

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 [Cite](#) Cited by 120750 [Related articles](#) All 56 versions Web of Science: 77938 [↔](#)

Highly accurate protein structure prediction with **AlphaFold**

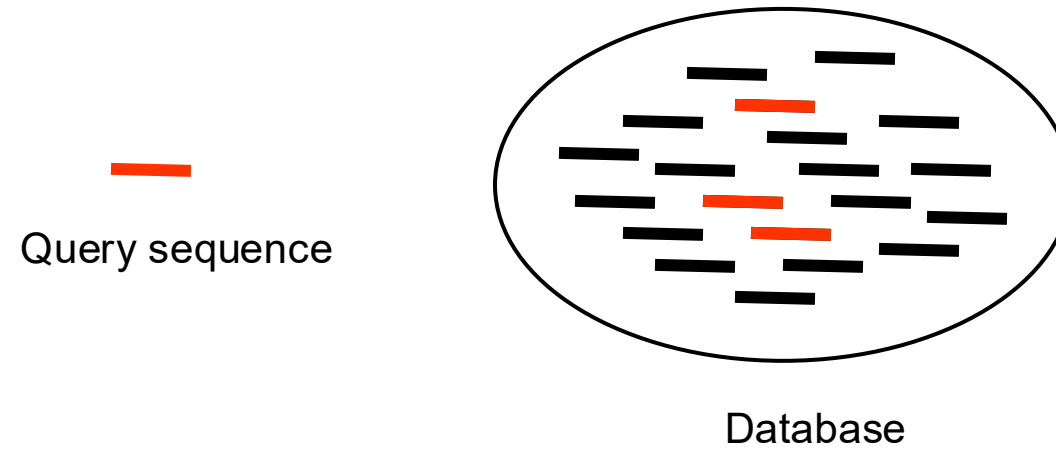
[J Jumper](#), [R Evans](#), [A Pritzel](#), [T Green](#), [M Figurnov](#)... - nature, 2021 - nature.com

... 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 [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

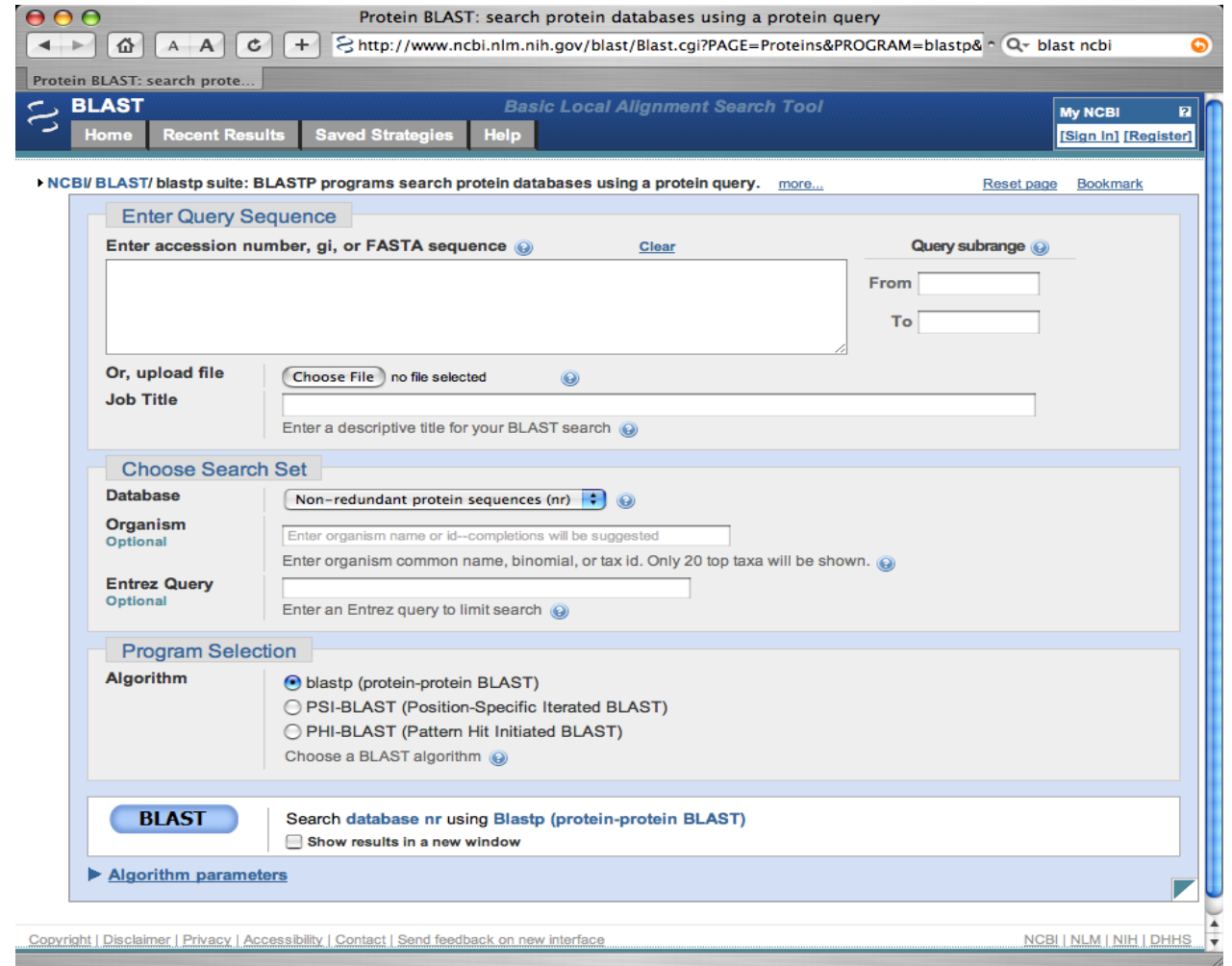
Searching on the web: BLAST at NCBI

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



The screenshot shows the NCBI Protein BLAST web interface. The browser address bar displays the URL: `http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&`. The page title is "Protein BLAST: search protein databases using a protein query". The interface includes a navigation bar with "Home", "Recent Results", "Saved Strategies", and "Help" buttons, along with a "My NCBI" link and "Sign In" and "Register" buttons. The main content area is titled "NCBI/BLAST/blastp suite: BLASTP programs search protein databases using a protein query." and includes a "Reset page" and "Bookmark" link. The interface is divided into several sections: "Enter Query Sequence" with a text input field for "Enter accession number, gi, or FASTA sequence", a "Clear" button, and a "Query subrange" section with "From" and "To" input fields; "Or, upload file" with a "Choose File" button and a "Job Title" input field; "Choose Search Set" with a "Database" dropdown menu set to "Non-redundant protein sequences (nr)", an "Organism" input field, and an "Entrez Query" input field; "Program Selection" with radio buttons for "blastp (protein-protein BLAST)", "PSI-BLAST (Position-Specific Iterated BLAST)", and "PHI-BLAST (Pattern Hit Initiated BLAST)"; and a "BLAST" button with a "Show results in a new window" checkbox. The footer contains a copyright notice and links for "Disclaimer", "Privacy", "Accessibility", "Contact", and "Send feedback on new interface", along with the text "NCBI | NLM | NIH | DHHS".

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**?

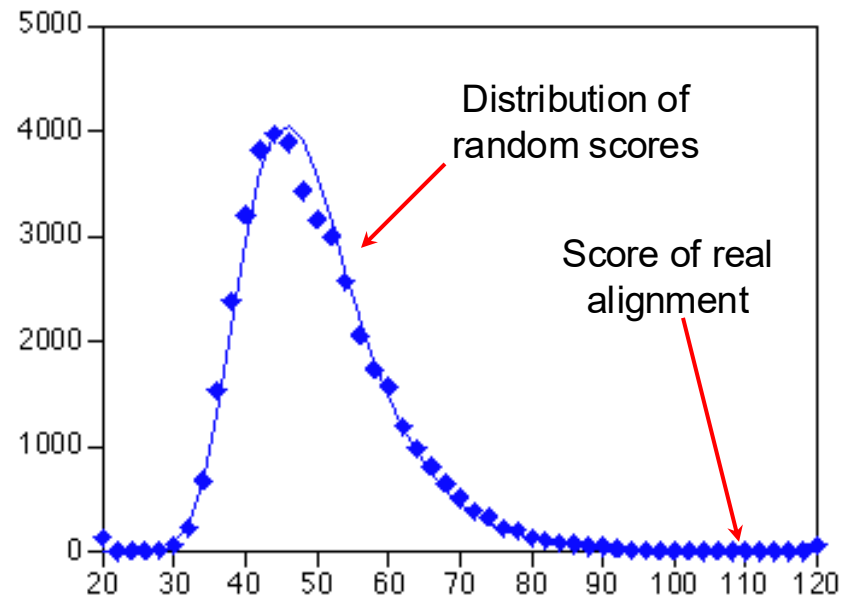
Distribution of random alignment scores

https://colab.research.google.com/drive/1rrzB1MLjk_IWaAuh-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.

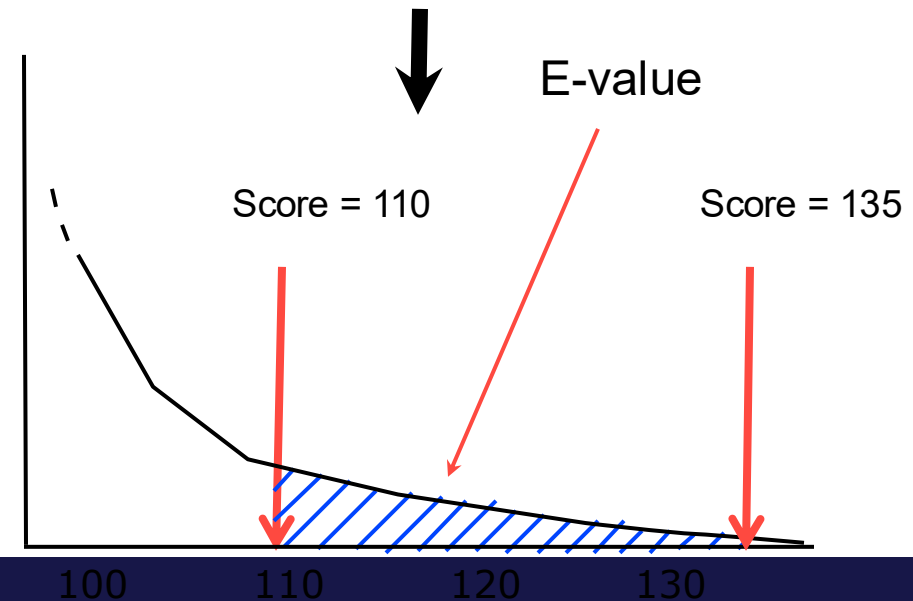
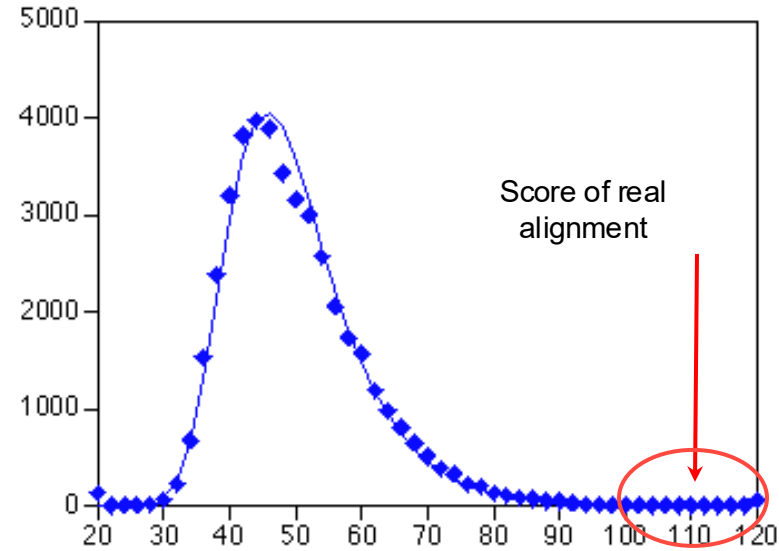
Example:

Alignment score = 110

E-value = 8.7

Alignment score = 135

E-value = 0.0001



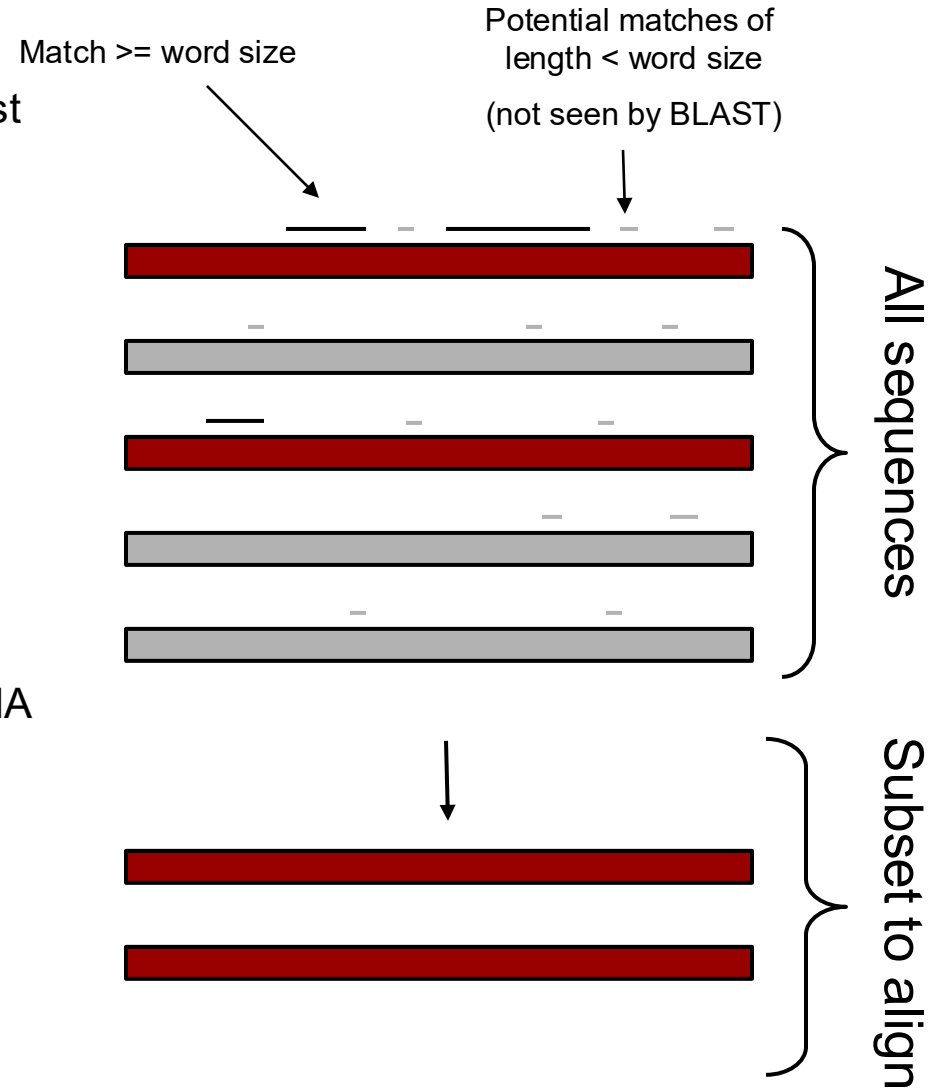
BLAST heuristics

- 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

Or, upload file 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 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 [?](#)

Search **database Nucleotide collection (nr/nt)** using **Blastn (Optimize for somewhat similar sequences)**

Show results in a new window

[+ Algorithm parameters](#)

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Notice: BLASTN is NOT default!

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

Short queries


Expect threshold

Word size

Max matches in a query range

Select the maximum number of aligned sequences to display


Automatically adjust parameters for short input sequences



Scoring Parameters

Match/Mismatch Scores

Gap Costs



Filters and Masking

Filter

Mask

Low complexity regions

Species-specific repeats for:

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

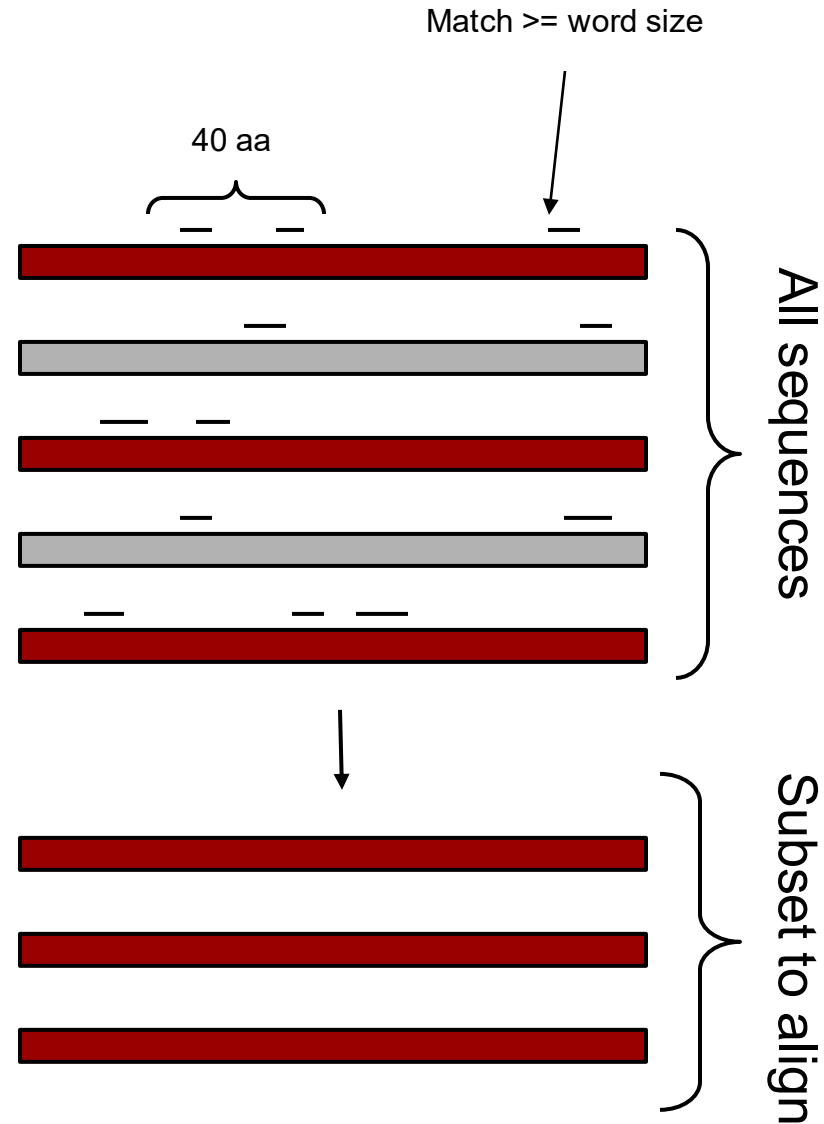
Technical University of Denmark

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](#)

[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

Or, upload file No file selected.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database exclude

Organism Optional 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

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

Show results in a new window

[+ Algorithm parameters](#)

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

BLASTP@NCBI

BLAST

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

 Show results in a new window

Algorithm parameters

General Parameters

[Restore default search parameters](#)

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

Short queries Automatically adjust parameters for short input sequences

Expect threshold ← Notice: deviates from the standard 3 aa

Word size ←

Max matches in a query range

Scoring Parameters

Matrix

Gap Costs

Compositional adjustments

Filters and Masking

Filter Low complexity regions ← Notice: Low complexity regions are by default *not* ignored

Mask Mask for lookup table only

Mask lower case letters

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
Sbjct 108		AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+ QD			167
Query 59		GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH			118
Sbjct 168		GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG S I G+EWA +N M			227
Query 119		VANLSLGSPPSATLEQAVNSATSRGVLVVAASGNSGAGS-----ISYPARYANAMAVGAT			174
Sbjct 228		V N+SLG P S L+ V+ A S+G++VVAA+GNSG+ + YPA+Y + +AVGA			287
Query 175		DQNNNRASFSSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS			234
Sbjct 288		D NN RASF S G+ LD++APGV++QST PGS Y S NGTSM A+PHVAGAAALV K+P+			347
Query 235		WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAAATR		269	
Sbjct 348		W+N Q+RN L+++TAT+LG++ YG GL+N +AA +		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	AQSVPWGISRVQAPAAHNRGLTGSQVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD				58
Sbjct 108	AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+ QD				167
Query 59	GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH				118
Sbjct 168	GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG S I G+EWA +N M				227
Query 119	VANLSLGSPPSATLEQAVNSATSRGVLVVAASGNSGAGS-----ISYPARYANAMAVGAT				174
Sbjct 228	V N+SLG P S L+ V+ A S+G++VVAA+GNSG+ + YPA+Y + +AVGA				287
Query 175	DQNNNRASFSSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS				234
Sbjct 288	D NN RASFSS G+ LD++APGV++QST PGS Y S NGTSM+PHVAGAAALV K+P+				347
Query 235	WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR				269
Sbjct 348	W+N Q+RN L++TAT+LG++ YG GL+N +AA +				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	AQSVPWGISRQAPAAHNRLTGSQVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD				58	
	AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+ QD					
Sbjct 108	AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVS GGASFVPSEPNPFQD				167	
Query 59	GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH				118	
	GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG S I G+EWA +N M					
Sbjct 168	GNSHGTHVAGTVAALNNSVGLGVAPSAELYAVKVLSSSGSGDYSWIINGIEWAISNNMD				227	
Query 119	VANLSLGSPPSATLEQAVNSATSRGVLVVAASGNSGAGS-----ISYPARYANAMAVGAT				174	
	V N+SLG P S L+ V+ A S+G++VVAA+GNSG+ + YPA+Y + +AVGA					
Sbjct 228	VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAGNSGSSGSTSTVGYPAKYPSVIAVGAV				287	
Query 175	DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS				234	
	D NN RASFS G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV K+P+					
Sbjct 288	DSNNQRASFSSAGSELDVMPAGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN				347	
Query 235	WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR		269			
	W+N Q+RN L++TAT+LG++ YG GL+N +AA +					
Sbjct 348	WTNTQVRNSLESTATNLGNSFYGKGLINVQAAAQ		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	AQSVPWGISRVQAPAAHNRLTGSVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD				58
	AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+ QD				
Sbjct 108	AQTVPYGISQIKAPAVHSQGYTGSNVKVAVIDSGIDSSHPDLKVS GGASFVPSEPDPFQD				167
Query 59	GNGHGHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH				118
	GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG S I G+EWA +N M				
Sbjct 168	GNSHGTHVAGTVAALNNSVGLGVAPSAELYAVKVLSSSGSGDYSWIINGIEWAISNNMD				227
Query 119	VANLSLGSPPSATLEQAVNSATSRGVLVVAASGNSGAGS-----ISYPARYANAMAVGAT				174
	V N+SLG P S L+ V+ A S+G++VVAA+GNSG+ + YPA+Y + +AVGA				
Sbjct 228	VINMSLGGPQGSTALKAVVDKAVSQGIVVVAAGNSGSSGSTSTVGYPAKYPSVIAVGAV				287
Query 175	DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGTYASLNGTSMATPHVAGAAALVKQKNPS				234
	D NN RASFS G+ LD++APGV++QST PGS Y S NGTSMA+PHVAGAAALV K+P+				
Sbjct 288	DSNNQRASFSSAGSELDVMPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN				347
Query 235	WSNVQIRNHLKNTATSLGSTNLYGSLVNAEAATR				269
	W+N Q+RN L++TAT+LG++ YG GL+N +AA +				
Sbjct 348	WTNTQVRNSLESTATNLGNSFYFGKGLINVQAAAQ				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		AQSVPWGISRVQAPAAHNRLTGSQVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD			58
Sbjct 108		AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+ QD			167
Query 59		GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH			118
Sbjct 168		GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG S I G+EWA +N M			227
Query 119		VANLSLGSPPSATLEQAVNSATSRGVLVVAASGNSGAGS-----ISYPARYANAMAVGAT			174
Sbjct 228		V N+SLG P S L+ V+ A S+G++VVAA+GNSG+ + YPA+Y + +AVGA			287
Query 175		DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS			234
Sbjct 288		D SNNQRASFSSAGSELDVMPGVSIQSTLPGSRYGSNNGTSMASPHVAGAAALVLSKHPN			347
Query 235		WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR		269	
Sbjct 348		W+N Q+RN L+++TAT+LG++ YG GL+N +AA +		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		AQSVPWGISRVQAPAAHNRLTGSVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD			58
Sbjct 108		AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+ QD			167
Query 59		GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH			118
Sbjct 168		GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG S I G+EWA +N M			227
Query 119		VANLSLGSPPSATLEQAVNSATSRGVLVVAASGNSGAGS-----ISYPARYANAMAVGAT			174
Sbjct 228		V N+SLG P S L+ V+ A S+G++VVAA+GNSG+ + YPA+Y + +AVGA			287
Query 175		DQNNNRASFSSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS			234
Sbjct 288		D NN RASF S G+ LD++APGV++QST PGS Y S NGTSM+PHVAGAAALV K+P+			347
Query 235		WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR	269		
Sbjct 348		W+N Q+RN L+++TAT+LG++ YG GL+N +AA +		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		AQSVPWGISRVQAPAAHNRLTGSVAVLDTGI-STHPDLNIRGGASFVPGEPS-TQD			58
Sbjct 108		AQ+VP+GIS+++APA H++G TGS VKVAV+D+GI S+HPDL + GGASFVP EP+ QD			167
Query 59		GNGHGTHVAGTIAALNNSIGVLGVAPSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMH			118
Sbjct 168		GN HGTHVAGT+AALNNS+GVLGVAPSA LYAVKVL +SGSG S I G+EWA +N M			227
Query 119		VANLSLGSPPSATLEQAVNSATSRGVLVVAASGNSGAGS-----ISYPARYANAMAVGAT			174
Sbjct 228		V N+SLG P S L+ V+ A S+G++VVAA+GNSG+ + YPA+Y + +AVGA			287
Query 175		DQNNNRASFSQYGAGLDIVAPGVNVQSTYPGSTYASLNGTSMATPHVAGAAALVKQKNPS			234
Sbjct 288		D NN RASFS G+ LD++APGV++QST PGS Y S NGTSM+PHVAGAAALV K+P+			347
Query 235		WSNVQIRNHLKNTATSLGSTNLYGSGLVNAEAATR		269	
Sbjct 348		W+N Q+RN L++TAT+LG++ YG GL+N +AA +		382	

BLASTP - search details

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

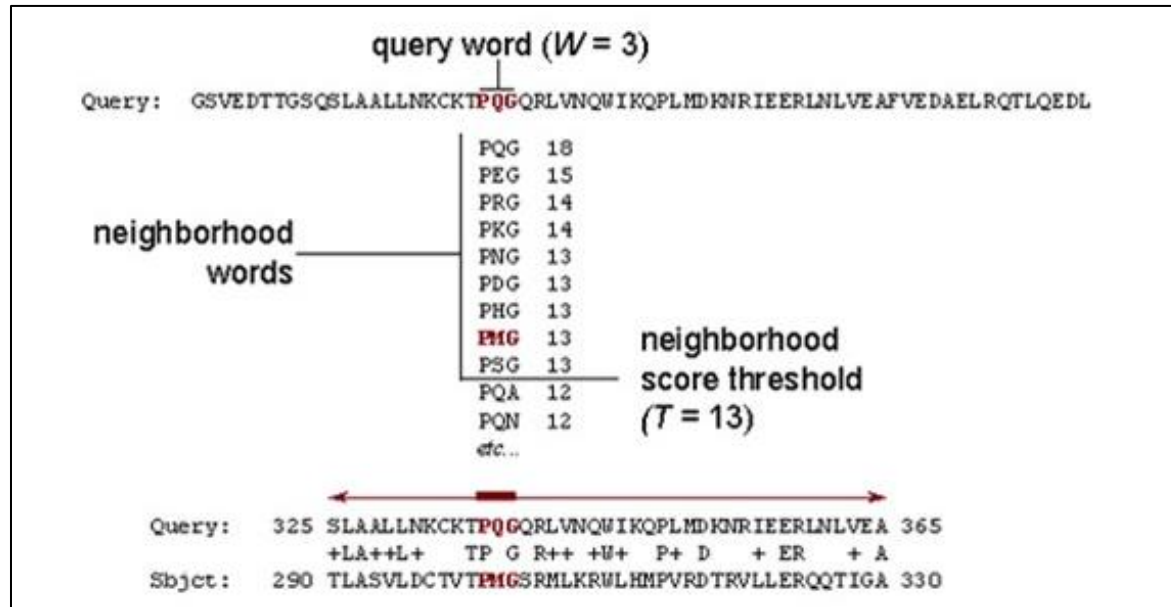
RecName: Full=Titin; AltName: Full=Connectin [Mus musculus]
Sequence ID: [A2ASS6.1](#) Length: 35213 Number of Matches: 776

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		SPIEAERRKLRPGSGGEEKPPDEAPFTYQLKAVPLKFVKEIKDIILTESEFVGSSAIFECL			12068
Sbjct 12871		SPIEAER+KLRPGSGGEEKPPDEAPFTYQLKAVPLKFVKEIKDI+LTE+E VGSSAIFECL			12930
Query 12069		VSPSTAITTWMKDGSNIRESPKHRFIADGKDRKLHIIDVQLSDAGEYTCVLRRLGNKEKTS			12128
Sbjct 12931		VSPSTAITTWMKDGSNIRESPKHRFIADGKDRKLHIIDVQLSDAGEYTCVLRRLGNKEKTS			12990
Query 12129		TAKLVEELPVRFvktleeevtvkvGQPLYLSCELNKERDVVWRKDGKIVVEKPGRIVPG			12188
Sbjct 12991		TAKLVEELPVRFVKTLEEEVTVVKGOPLYLSCELNKERDVVWRKDGKIVVEKPGRIVPG			13050
Query 12189		VIGLMRALTINDADDTAGTYTVTVENANNLECSSCVKVEVIRDWLVKPIRDQHVKPKG			12248
Sbjct 13051		VIGLMRALTINDADDTAGTYTVTVENANNLECSSCVKVEIIREWLVKPIRDQHVKPKG			13110
Query 12249		TAIFACDIAKDTPNIKWFKGYDEIPAEPNDKTEILRDGNHLYLKIKNAMPEDIAEYAVEI			12308
Sbjct 13111		TA+FCDIAKDTPNIKWFKGYDEIPEPNDKTEILKEGNHLFLKVKNAMPEDIDEYAVEI			13170

BLASTP - search details

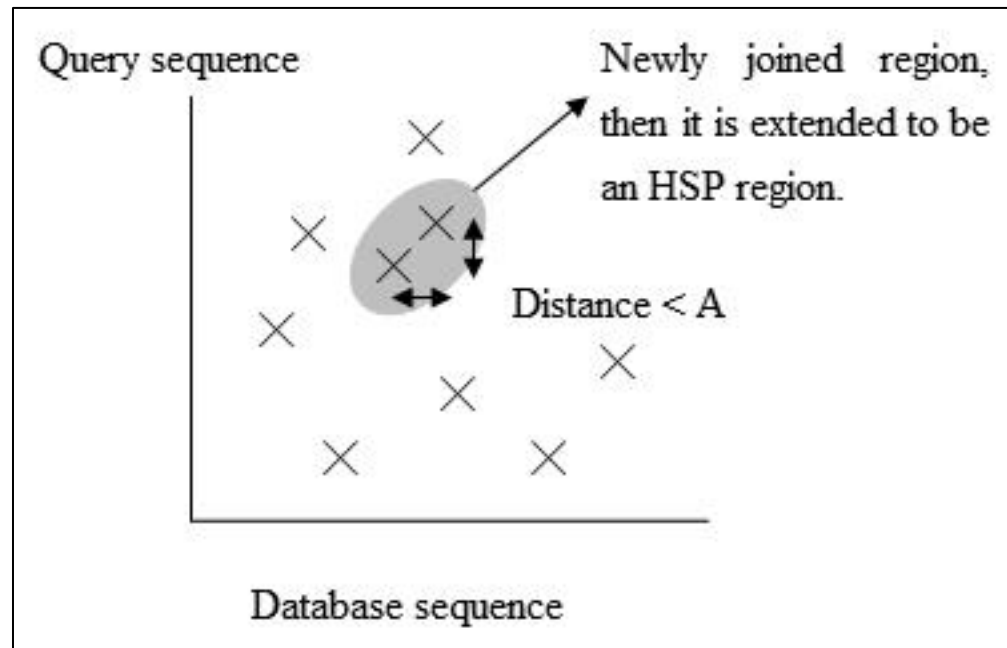
- **Step 2:** Finding "High-scoring segments"
- For each k -letter word (default 3 aa) in the query sequence, search for all permutations with an alignment score above a certain threshold.



- 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-value following p

m

n

K, λ

- From this t with a scor

- Remember 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	

and the

search strategy)

f alignments

e alignment

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 $\geq S$.
- The table to the left shows the $E \rightarrow p$ 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	<i>p</i> -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**:

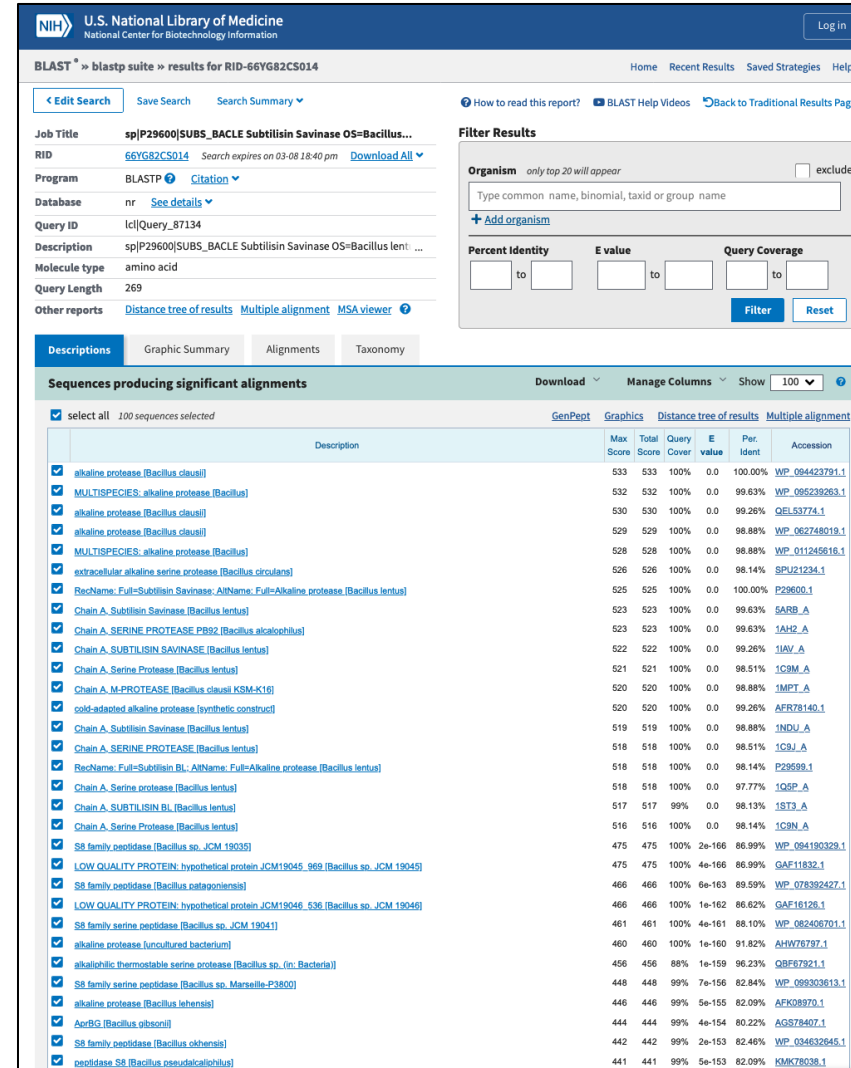
$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

- 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:

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

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)



U.S. National Library of Medicine
National Center for Biotechnology Information

BLASTP » blastp suite » results for RID-66YG82CS014

Job Title: sp|P29600|SUBS_BACLE Subtilisin Savinase OS=Bacillus...
 RID: 66YG82CS014
 Program: BLASTP
 Database: nr
 Query ID: lcl|Query_87134
 Description: sp|P29600|SUBS_BACLE Subtilisin Savinase OS=Bacillus lent...
 Molecule type: amino acid
 Query Length: 269

Filter Results
 Organism: only top 20 will appear
 Percent Identity: [] to []
 E value: [] to []
 Query Coverage: [] to []

Sequences producing significant alignments

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
alkaline protease (Bacillus clausii)	533	533	100%	0.0	100.00%	WP_064423781.1
MULTISPECIES: alkaline protease (Bacillus)	532	532	100%	0.0	99.63%	WP_095239263.1
alkaline protease (Bacillus clausii)	530	530	100%	0.0	99.28%	QEL53774.1
alkaline protease (Bacillus clausii)	529	529	100%	0.0	99.88%	WP_062748019.1
MULTISPECIES: alkaline protease (Bacillus)	528	528	100%	0.0	99.88%	WP_011245816.1
extracellular alkaline serine protease (Bacillus circulans)	526	526	100%	0.0	98.14%	SPU21234.1
RecName: Full=Subtilisin Savinase; AName: 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%	GARB_A
Chain A: SERINE PROTEASE PB92 (Bacillus alkalophilus)	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	99.88%	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	99.88%	1NDU_A
Chain A: SERINE PROTEASE (Bacillus lentus)	518	518	100%	0.0	98.51%	1C9J_A
RecName: Full=Subtilisin BL; AName: Full=Alkaline protease (Bacillus lentus)	518	518	100%	0.0	98.14%	P29599.1
Chain A: Serine protease (Bacillus lentus)	518	518	100%	0.0	97.77%	1QSP_A
Chain A: SUBTILISIN BL (Bacillus lentus)	517	517	99%	0.0	98.13%	1ST3_A
Chain A: Serine Protease (Bacillus lentus)	516	516	100%	0.0	98.14%	1C9N_A
S8 family peptidase (Bacillus sp. JCM 19035)	475	475	100%	2e-166	86.99%	WP_064190329.1
LOW QUALITY PROTEIN: hypothetical protein JCM19045_969 (Bacillus sp. JCM 19045)	475	475	100%	4e-166	86.99%	GAF11832.1
S8 family peptidase (Bacillus pasteurianus)	466	466	100%	6e-163	89.59%	WP_078392427.1
LOW QUALITY PROTEIN: hypothetical protein JCM19046_536 (Bacillus sp. JCM 19046)	466	466	100%	1e-162	86.62%	GAF16128.1
S8 family serine peptidase (Bacillus sp. JCM 19041)	461	461	100%	4e-161	88.10%	WP_082406701.1
alkaline protease (uncultured bacterium)	460	460	100%	1e-160	91.82%	AHW76787.1
alkaliphilic thermostable serine protease (Bacillus sp. (in: Bacteria))	456	456	88%	1e-159	96.23%	QRE67921.1
S8 family serine peptidase (Bacillus sp. Marseille-P3800)	448	448	99%	7e-156	82.84%	WP_099503813.1
alkaline protease (Bacillus lehensis)	446	446	99%	5e-155	82.09%	AFK08970.1
AcroBG (Bacillus gilsonii)	444	444	99%	4e-154	80.22%	AGS78407.1
S8 family peptidase (Bacillus oshensis)	442	442	99%	2e-153	82.46%	WP_034632845.1
peptidase S8 (Bacillus neustalticillus)	441	441	99%	5e-153	82.09%	KMK76038.1

NCBI BLASTP report

Descriptions							
Graphic Summary		Alignments		Taxonomy			
Sequences producing significant alignments				Download	Manage Columns	Show	100
<input checked="" type="checkbox"/> select all <i>100 sequences selected</i>				GenPept	Graphics	Distance tree of results	Multiple alignment
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	alkaline protease [Bacillus clausii]	533	533	100%	0.0	100.00%	WP_094423791.1
<input checked="" type="checkbox"/>	MULTISPECIES: alkaline protease [Bacillus]	532	532	100%	0.0	99.63%	WP_095239263.1
<input checked="" type="checkbox"/>	alkaline protease [Bacillus clausii]	530	530	100%	0.0	99.26%	QEL53774.1
<input checked="" type="checkbox"/>	alkaline protease [Bacillus clausii]	529	529	100%	0.0	98.88%	WP_062748019.1
<input checked="" type="checkbox"/>	MULTISPECIES: alkaline protease [Bacillus]	528	528	100%	0.0	98.88%	WP_011245616.1
<input checked="" type="checkbox"/>	extracellular alkaline serine protease [Bacillus circulans]	526	526	100%	0.0	98.14%	SPU21234.1
<input checked="" type="checkbox"/>	RecName: Full=Subtilisin Savinase; AltName: Full=Alkaline protease [Bacillus lentus]	525	525	100%	0.0	100.00%	P29600.1
<input checked="" type="checkbox"/>	Chain A, Subtilisin Savinase [Bacillus lentus]	523	523	100%	0.0	99.63%	5ARB_A
<input checked="" type="checkbox"/>	Chain A, SERINE PROTEASE PB92 [Bacillus alcalophilus]	523	523	100%	0.0	99.63%	1AH2_A
<input checked="" type="checkbox"/>	Chain A, SUBTILISIN SAVINASE [Bacillus lentus]	522	522	100%	0.0	99.26%	1IAV_A
<input checked="" type="checkbox"/>	Chain A, Serine Protease [Bacillus lentus]	521	521	100%	0.0	98.51%	1C9M_A
<input checked="" type="checkbox"/>	Chain A, M-PROTEASE [Bacillus clausii KSM-K16]	520	520	100%	0.0	98.88%	1MPT_A
<input checked="" type="checkbox"/>	cold-adapted alkaline protease [synthetic construct]	520	520	100%	0.0	99.26%	AFR78140.1
<input checked="" type="checkbox"/>	Chain A, Subtilisin Savinase [Bacillus lentus]	519	519	100%	0.0	98.88%	1NDU_A

NCBI BLASTP report

[Descriptions](#) | **Graphic Summary** | [Alignments](#) | [Taxonomy](#)

[hover to see the title](#) | [click to show alignments](#) | Show Conserved Domains

Alignment Scores: ■ < 40 | ■ 40 - 50 | ■ 50 - 80 | ■ 80 - 200 | ■ >= 200

100 sequences selected **?** Putative conserved domains have been detected, click on the image below for detailed results.

Query seq. 50 100 150 200 250 269
 active site
 catalytic residues
 Specific hits: Peptidases_S8_Subtilisin_subset
 Superfamilies: Peptidases_S8_S53 superfamily

Distribution of the top 100 Blast Hits on 100 subject sequences

How to work on BLAST@NCBI with previous searches

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

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST® Home **Recent Results** Saved Strategies Help

i Important update
The core nucleotide database (*core_nt*) is now the default nucleotide BLAST database. [Learn more about core_nt.](#)

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](#)

NEWS
Non-interactive searches of nt switch to core_nt
Starting late September 2024 all non-interactive WebBLAST and PrimerBLAST searches of ``nt`` will
Tue, 24 Sep 2024 [More BLAST news...](#)

Web BLAST

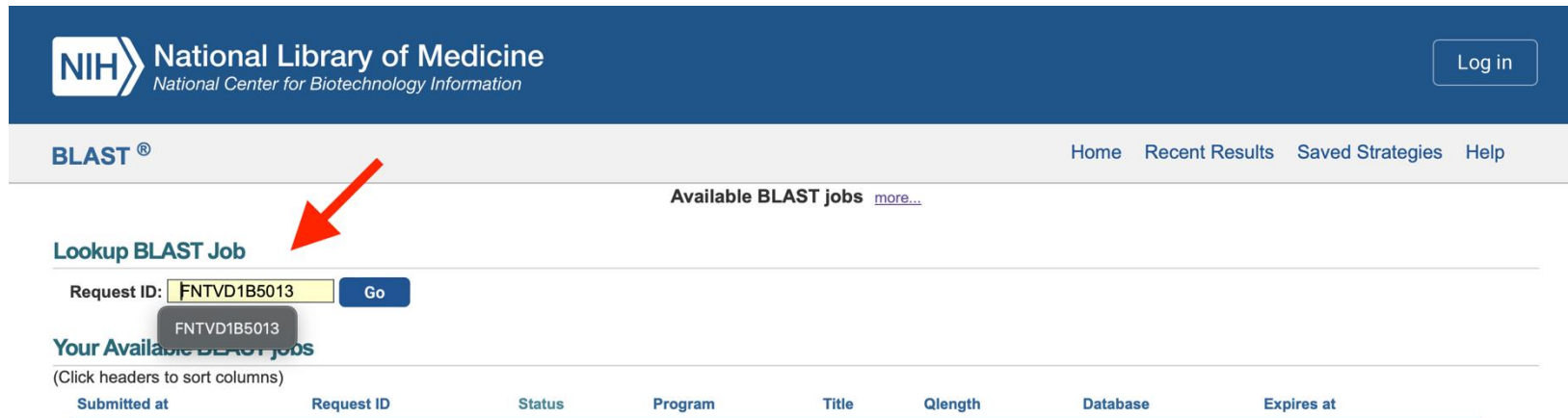
Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

How to work on BLAST@NCBI with previous searches



The screenshot shows the NCBI BLAST@NCBI interface. At the top, there is a blue header with the NIH logo and the text "National Library of Medicine National Center for Biotechnology Information". A "Log in" button is located in the top right corner. Below the header, the word "BLAST®" is displayed on the left, and navigation links for "Home", "Recent Results", "Saved Strategies", and "Help" are on the right. A link for "Available BLAST jobs" with a "more..." link is centered below the navigation. The "Lookup BLAST Job" section features a text input field containing "FNTVD1B5013" and a "Go" button. A red arrow points to the input field. Below this, the "Your Available BLAST Jobs" section is shown, with a tooltip displaying "FNTVD1B5013" over the input field. A table header is visible below, with columns: Submitted at, Request ID, Status, Program, Title, Qlength, Database, and Expires at.

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST® Home Recent Results Saved Strategies Help

Available BLAST jobs [more...](#)

Lookup BLAST Job

Request ID: Go

Your Available BLAST Jobs

(Click headers to sort columns)

Submitted at	Request ID	Status	Program	Title	Qlength	Database	Expires at
--------------	------------	--------	---------	-------	---------	----------	------------