
DTU

Introduction to Bioinformatics 22111

Basic Local Alignment Search Tool (BLAST)

BLAST is the most popular algorithm in bioinformatics

Basic local alignment search tool

SF **Altschul**, W Gish, W Miller, EW Myers… - Journal of molecular …, **1990** - Elsevier

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal …

☆ Save 99 Cite Cited by 120750 Related articles All 56 versions Web of Science: 77938 »

Highly accurate protein structure prediction with **AlphaFold**

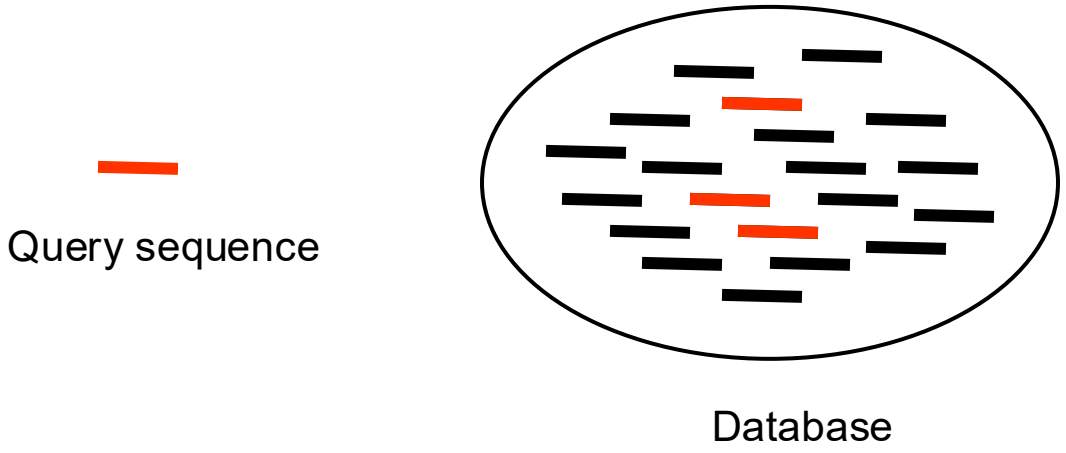
J Jumper, R Evans, A Pritzel, T Green, M Figurnov… - nature, 2021 - nature.com

… The neural network **AlphaFold** that we developed was … different model from our CASP13 **AlphaFold** system 10). The CASP … In CASP14, **AlphaFold** structures were vastly more accurate …

☆ Save 99 Cite Cited by 40704 Related articles All 35 versions Web of Science: 24784 »

Database searching

Using pairwise alignments to search databases for similar sequences



Database searching

Most common use of pairwise sequence alignments is to search databases for related sequences. For instance: find probable function of newly isolated protein by identifying similar proteins with known function.

Most often, ***local*** alignment (“Smith-Waterman”) is used for database searching: you are interested in finding out if ANY domain in your protein looks like something that is known.

Often, full Smith-Waterman is too time-consuming for searching large databases, so heuristic methods are used (fasta, BLAST).

Database searching: heuristic search algorithms

FASTA (Pearson 1985, 1988)

Uses heuristics to avoid calculating the full dynamic programming matrix

Speed up searches by an order of magnitude compared to full Smith-Waterman

The statistical side of FASTA is still stronger than BLAST

BLAST (Altschul 1990, 1997)

Uses rapid word lookup methods to completely skip most of the database entries

Extremely fast

One order of magnitude faster than FASTA

Two orders of magnitude faster than Smith-Waterman

Almost as sensitive as FASTA

BLAST flavors

BLASTN
Nucleotide query sequence
Nucleotide database

BLASTP
Protein query sequence
Protein database

BLASTX
Nucleotide query sequence
Protein database
Compares all six reading frames with the database

TBLASTN
Protein query sequence
Nucleotide database
”On the fly” six frame translation of database

TBLASTX
Nucleotide query sequence
Nucleotide database
Compares all reading frames of query with all reading frames of the database

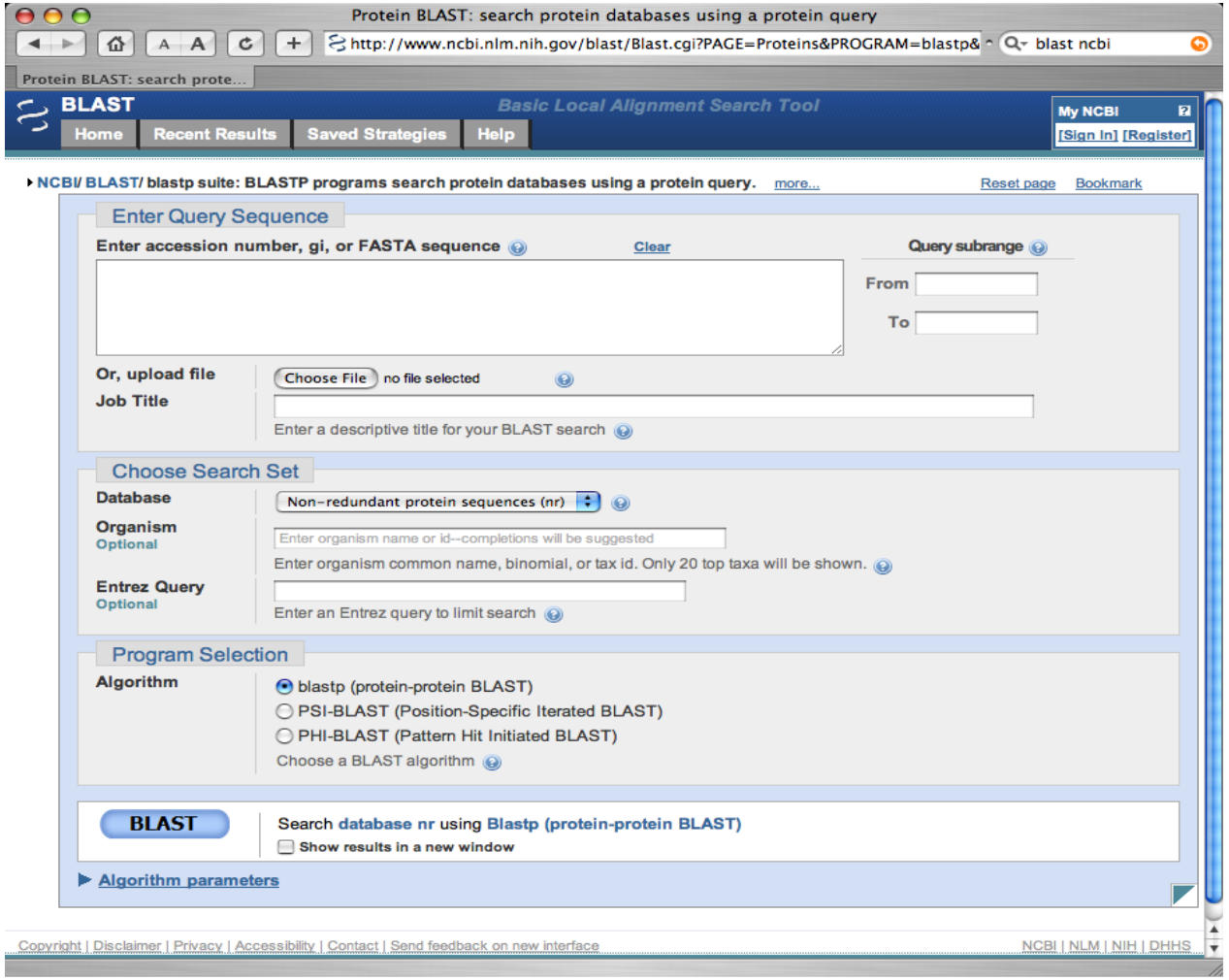
Searching on the web: BLAST at NCBI

Very fast computers dedicated to running BLAST searches

Many databases that are always up to date (e.g. NR and Human Genome)

Nice simple web interface

But you still need knowledge about BLAST to use it properly



When is a database hit significant?

- **Problem**:
  - Even unrelated sequences can be aligned (yielding a low score)
  - How do we know if a database hit is meaningful?
  - When is an alignment score sufficiently high?
- **Solution**:
  - Determine the range of alignment scores you would expect to get for random reasons (i.e., when aligning unrelated sequences).
  - Compare actual scores to the distribution of random scores.
  - Is the real score much higher than you'd expect by chance?

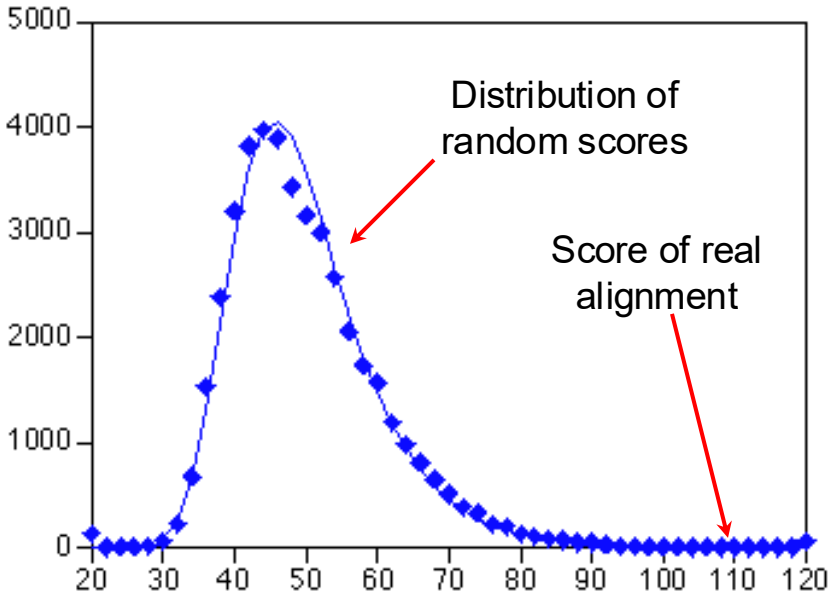
Distribution of random alignment scores

https://colab.research.google.com/drive/1rrzB1MLjk_lWaAuh-8WlgZeKCgGB9Q1u#scrollTo=si1_HtmJURK0

Significance of alignment score expressed as E-value

Searching a database of **unrelated** sequences results in scores following an extreme value distribution

The exact shape and location of the distribution depends on the exact nature of the database and the query sequence



E-value: the number of **random hits** to **expect** for any given score

Want E-values below 1 (the lower the better)

Significance of alignment score expressed as E-value

E-value / Expect-value:

Number of unrelated hits with an equal or better alignment score to expect due to strictly stochastic reasons.

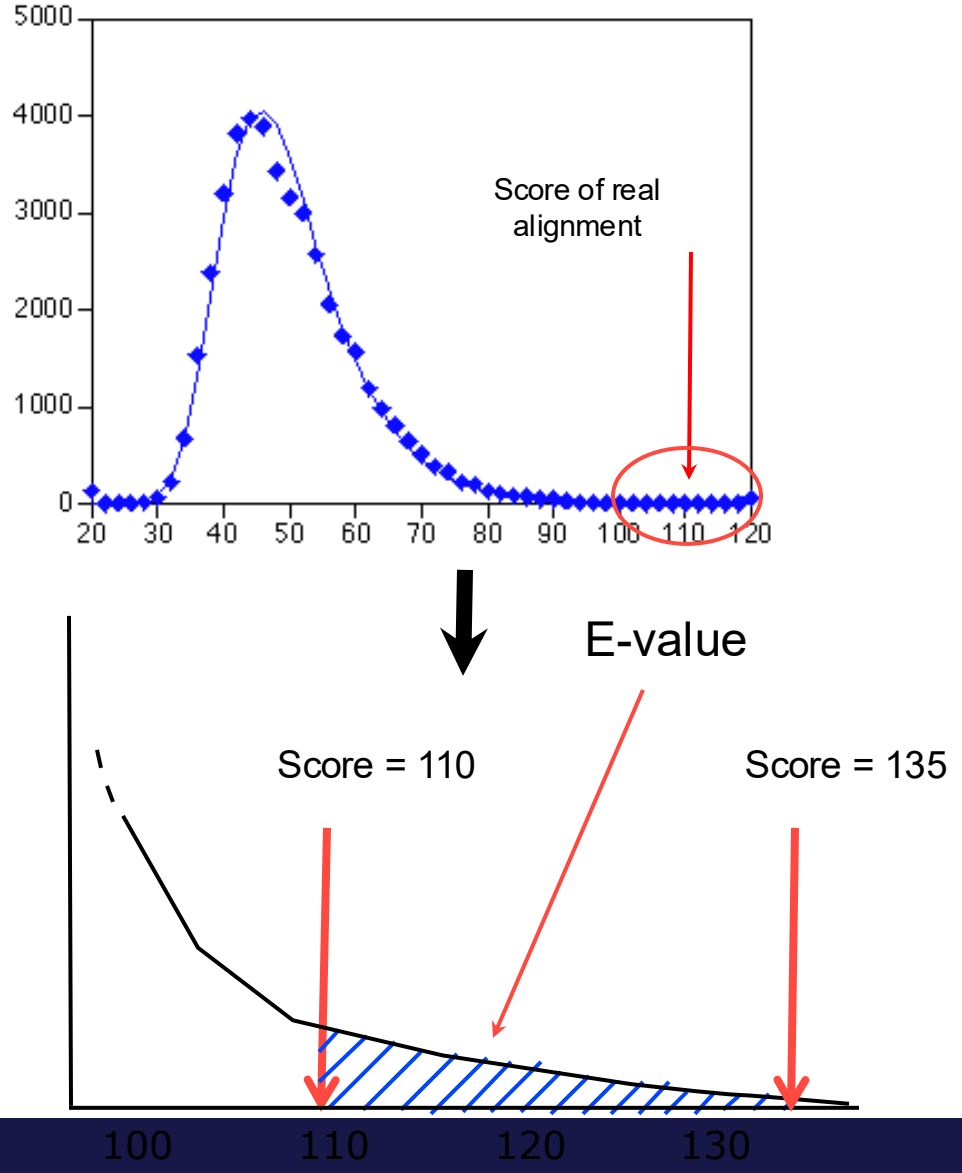
Example:

Alignment score = 110

E-value = 8.7

Alignment score = 135

E-value = 0.0001



BLAST heuristics

- BLAST speeds up the search >100x by pre-screening the database sequences and only performing the full Dynamic Programming on "*promising*" sequences.

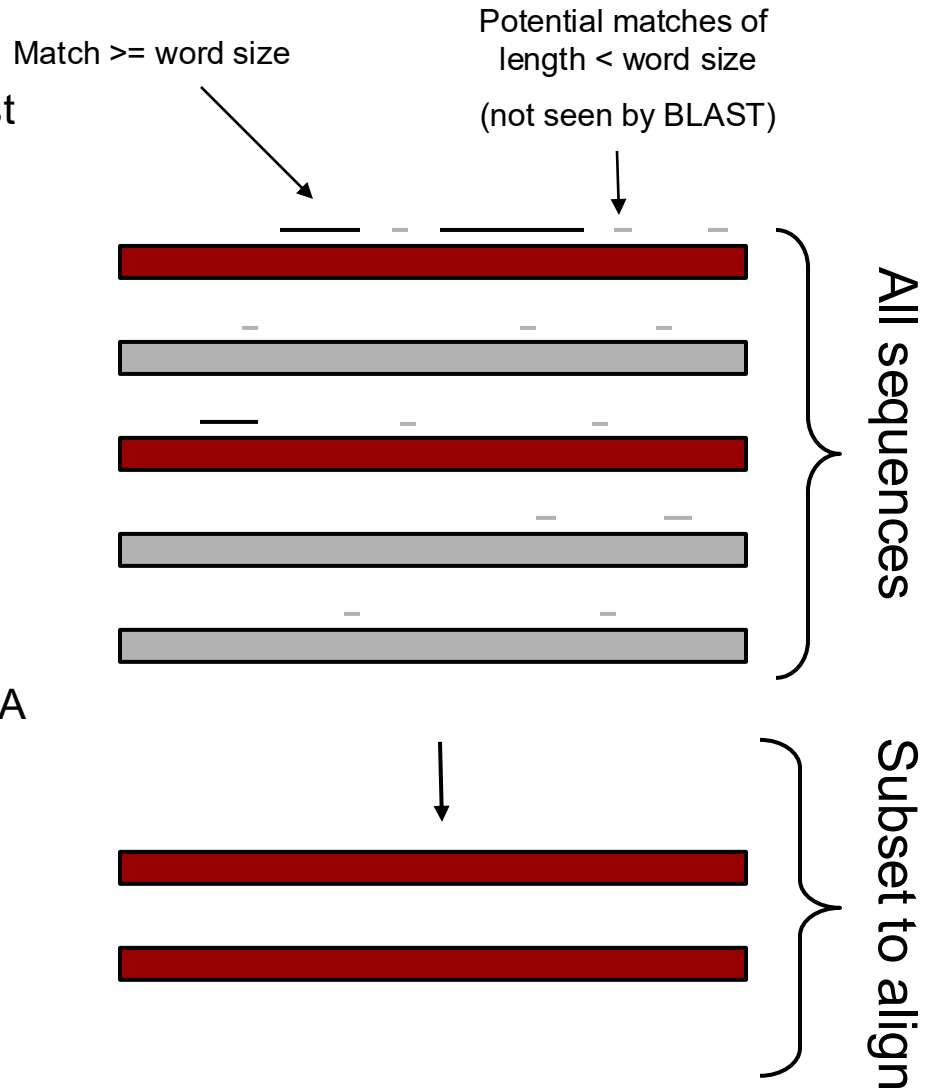
- Promising sequences: database sequences that have sub-strings ("words") which also occur in the query sequence.

- **BLASTN** and **BLASTP** use different criteria for overlap required for a sequence to be deemed promising.

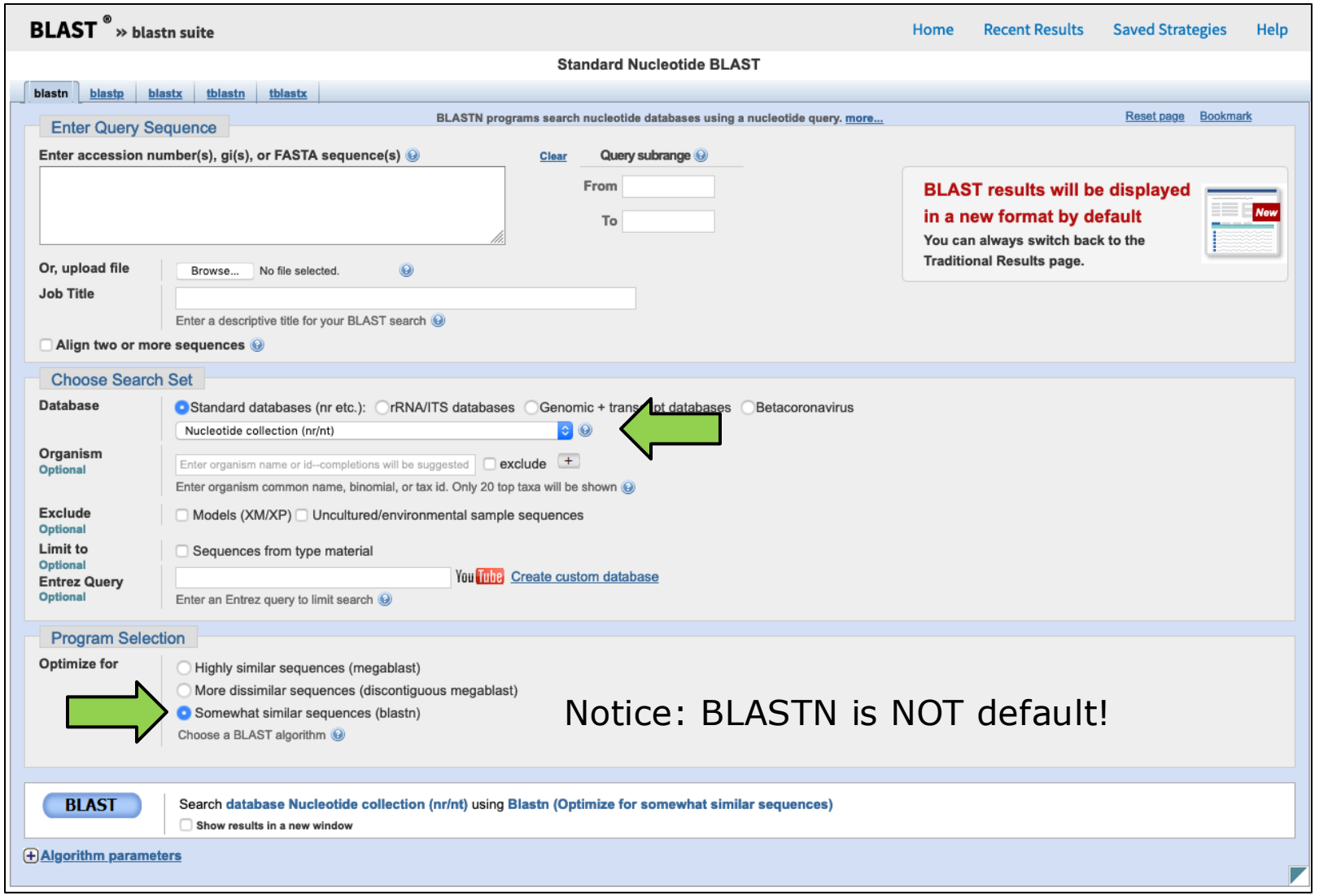
BLASTN - conceptual

- Heuristics:
  - Perfect match “word” of at least size: 7, 11 (default) or 15.
- Alignment matrix:
  - Match: **2**
  - Mismatch: **-3**
- Notice: All mismatches are equally penalized:
  - E.g. A:G == A:C == A:T
  - More advanced models for DNA evolution do exist.

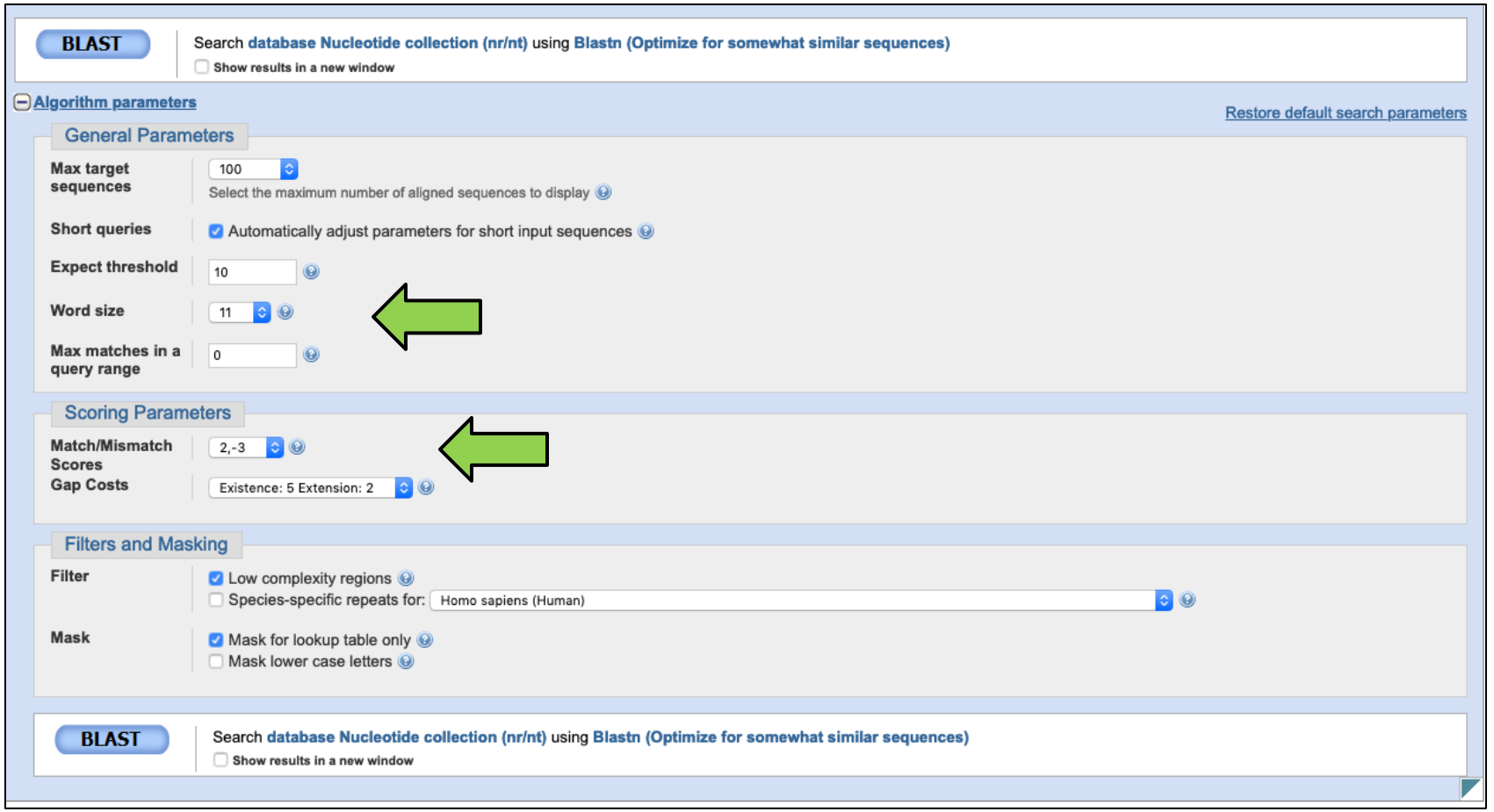
	A	C	G	T
A	2	-3	-3	-3
C	-3	2	-3	-3
G	-3	-3	2	-3
T	-3	-3	-3	2



BLASTN@NCBI

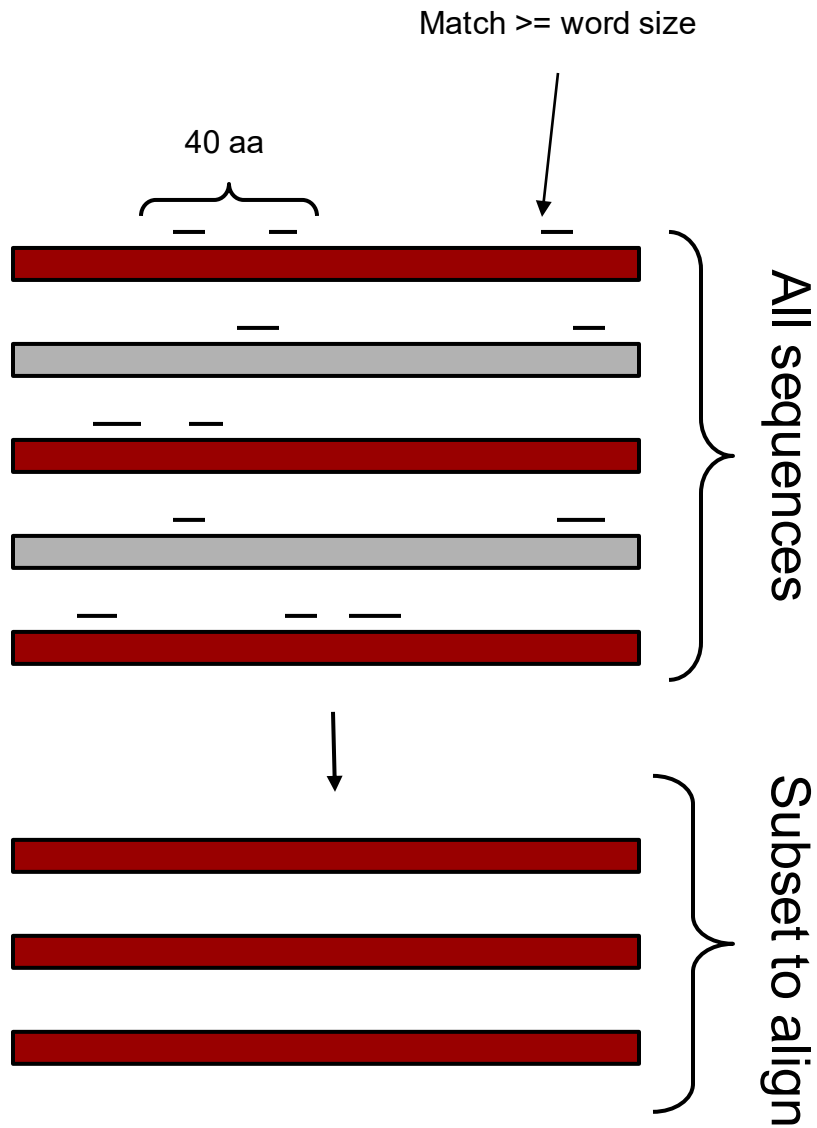

BLAST ® » blastn suite
Home
Recent Results
Saved Strategies
Help
Standard Nucleotide BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
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
Browse...
No file selected.
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Standard databases (nr etc.):
rRNA/ITS databases
Genomic + transcript databases
Betacoronavirus
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
YouTube
Create custom database
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
Notice: BLASTN is NOT default!
BLAST
Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
Show results in a new window
Algorithm parameters

BLASTN@NCBI

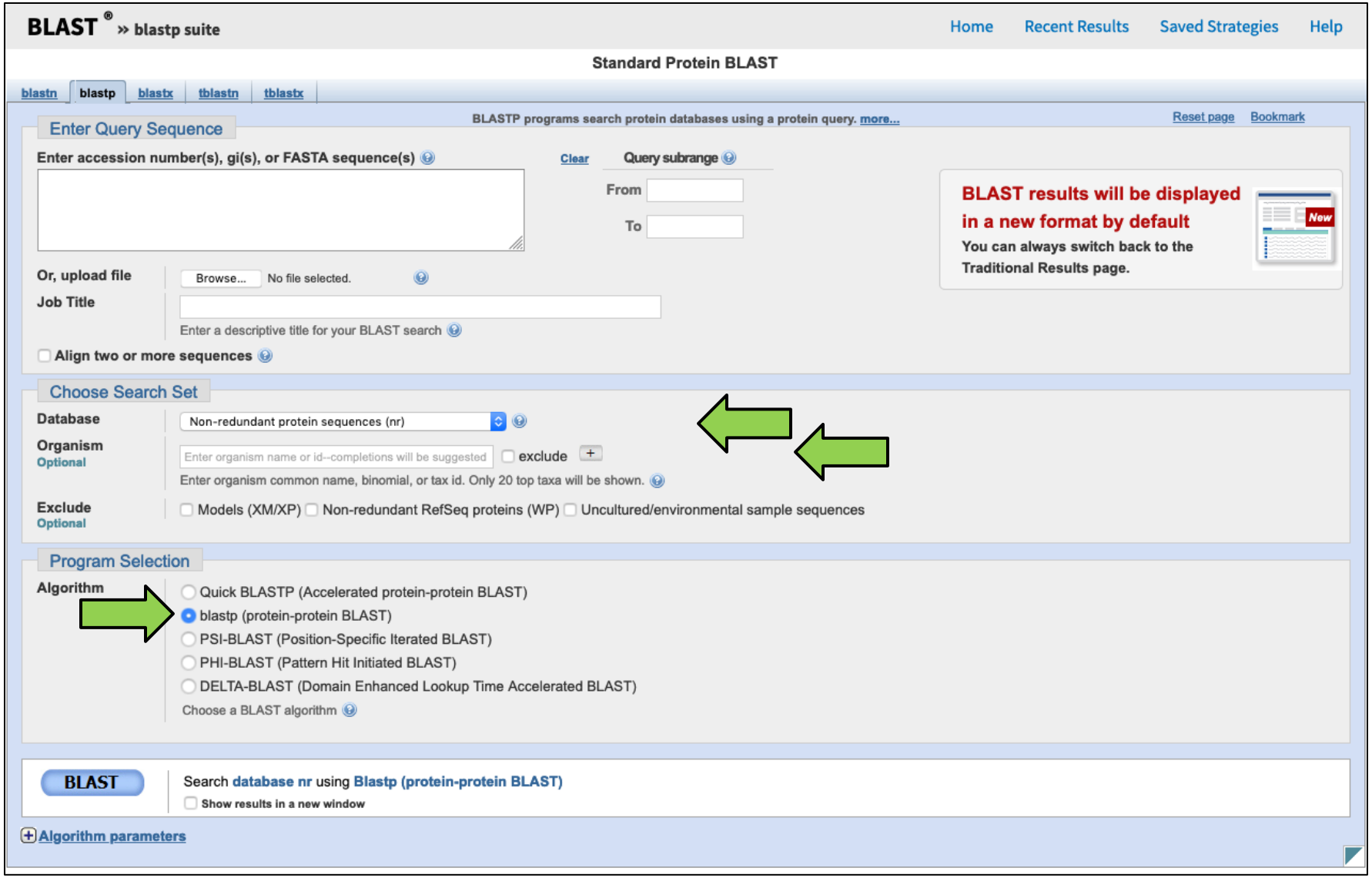

BLAST
Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
Show results in a new window
Algorithm parameters
Restore default search parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
11
Max matches in a query range
0
Scoring Parameters
Match/Mismatch Scores
2,-3
Gap Costs
Existence: 5 Extension: 2
Filters and Masking
Filter
Low complexity regions
Species-specific repeats for:
Homo sapiens (Human)
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
Show results in a new window

BLASTP - conceptual

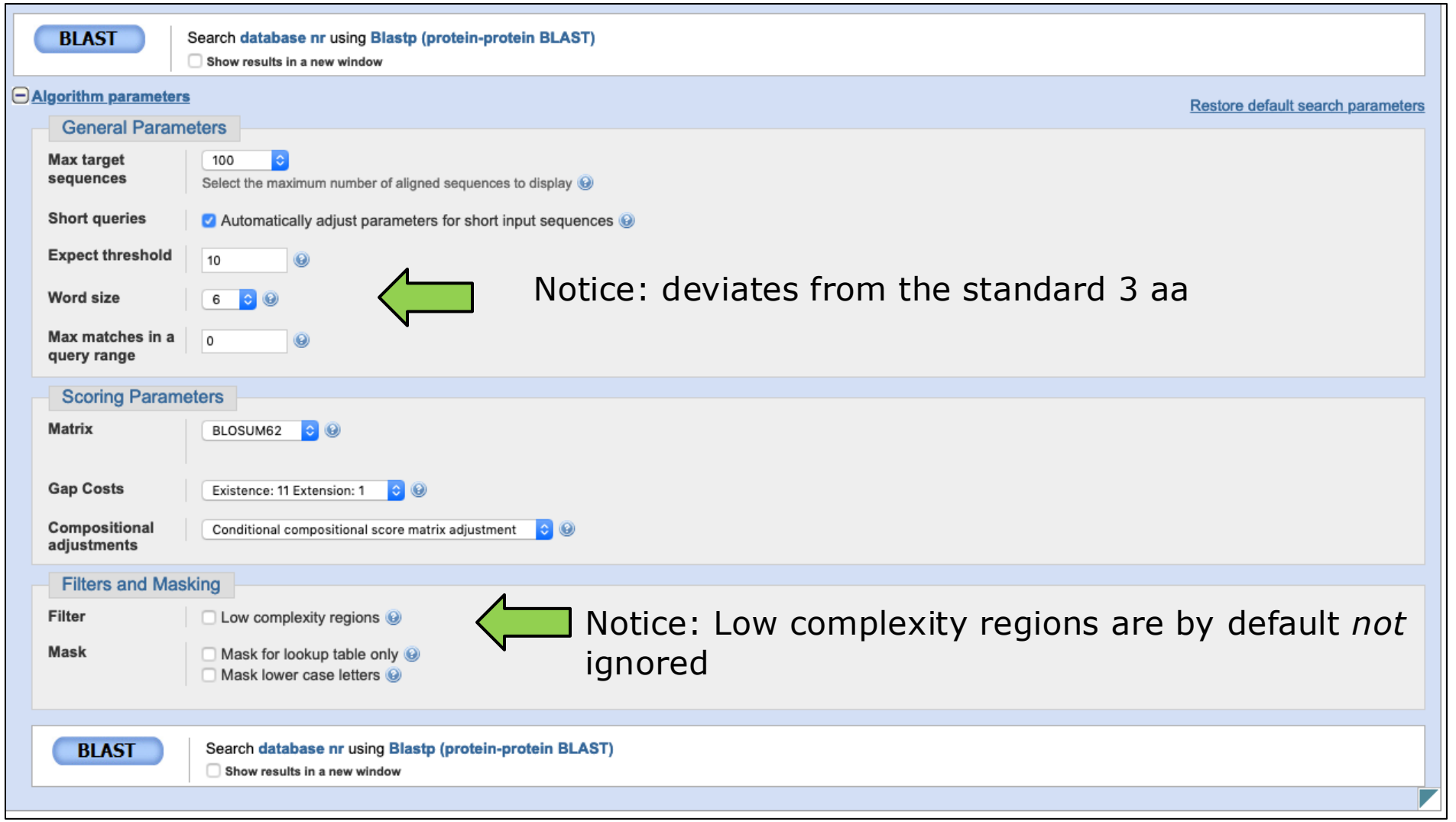
- Heuristics:
  - 2 x “Near match” within a window.
  - Default word length: 3 aa
    - (6 aa @ NCBI)
  - Default window length: 40 aa
- Alignment matrix:
  - PAM and BLOSUM-series (default: BLOSUM 62)
- Notice: These alignment matrices incorporate knowledge about protein evolution.



BLASTP@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
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
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)
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 sample sequences
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
Algorithm parameters

BLASTP@NCBI


BLAST
Search database nr using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Restore default search 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
6
Max matches in a query range
0
Notice: deviates from the standard 3 aa
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
Notice: Low complexity regions are by default not ignored
BLAST
Search database nr using Blastp (protein-protein BLAST)
Show results in a new window

BLASTP – alignment details

S8 family peptidase [Bacillus atrophaeus]

Sequence ID: WP_106045332.1 **Length:** 382 **Number of Matches:** 1

See 1 more title(s)

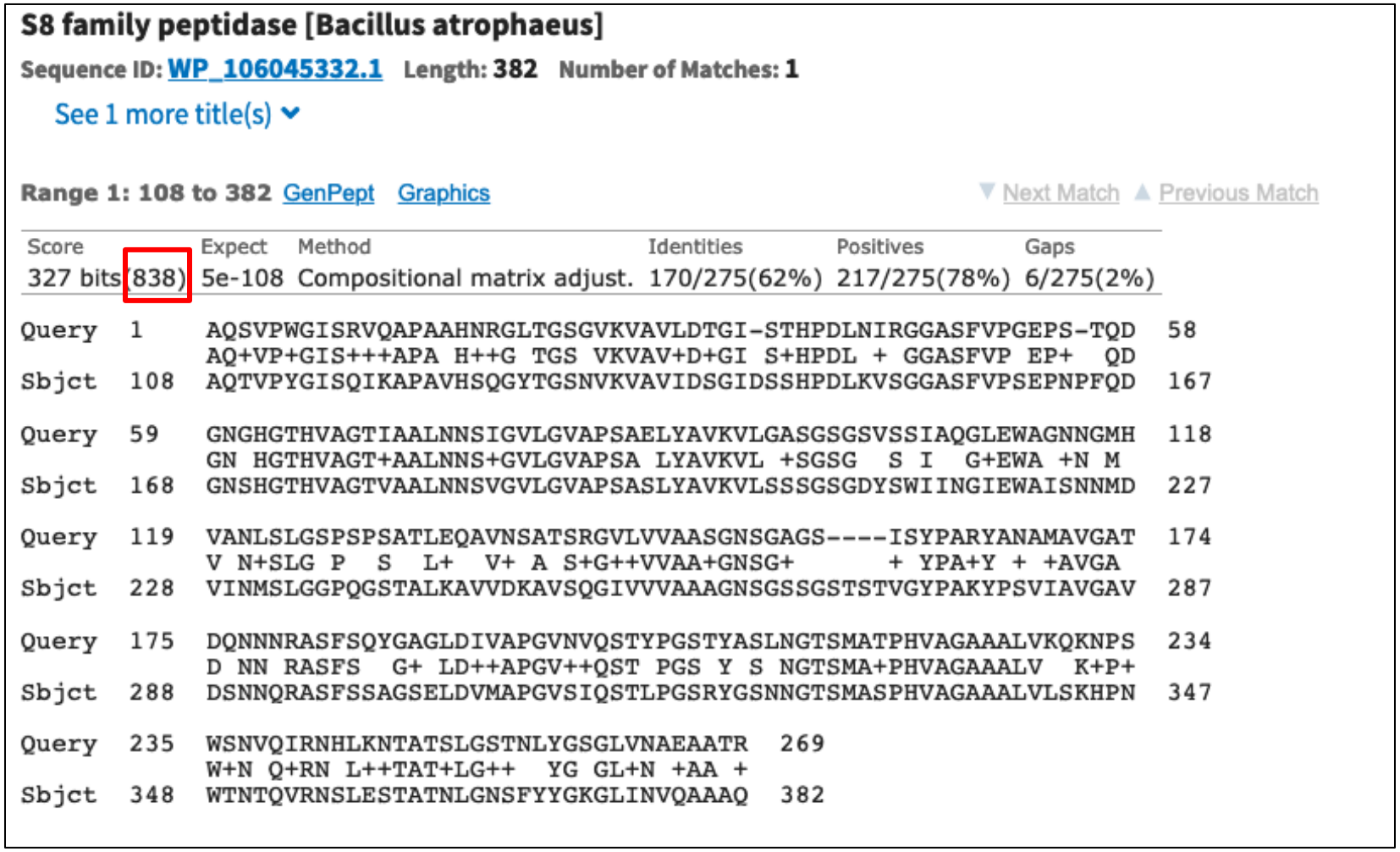
Range 1: 108 to 382 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
327 bits(838)	5e-108	Compositional matrix adjust.	170/275(62%)	217/275(78%)	6/275(2%)

```
Query  1    AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD  58
            AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+  QD
Sbjct  108  AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVSGGASFVPSEPNPFQD  167

Query  59   GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH  118
            GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG  S I  G+EWA +N M
Sbjct  168  GNSHGTHVAGTVAALNNSVGVLGVAPSASLYAVKVLSSSGSGDYSWIINGIEWAISNNMD  227

Query  119  VANLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGS----ISYPARYANAMAVGAT  174
            V N+SLG P  S  L+  V+ A S+G++VVAA+GNSG+      + YPA+Y + +AVGA
Sbjct  228  VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAAGNSGSSGSTSTVGYPAKYPSVIAVGAV  287

Query  175  DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS  234
            D NN RASFS  G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV  K+P+
Sbjct  288  DSNNQRASFSSAGSELDVMAPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN  347

Query  235  WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR  269
            W+N Q+RN L++TAT+LG++  YG GL+N +AA +
Sbjct  348  WTNTQVRNSLESTATNLGNSFYYGKGLINVQAAAQ  382
```

BLASTP – alignment details

Alignment score



S8 family peptidase [Bacillus atrophaeus]

Sequence ID: WP_106045332.1 **Length:** 382 **Number of Matches:** 1

See 1 more title(s)

Range 1: 108 to 382 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
327 bits(838)	5e-108	Compositional matrix adjust.	170/275(62%)	217/275(78%)	6/275(2%)

```
Query  1    AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD  58
            AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+  QD
Sbjct  108  AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVSGGASFVPSEPNPFQD  167

Query  59   GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH  118
            GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG  S I  G+EWA +N M
Sbjct  168  GNSHGTHVAGTVAALNNSVGVLGVAPSASLYAVKVLSSSGSGDYSWIINGIEWAISNNMD  227

Query  119  VANLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGS----ISYPARYANAMAVGAT  174
            V N+SLG P  S  L+  V+ A S+G++VVAA+GNSG+      + YPA+Y + +AVGA
Sbjct  228  VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAAGNSGSSGSTSTVGYPAKYPSVIAVGAV  287

Query  175  DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS  234
            D NN RASFS  G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV  K+P+
Sbjct  288  DSNNQRASFSSAGSELDVMAPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN  347

Query  235  WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR  269
            W+N Q+RN L++TAT+LG++  YG GL+N +AA +
Sbjct  348  WTNTQVRNSLESTATNLGNSFYYGKGLINVQAAAQ  382
```

BLASTP – alignment details

Bit-score (normalized alignment score)

S8 family peptidase [Bacillus atrophaeus]

Sequence ID: WP_106045332.1 **Length:** 382 **Number of Matches:** 1

See 1 more title(s)

Range 1: 108 to 382 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
327 bits(838)	5e-108	Compositional matrix adjust.	170/275(62%)	217/275(78%)	6/275(2%)

```
Query  1    AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD  58
            AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+  QD
Sbjct  108  AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVSGGASFVPSEPNPFQD  167

Query  59   GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH  118
            GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG  S I  G+EWA +N M
Sbjct  168  GNSHGTHVAGTVAALNNSVGVLGVAPSASLYAVKVLSSSGSGDYSWIINGIEWAISNNMD  227

Query  119  VANLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGS----ISYPARYANAMAVGAT  174
            V N+SLG P  S  L+  V+ A S+G++VVAA+GNSG+      + YPA+Y + +AVGA
Sbjct  228  VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAAGNSGSSGSTSTVGYPAKYPSVIAVGAV  287

Query  175  DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS  234
            D NN RASFS  G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV  K+P+
Sbjct  288  DSNNQRASFSSAGSELDVMAPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN  347

Query  235  WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR  269
            W+N Q+RN L++TAT+LG++  YG GL+N +AA +
Sbjct  348  WTNTQVRNSLESTATNLGNSFYYGKGLINVQAAAQ  382
```

BLASTP – alignment details

E-value

S8 family peptidase [Bacillus atrophaeus]

Sequence ID: WP_106045332.1 **Length:** 382 **Number of Matches:** 1

See 1 more title(s)

Range 1: 108 to 382 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
327 bits(838)	5e-108	Compositional matrix adjust.	170/275(62%)	217/275(78%)	6/275(2%)

```
Query  1    AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD  58
            AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+  QD
Sbjct  108  AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVSGGASFVPSEPNPFQD  167

Query  59   GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH  118
            GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG  S I  G+EWA +N M
Sbjct  168  GNSHGTHVAGTVAALNNSVGVLGVAPSASLYAVKVLSSSGSGDYSWIINGIEWAISNNMD  227

Query  119  VANLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGS----ISYPARYANAMAVGAT  174
            V N+SLG P  S  L+  V+ A S+G++VVAA+GNSG+      + YPA+Y + +AVGA
Sbjct  228  VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAAGNSGSSGSTSTVGYPAKYPSVIAVGAV  287

Query  175  DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS  234
            D NN RASFS  G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV  K+P+
Sbjct  288  DSNNQRASFSSAGSELDVMAPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN  347

Query  235  WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR  269
            W+N Q+RN L++TAT+LG++  YG GL+N +AA +
Sbjct  348  WTNTQVRNSLESTATNLGNSFYYGKGLINVQAAAQ  382
```

BLASTP – alignment details

Identities (perfect matches)

S8 family peptidase [Bacillus atrophaeus]

Sequence ID: WP_106045332.1 **Length:** 382 **Number of Matches:** 1

See 1 more title(s)

Range 1: 108 to 382 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
327 bits(838)	5e-108	Compositional matrix adjust.	170/275(62%)	217/275(78%)	6/275(2%)

```
Query  1    AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD  58
            AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+  QD
Sbjct  108  AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVSGGASFVPSEPNPFQD  167

Query  59   GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH  118
            GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG  S I  G+EWA +N M
Sbjct  168  GNSHGTHVAGTVAALNNSVGVLGVAPSASLYAVKVLSSSGSGDYSWIINGIEWAISNNMD  227

Query  119  VANLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGS----ISYPARYANAMAVGAT  174
            V N+SLG P  S  L+  V+ A S+G++VVAA+GNSG+      + YPA+Y + +AVGA
Sbjct  228  VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAAGNSGSSGSTSTVGYPAKYPSVIAVGAV  287

Query  175  DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS  234
            D NN RASFS  G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV  K+P+
Sbjct  288  DSNNQRASFSSAGSELDVMAPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN  347

Query  235  WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR  269
            W+N Q+RN L++TAT+LG++  YG GL+N +AA +
Sbjct  348  WTNTQVRNSLESTATNLGNSFYYGKGLINVQAAAQ  382
```

BLASTP – alignment details

Positives (AKA similarities – have a positive BLOSUM62 value)

S8 family peptidase [Bacillus atrophaeus]

Sequence ID: WP_106045332.1 **Length:** 382 **Number of Matches:** 1

See 1 more title(s)

Range 1: 108 to 382 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
327 bits(838)	5e-108	Compositional matrix adjust.	170/275(62%)	217/275(78%)	6/275(2%)

```
Query  1    AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD  58
            AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+  QD
Sbjct  108  AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVSGGASFVPSEPNPFQD  167

Query  59   GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH  118
            GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG  S I  G+EWA +N M
Sbjct  168  GNSHGTHVAGTVAALNNSVGVLGVAPSASLYAVKVLSSSGSGDYSWIINGIEWAISNNMD  227

Query  119  VANLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGS----ISYPARYANAMAVGAT  174
            V N+SLG P  S  L+  V+ A S+G++VVAA+GNSG+      + YPA+Y + +AVGA
Sbjct  228  VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAAGNSGSSGSTSTVGYPAKYPSVIAVGAV  287

Query  175  DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS  234
            D NN RASFS  G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV  K+P+
Sbjct  288  DSNNQRASFSSAGSELDVMAPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN  347

Query  235  WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR  269
            W+N Q+RN L++TAT+LG++  YG GL+N +AA +
Sbjct  348  WTNTQVRNSLESTATNLGNSFYYGKGLINVQAAAQ  382
```

BLASTP – alignment details

Gaps

S8 family peptidase [Bacillus atrophaeus]

Sequence ID: WP_106045332.1 **Length:** 382 **Number of Matches:** 1

See 1 more title(s)

Range 1: 108 to 382 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
327 bits(838)	5e-108	Compositional matrix adjust.	170/275(62%)	217/275(78%)	6/275(2%)

```
Query  1    AQSVPWGISRVQAPAAHNRGLTGSGVKVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD  58
            AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+  QD
Sbjct  108  AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVSGGASFVPSEPNPFQD  167

Query  59   GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH  118
            GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG  S I  G+EWA +N M
Sbjct  168  GNSHGTHVAGTVAALNNSVGVLGVAPSASLYAVKVLSSSGSGDYSWIINGIEWAISNNMD  227

Query  119  VANLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSGAGS----ISYPARYANAMAVGAT  174
            V N+SLG P  S  L+  V+ A S+G++VVAA+GNSG+      + YPA+Y + +AVGA
Sbjct  228  VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAAGNSGSSGSTSTVGYPAKYPSVIAVGAV  287

Query  175  DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS  234
            D NN RASFS  G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV  K+P+
Sbjct  288  DSNNQRASFSSAGSELDVMAPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN  347

Query  235  WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR  269
            W+N Q+RN L++TAT+LG++  YG GL+N +AA +
Sbjct  348  WTNTQVRNSLESTATNLGNSFYYGKGLINVQAAAQ  382
```

BLASTP - search details

- **Step 1:** (optional) Mask out “low-complexity” regions in the query sequence.

RecName: Full=Titin; AltName: Full=Connectin [Mus musculus]

Sequence ID: A2ASS6.1 **Length:** 35213 **Number of Matches:** 776

Range 1: 12871 to 35213 GenPept Graphics ▼ Next Match ▲ Previous Match

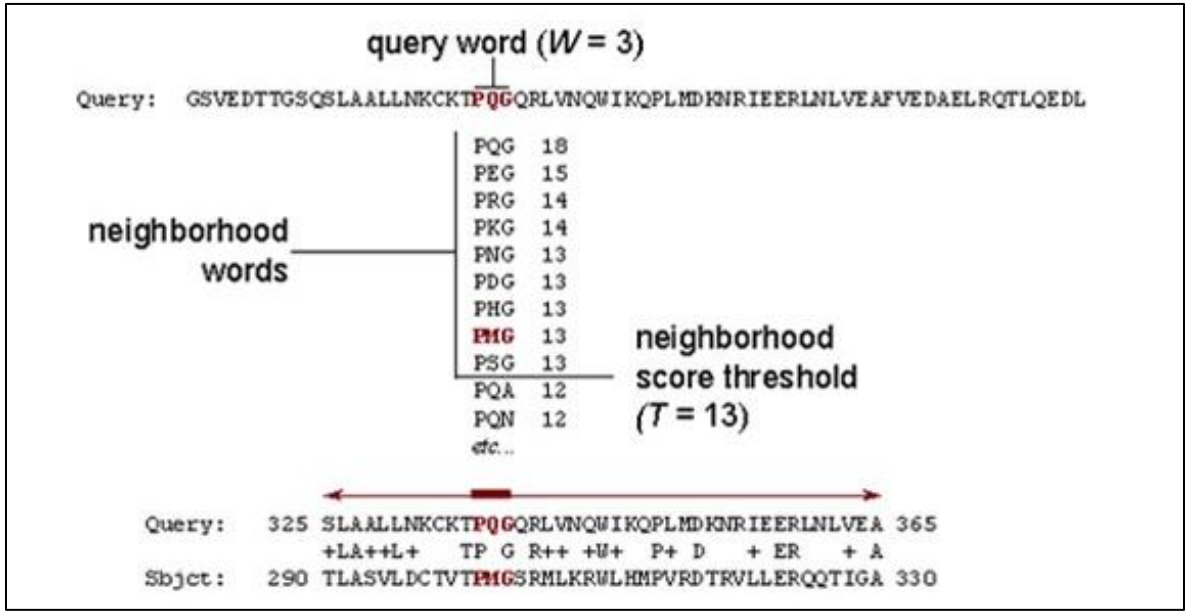
Score	Expect	Method	Identities	Positives	Gaps
42468 bits(110239)	0.0	Compositional matrix adjust.	21066/22346(94%)	21832/22346(97%)	7/22346(0%)

```
Query  12009  SPIEAERRKLRPGSGGEKPPDEAPFTYQLKAVPLKFVKEIKDIILTESEFVGSSAIFECL  12068
              SPIEAER+KLRPGSGGEKPPDEAPFTYQLKAVPLKFVKEIKDI+LTE+E VGSSAIFECL
Sbjct  12871  SPIEAERKKLRPGSGGEKPPDEAPFTYQLKAVPLKFVKEIKDIVLTEAESVGSSAIFECL  12930

Query  12069  VSPSTAITTWMKDGSNIRESPKHRFIADGKDRKLHIIDVQLSDAGEYTCVLRLGNKEKTS  12128
              VSPSTAITTWMKDGSNIRESPKHRFIADGKDRKLHIIDVQLSDAGEYTCVLRLGNKEKTS
Sbjct  12931  VSPSTAITTWMKDGSNIRESPKHRFIADGKDRKLHIIDVQLSDAGEYTCVLRLGNKEKTS  12990

Query  12129  TAKLVVEELPVRFvktleeevtvvkGQPLYLSCELNKERDVVWRKDGKIVVEKPGRIVPG  12188
              TAKL+VEELPVRFVKTLEEEVTVVKGQPLYLSCELNKERDVVWRKDGKIVVEKPGRIVPG
Sbjct  12991  TAKLIVEELPVRFVKTLEEEVTVVKGQPLYLSCELNKERDVVWRKDGKIVVEKPGRIVPG  13050

Query  12189  VIGLMRALTINDADDTDAGTYTVTVENANNLECSSCVKVVEVIRDWLVKPIRDQHVKPKG  12248
              VIGLMRALTINDADDTDAGTYTVTVENANNLECSSCVKVVE+IR+WLVKPIRDQHVKPKG
Sbjct  13051  VIGLMRALTINDADDTDAGTYTVTVENANNLECSSCVKVVEIIREWLVKPIRDQHVKPKG  13110

Query  12249  TAIFACDIAKDTPNIKWFKGYDEIPAEPNDKTEILRDGNHLYLKIKNAMPEDIAEYAVEI  12308
              TA+F CDIAKDTPNIKWFKGYDEIP EPNDKTEIL++GNHL+LK+KNAMPEDI EYAVEI
Sbjct  13111  TAVFTCDIAKDTPNIKWFKGYDEIPLEPNDKTEILKEGNHLFLKVKNAMPEDIDEYAVEI  13170
```

BLASTP - search details

- **Step 2:** Finding ”High-scoring segments”

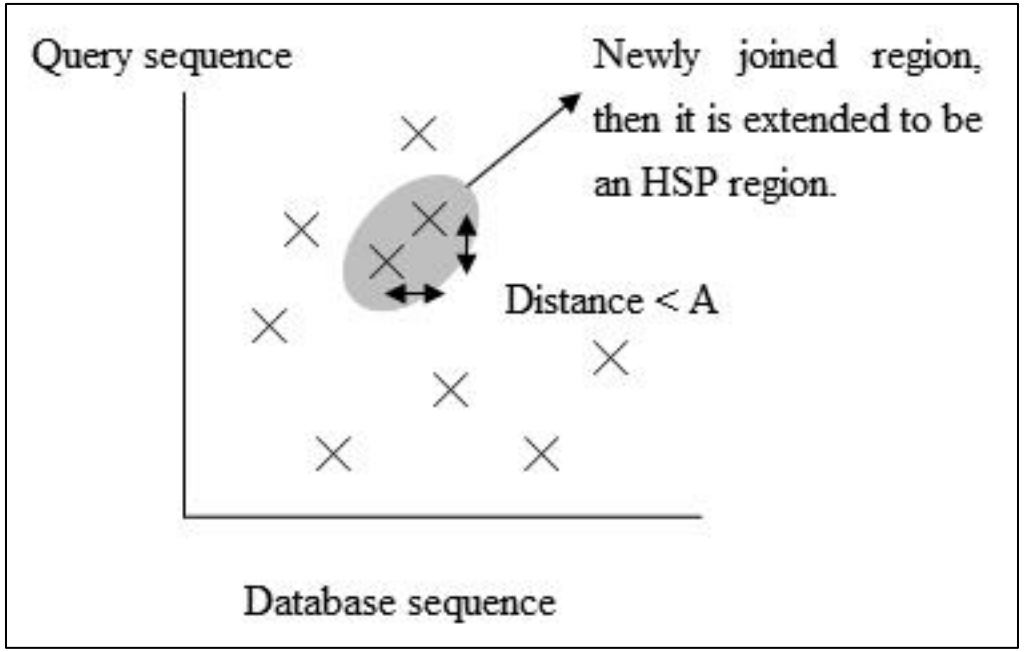
- For each *k*-letter word (default 3 aa) in the query sequence, search for all permutations with an alignment score above a certain threshold.



- …. The alignment is then extended to the sides, as long as the score is positive.
- (We are glossing over some finer details of algorithm optimization here)

BLASTP – search details

- **Step 3:** Evaluating “High-scoring segment pairs” (HSPs)



- … the alignment is extended as part of the process, and the algorithm saves time by not having to perform a full local alignment on both sequences again.

BLASTP – search details

- **Step 4:** Statistical assessment: **E-values**

- The E-value is calculated based on the alignment score (S) and the following parameters describing the search space:

m	length of query sequence
n	length of ALL sequences in the database
K, λ	statistical parameters (tailored to database type/search strategy)

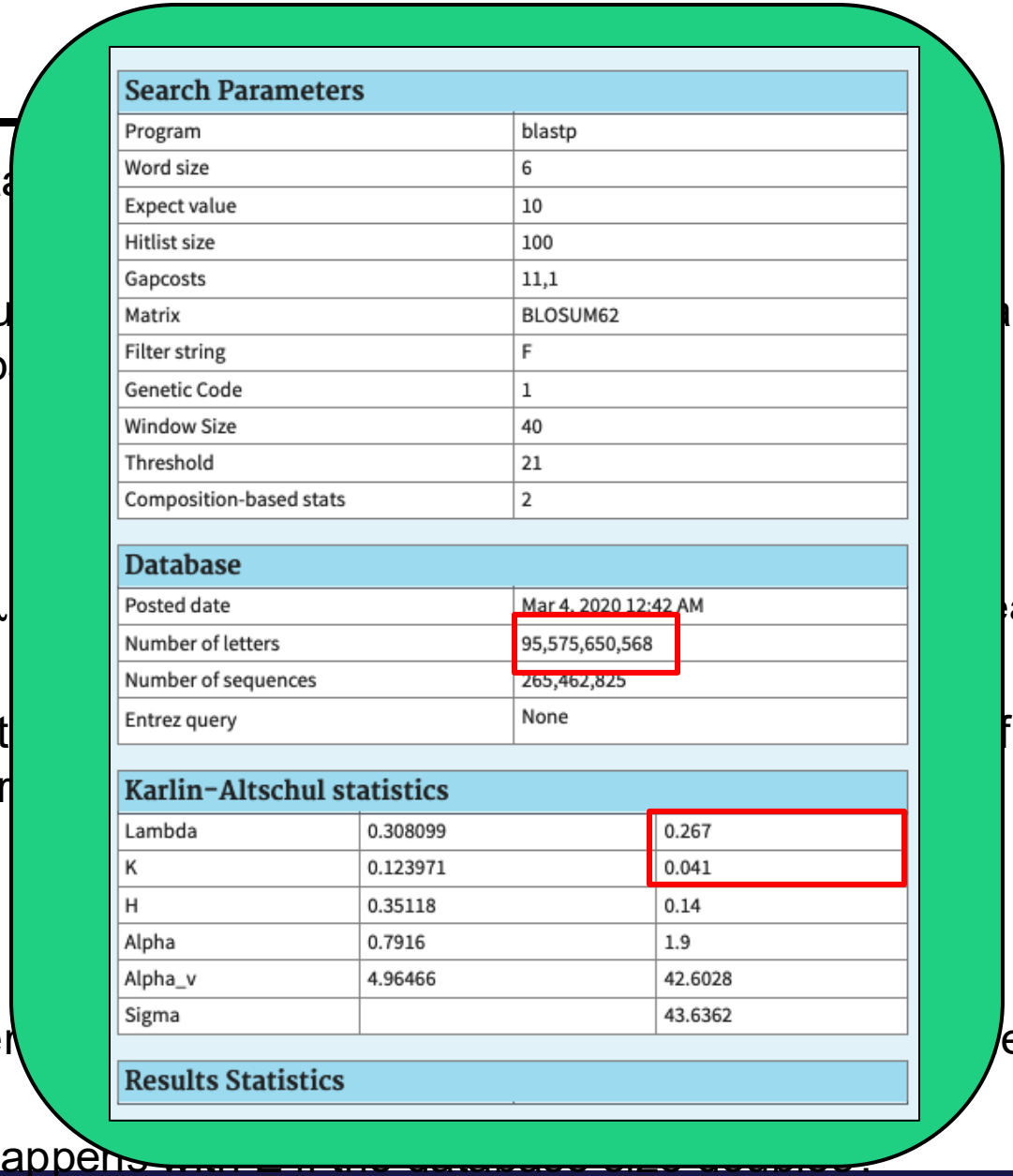
- From this the E-value can be calculated (expected number of alignments with a score equal or greater than S):

$$E = K\,mn\,e^{-\lambda S}$$

- Remember that the E-value is a statistical interpretation of the alignment score (S).
- **Q:** What happens with E if the database size doubles?

BLASTP –

- **Step 4:** Sta
- The E-valu nd the following p
  - m
  - n
  - K, λ earch strategy)
- From this t f alignments with a scor
- Remember e alignment score (S).
- **Q:** What happens



Search Parameters	
Program	blastp
Word size	6
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

Database	
Posted date	Mar 4, 2020 12:42 AM
Number of letters	95,575,650,568
Number of sequences	265,462,825
Entrez query	None

Karlin-Altschul statistics		
Lambda	0.308099	0.267
K	0.123971	0.041
H	0.35118	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Results Statistics

BLASTP – search details

- **Step 4:** Statistical assessment: ***p*-values**

- A more traditional statistical measure would be a *p*-value. The *p*-value of the significances of an alignment, can be derived from the E-value:

$$p = 1 - e^{-E}$$

- Here *p* will be the probability observing **zero** random alignments with score >= S.

- The table to the left shows the E →*p* relationship for some example E-values.

- Observations:
- An E-value > 3 is close to a *p*-value of 1.0
- As the E-values get lower, the E and *p*-value will get close to each other.

E-value	*p* -value
250	1
10	0.9999546
3	0.95021293
1	0.63212056
0.05	0.04877058
1.0E-03	0.0009995
1.0E-05	1.0E-05
1.0E-10	1.0E-10
1.0E-20	1.0E-20
1.0E-50	1.0E-50

BLASTP – search details

- **Step 4:** Statistical assessment: **bit-scores**

- The alignment score (S) can be evaluated in the context of the data-base by converting it to a **bit-score**:

$$S' = \frac{\lambda S - ln(K)}{ln(2)}$$

- The unit is in bits (base 2), and can be interpreted as a an estimate of the effective search space needed to generate S' by random. For example:

$$S' = 30 \rightarrow search\ space = 2^{30} \cong 1.07\ billion$$

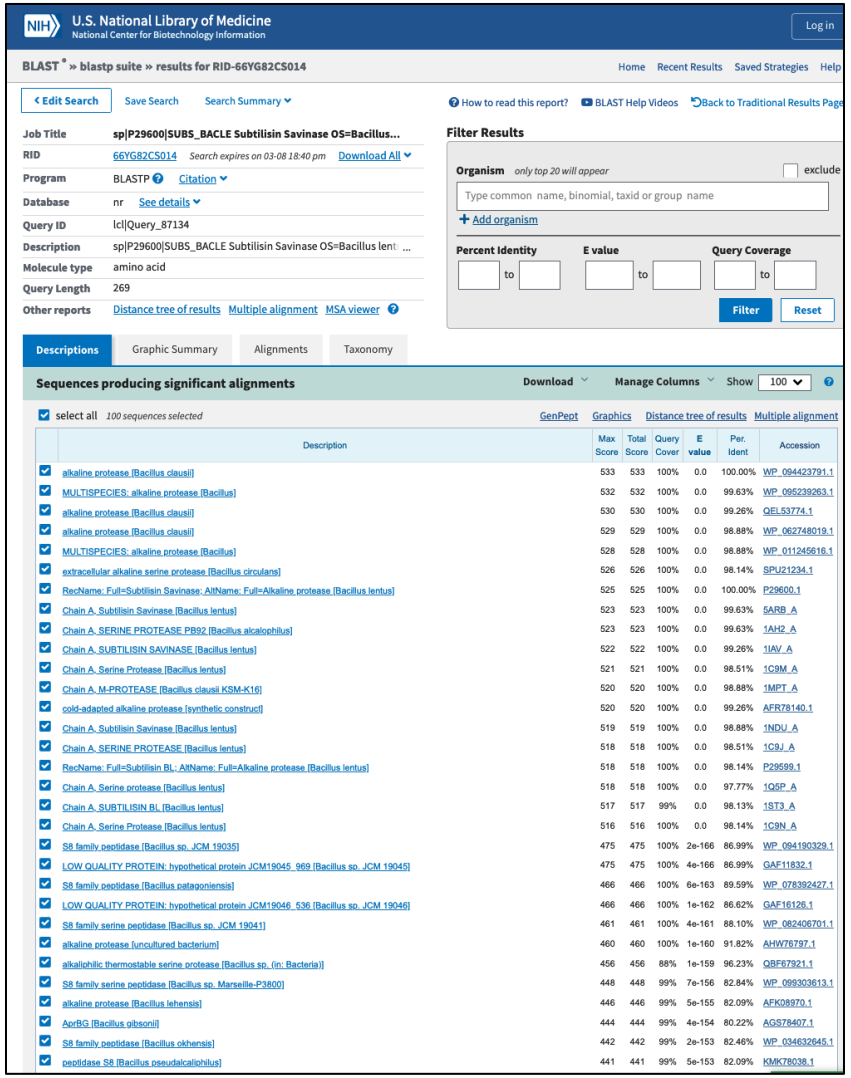
BLASTP – search details

- **Step 5:** Reporting

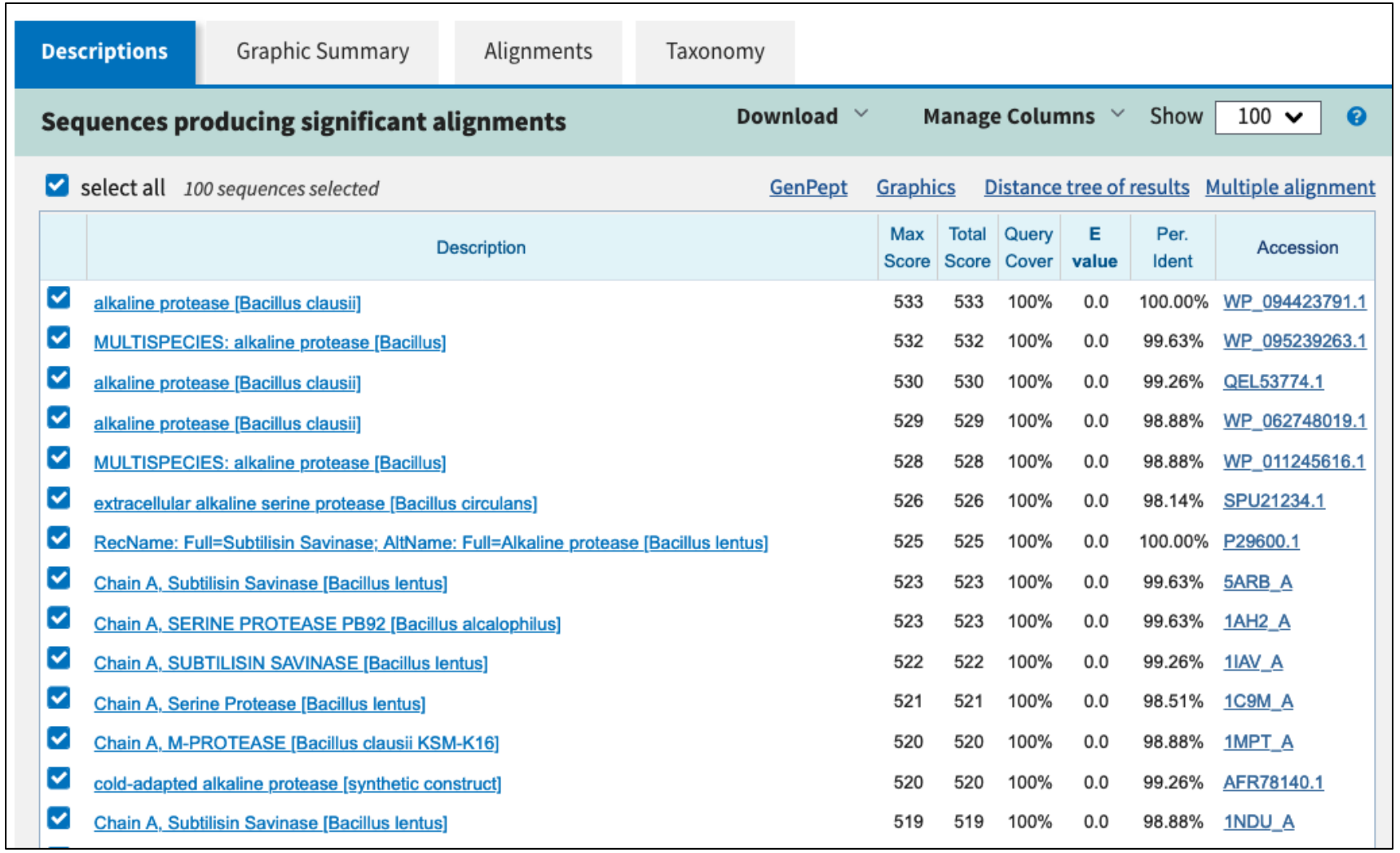
- Only hits with an E-value < 10 (NCBI default) are kept and the top 100 (NCBI default) is reported.

- The low-level output is a `text-based` report.

- The NCBI web-page wraps all information in a much more interactive manner, and links in additional data resources.
  (See next slides)



NCBI BLASTP report



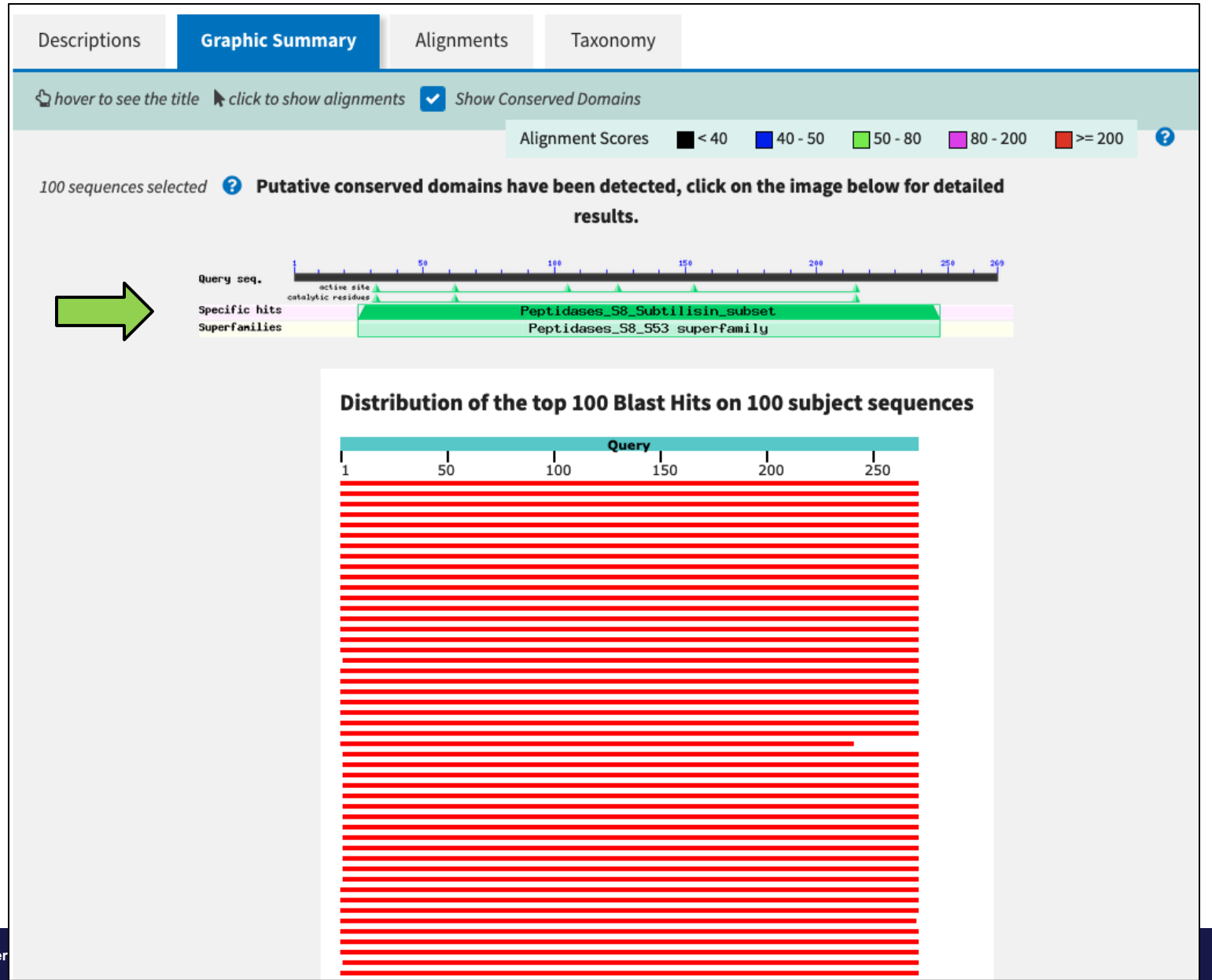
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

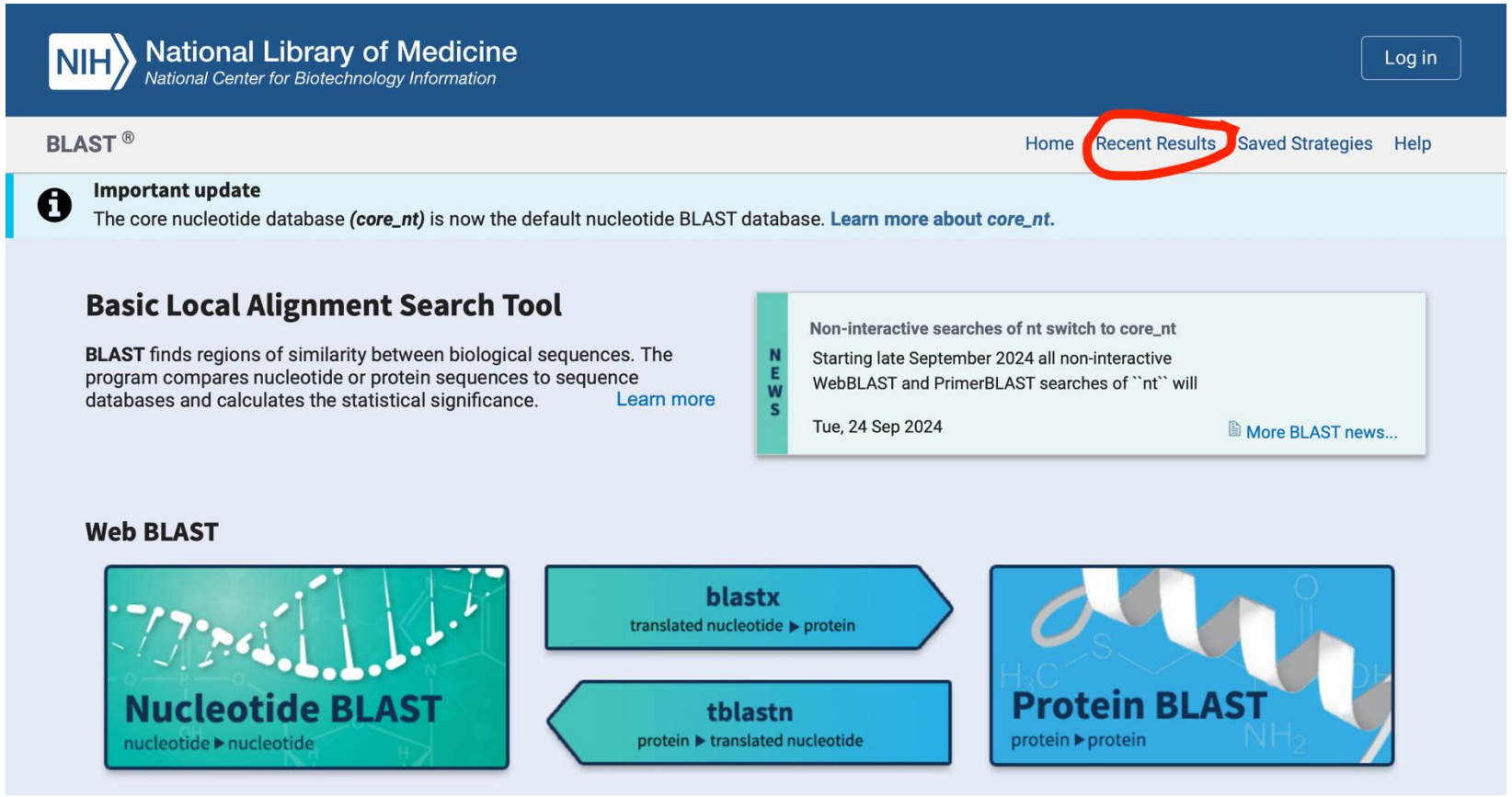
Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
alkaline protease [Bacillus clausii]	533	533	100%	0.0	100.00%	WP_094423791.1
MULTISPECIES: alkaline protease [Bacillus]	532	532	100%	0.0	99.63%	WP_095239263.1
alkaline protease [Bacillus clausii]	530	530	100%	0.0	99.26%	QEL53774.1
alkaline protease [Bacillus clausii]	529	529	100%	0.0	98.88%	WP_062748019.1
MULTISPECIES: alkaline protease [Bacillus]	528	528	100%	0.0	98.88%	WP_011245616.1
extracellular alkaline serine protease [Bacillus circulans]	526	526	100%	0.0	98.14%	SPU21234.1
RecName: Full=Subtilisin Savinase; AltName: Full=Alkaline protease [Bacillus lentus]	525	525	100%	0.0	100.00%	P29600.1
Chain A, Subtilisin Savinase [Bacillus lentus]	523	523	100%	0.0	99.63%	5ARB_A
Chain A, SERINE PROTEASE PB92 [Bacillus alcalophilus]	523	523	100%	0.0	99.63%	1AH2_A
Chain A, SUBTILISIN SAVINASE [Bacillus lentus]	522	522	100%	0.0	99.26%	1IAV_A
Chain A, Serine Protease [Bacillus lentus]	521	521	100%	0.0	98.51%	1C9M_A
Chain A, M-PROTEASE [Bacillus clausii KSM-K16]	520	520	100%	0.0	98.88%	1MPT_A
cold-adapted alkaline protease [synthetic construct]	520	520	100%	0.0	99.26%	AFR78140.1
Chain A, Subtilisin Savinase [Bacillus lentus]	519	519	100%	0.0	98.88%	1NDU_A

NCBI BLASTP report



How to work on BLAST@NCBI with previous searches

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



How to work on BLAST@NCBI with previous searches