKAHEVRIQIIATSLCHTDATVIDSKFE-GLAFPVIVGHEAAGIV 75  
      E  + L + E ++ PK ++V ++I+A + D I K+ FP + G+E V  
Sbjct 33 EPQEVQLQLVEDKLPDPKDNQVLVKILAAPINPADINTIQGKYPVKPKFPAVGGNECVAEV 92  
  
Query 76 ESIGPGVTNVKPGDKVIPLYAPL 98  
      +G V + G VIPL + L  
Sbjct 93 ICVGDKVKGFEGQHVIPLASGL 115
```

Query – это последовательность запроса, Subject – последовательность из банка (в данном случае MECR_DROME). Выравнивание локальное, сопоставляет участок 17–98 из запроса с участком 33–115 из находки. Знаком “+” обозначены сопоставления неодинаковых букв, вносящие тем не менее положительный вклад в вес выравнивания (то есть для которых значение из матрицы > 0).

«Хвост» выходного файла: использованные параметры

Lambda	K	H	a	alpha
0.320	0.138	0.418	0.792	4.96

Gapped

Lambda	K	H	a	alpha	sigma
0.267	0.0410	0.140	1.90	42.6	43.6

Effective search space used: 484173366

Параметры подсчёта
bit score и E-value

Вместо *mn*

Database: drome.fasta

Posted date: Jan 22, 2014 4:20 PM

Number of letters in database: 1,948,893

Number of sequences in database: 3,203

Matrix: BLOSUM62

Gap Penalties: Existence: 11, Extension: 1

Neighboring words threshold: 11

Window for multiple hits: 40

Параметры поиска

Полезные параметры

`–outfmt 6` или `–outfmt 7`

табличная выдача (часто бывает удобна для автоматической обработки, особенно если запросов много)

Кстати: если поместить во входной файл (query) не одну последовательность, а много, то будет выдан результат для всех, причём существенно быстрее, чем при поиске по каждой последовательности отдельно.

`–outfmt 5`

выдача в формате xml (для автоматической обработки любителями xml)

`–evaluate 0.001`

ограничение не E-value, чтобы не перегружать выдачу слабыми находками

`–num_alignments 1000`

если в 250, выдаваемые по умолчанию, не помещаются интересные находки

Полный список параметров выдаётся по команде `blastp –help`, подробности – на <https://www.ncbi.nlm.nih.gov/books/NBK279675/>

Online BLAST на NCBI

The screenshot shows the NCBI BLAST website. The browser address bar displays 'blast.ncbi.nlm.nih.gov/Blast.cgi'. The page header includes the NIH logo, 'U.S. National Library of Medicine', the NCBI logo, 'National Center for Biotechnology Information', and a 'Sign in to NCBI' link. The main navigation bar features 'BLAST®' and links for 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The main content area is titled 'Basic Local Alignment Search Tool'. It describes BLAST as a tool for finding regions of similarity between biological sequences. A 'Learn more' link is provided. A news box on the right announces 'BLAST+ 2.5.0 released' with details about new features like HTTPS support and a new taxonomic organism report. Below the main content, the 'Web BLAST' section offers three options: 'Nucleotide BLAST' (nucleotide to nucleotide), 'blastx' (translated nucleotide to protein), and 'tblastn' (protein to translated nucleotide). A 'Protein BLAST' option (protein to protein) is also shown. At the bottom, a link for 'BLAST Genomes' is visible.

BLAST: Basic Local Alignment ...

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

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

BLAST+ 2.5.0 released

NEWS

The new version offers support for HTTPS, accession.version as the primary sequence identifier, support for composition-based statistics with RPSTBLASTN, and a new taxonomic organism report.

Fri, 23 Sep 2016 17:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

Protein BLAST
protein ► protein

BLAST Genomes

BLASTP на NCBI

Protein BLAST: search protein... x +

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LI

Search

NIH U.S. National Library of Medicine NCBI Sign in to NCBI

BLAST >> blastp suite Home Recent Results Saved Strategies Help

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

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

Enter Query Sequence

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

Or, upload file [Browse...](#) No file selected.

Job Title [Enter a descriptive title for your BLAST search](#)

☒ Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr) [v](#) [?](#)

Organism [Optional](#) [Enter organism name or id--completions will be suggested](#) ☐ Exclude [+](#)

Exclude [Optional](#) ☐ Models (X/M/X/P) ☐ Uncultured/environmental sample sequences

Entrez Query [Optional](#) [Enter an Entrez query to limit search](#) [YouTube](#) [Create custom database](#)

Выберите банк
(для начала лучше
Swiss-Prot)

Можно ограничить поиск
заданным таксоном

BLASTP на NCBI

Protein BLAST: search protein... x +

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LI

Search

NIH U.S. National Library of Medicine NCBI Sign in to NCBI

BLAST >> blastp suite Home Recent Results Saved Strategies Help

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

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

Enter Query Sequence

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

Or, upload file [Browse...](#) No file selected.

Job Title [Enter a descriptive title for your BLAST search](#)

☐ Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr) [v](#) [?](#)

Organism [Optional](#) [Enter organism name or id--completions will be suggested](#) ☐ Exclude [+](#)

[Exclude](#) [Optional](#) ☐ Models (X/M/X/P) ☐ Uncultured/environmental sample sequences

Entrez Query [Optional](#) [Enter an Entrez query to limit search](#) [YouTube](#) [Create custom database](#)

Выберите банк
(для начала лучше
Swiss-Prot)

Можно ограничить поиск
заданным таксоном

Нижняя часть страницы: параметры

Protein BLAST: search protein...

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LIT

Search

Choose Search Set

Database: UniProtKB/Swiss-Prot(swissprot)

Organism: Fungi (taxid:4751) ☐ Exclude ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Exclude: ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query: [YouTube](#) [Create custom database](#)

Program Selection

Algorithm: ☒ blastp (protein-protein BLAST) ☐ PSI-BLAST (Position-Specific Iterated BLAST) ☐ PHI-BLAST (Pattern Hit Initiated BLAST) ☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

BLAST

Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)

☐ Show results in a new window

[Algorithm parameters](#)

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

BLAST is a registered trademark of the National Library of Medicine.

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | NLM | NIH | DHHS

Часто имеет смысл
посмотреть и изменить
параметры

Параметры Online-BLAST

The screenshot shows the NCBI Protein BLAST search page. The browser address bar displays the URL: `blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasth`. The page title is "Protein BLAST: search protein...".

Algorithm parameters

General Parameters

- Max target sequences:** 100 (indicated by a red arrow). Below the input field, it says "Select the maximum number of aligned sequences to display".
- Short queries:** ☒ Automatically adjust parameters for short input sequences.
- Expect threshold:** 10 (indicated by a red arrow). Below the input field, it says "Select the maximum number of aligned sequences to display".
- Word size:** 6 (indicated by a red arrow). Below the input field, it says "Select the maximum number of aligned sequences to display".
- 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.

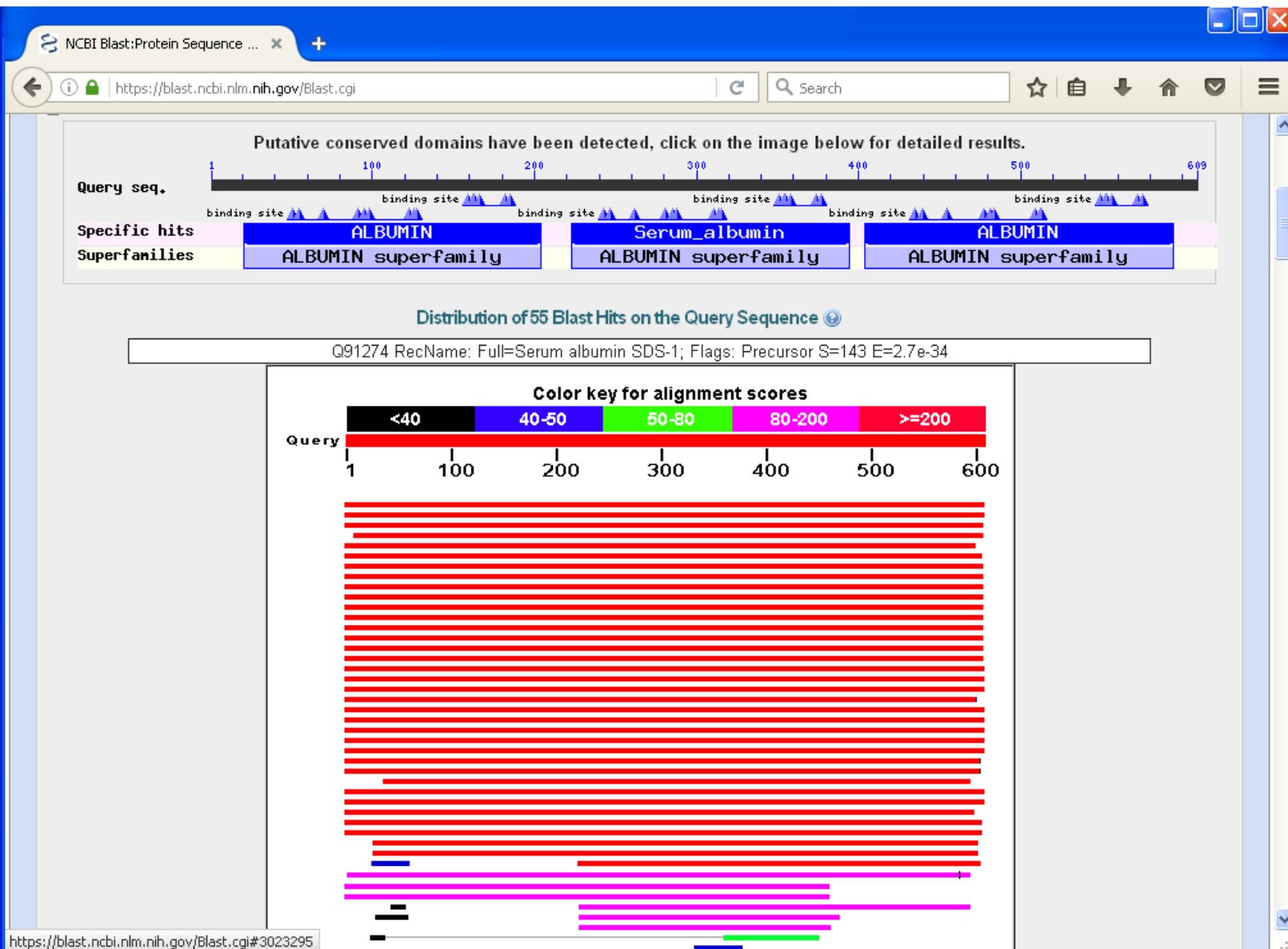
BLAST Search **database UniProtKB/Swiss-Prot(swissprot)** using **Blastp (protein-protein BLAST)**

☐ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign. [Restore default search parameters](#)

Длина слова оптимизирована под скорость, а не под чувствительность!
Ищете слабосходные последовательности – ставьте 2 или 3.

Выдача Online BLAST



TBLASTN и BLASTX

Всё практически так же, как у BLASTP, только белковые последовательности запроса (у TBLASTN) или банка (у BLASTX) возникают как формальные трансляции в шести рамках соответственно нуклеотидного запроса или нуклеотидного банка.

BLASTX – способ черновой аннотации кодирующих участков в вашей нуклеотидной последовательности.

TBLASTN помогает идентифицировать гены гомологов интересующего вас белка в неаннотированном геноме, или даже в сырых ридсах, или в коллекции EST.

Online – те же проблемы с длиной слова и размером выдачи по умолчанию.

Три подвида BLASTN

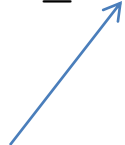
Программа	Затравка*	Цели
megablast	28 нк	Найти запрос в банке
discontiguous megablast	11 нк	Найти близкие гомологи
blastn	11 нк	Найти гомологи

* по умолчанию

BLASTN

```
> blastn -query trna_leu.fasta -db ec -out tl.blastn
```

Файл с
запросом



Название банка



Выходной файл



Казалось бы, всё как в `blastp`, но в данном примере будет запущен **megablast**: очень быстрый, но предназначенный для поиска в банке только последовательностей, совпадающих или почти совпадающих с запросом.

BLASTN

```
> blastn -task blastn query trna_leu.fasta -db ec -out tl.blastn
```



В таком варианте уже можно искать гомологи.

Но длина слова по умолчанию 11, что многовато.

Вес совпадения по умолчанию 2, вес несовпадения -3, что позволяет находить лишь гомологи с идентичностью > 70%.

BLASTN

```
> blastn -task blastn query trna_leu.fasta -db ec -out tl.blastn
```



В таком варианте уже можно искать гомологи.

Но длина слова по умолчанию 11, что многовато.

Вес совпадения по умолчанию 2, вес несовпадения –3, что позволяет находить лишь гомологи с идентичностью > 70%.

Поэтому лучше добавить опции: `–word_size 4 –penalty –1 –reward 1`

Длина слова 4 заметно замедляет работу. Разумный компромисс между чувствительностью и скоростью – длина слова 7.

Для гомологов с совсем низкой (порядка 50%) идентичностью разумно поставить `–penalty –4 –reward 5`, но при этом придётся поменять значения `–gapopen` и `–gapextend`

(см. о том, какие сочетания четырёх параметров допустимы, в

<https://www.ncbi.nlm.nih.gov/books/NBK279678/>).

BLASTN на NCBI

The screenshot shows the NCBI BLASTN web interface. The browser address bar displays the URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_. The page header includes the NIH logo, "U.S. National Library of Medicine", and "NCBI National Center for Biotechnology Information". The main heading is "BLAST® >> blastn suite". Below this, there are tabs for "blastn", "blastp", "blastx", "tblastn", and "tblastx". The "blastn" tab is selected. The page title is "Standard Nucleotide BLAST". The interface is divided into several sections: "Enter Query Sequence", "Choose Search Set", and "Program Selection". The "Enter Query Sequence" section has a text input field for "Enter accession number(s), gi(s), or FASTA sequence(s)", a "Clear" button, and a "Query subrange" section with "From" and "To" input fields. Below this is a "Browse..." button for uploading a file. The "Choose Search Set" section has a "Database" section with radio buttons for "Human genomic + transcript", "Mouse genomic + transcript", and "Others (nr etc.):". The "Others (nr etc.):" option is selected, and a dropdown menu shows "Nucleotide collection (nr/nt)". There is also an "Organism" section with a text input field for "Enter organism name or id--completions will be suggested" and an "Exclude" checkbox. The "Program Selection" section has radio buttons for "Highly similar sequences (megablast)", "More dissimilar sequences (discontiguous megablast)", and "Somewhat similar sequences (blastn)". The "Highly similar sequences (megablast)" option is selected. Red arrows point from Russian text boxes to specific elements: one to the "Nucleotide collection (nr/nt)" dropdown, one to the "Organism" input field, one to the "More dissimilar sequences (discontiguous megablast)" radio button, and one to the "blastn" tab.

Enter Query Sequence

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

Clear

Query subrange

From

To

Or, upload file

Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database

☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):

Nucleotide collection (nr/nt)

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) ☐ Uncultured/environmental sample sequences

Limit to

Optional

☐ Sequences from type material

Entrez Query

Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

☒ Highly similar sequences (megablast)

☐ More dissimilar sequences (discontiguous megablast)

☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Выберите банк
(для начала лучше
RefSeq Genomic)

Можно ограничить поиск
заданным таксоном

Выберите алгоритм
(в зависимости от задачи)

BLASTN на NCBI: параметры

The screenshot shows the NCBI Nucleotide BLAST search page. The browser address bar displays the URL: `https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_`. The page features a search bar with a "BLAST" button and a checkbox for "Show results in a new window". Below the search bar, the "Algorithm parameters" section is expanded, showing three sub-sections: "General Parameters", "Scoring Parameters", and "Filters and Masking".

General Parameters

- Max target sequences:** 100 (dropdown menu). Description: Select the maximum number of aligned sequences to display.
- Short queries:** ☒ Automatically adjust parameters for short input sequences.
- Expect threshold:** 10 (text input).
- Word size:** 11 (dropdown menu). A red arrow points to this field with the annotation: "Минимум длины слова в Online версии равен 7."
- Max matches in a query range:** 0 (text input).

Scoring Parameters

- Match/Mismatch Scores:** 2,-3 (dropdown menu). A red arrow points to this field with the annotation: "Для максимальной чувствительности ставьте 1, -1".
- Gap Costs:** Existence: 5 Extension: 2 (dropdown menu).

Filters and Masking

- Filter:** ☒ Low complexity regions. A red arrow points to this checkbox with the annotation: "По умолчанию включена фильтрация участков малой сложности". ☐ Species-specific repeats for: Homo sapiens (Human) (dropdown menu).
- Mask:** ☒ Mask for lookup table only. ☐ Mask lower case letters.

At the bottom, there is another "BLAST" button and a search bar with the text "Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)".

Проблема малой сложности

... to appear soon

Удаленный бласт

Имя банка в NCBI

```
> blastn -task blastn -query query.fasta -db refseq_rna -out blast.out -evaluate  
0.001 -word_size 7 -outfmt 7 -remote -entrez_query 'arabidopsis[orgn]'
```

Ограничение по организму возможно
только в сочетании с `-remote`

Опция `-remote` означает «запускать на сервере NCBI по банку NCBI»

Бласт двух последовательностей

```
> blastn -task blastn -query seq1.fasta -subject seq2.fasta
```

Фактически программа локального выравнивания.

В online-версии выдаёт, кроме самих выравниваний, ещё и удобную графическую карту локального сходства двух последовательностей.