BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 **Lecture 13**

1

Today's topics

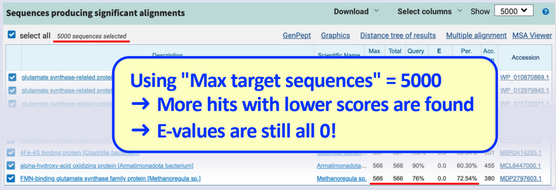
- Assignment 4 Review
- BLAST website and options
  - blastp and blastn
- BLAST and FASTA statistics

2

---

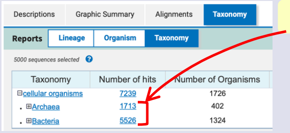
blastp search parameters:
Max target sequences

Query: Q58746.1 (AGLUS_METJA)



	Description	Scientific Name	Max	Total	Query	E	Per.	Acc. Len	Accession
☑	glutamate synthase-related protei								WP_010870869.1
☑	glutamate synthase-related protei								WP_012979840.1
☑	glutamate synthase-related protei								WP_015732971.1
☑	4Fe-4S binding protein [Clostridia bacterium]								MBR0414295.1
☑	alpha-hydroxy-acid oxidizing protein [Armatimonadota bacterium]	Armatimonadota...	566	566	90%	0.0	60.30%	455	MCL6447000.1
☑	FMN-binding glutamate synthase family protein [Methanoregula sp.]	Methanoregula sp.	566	566	76%	0.0	72.54%	380	MDP2797603.1

Using "Max target sequences" = 5000
→ More hits with lower scores are found
→ E-values are still all 0!



Taxonomy	Number of hits	Number of Organisms
⊟cellular organisms	7239	1726
. ⊞Archaea	1713	402
. ⊞Bacteria	5526	1324

All hits are from bacteria/archaea

Even using ClusteredNR!

⊟cellular organisms	4999	1474
. ⊞Archaea	1104	355
. ⊞Bacteria	3895	1119

Any similar proteins in eukaryotes?

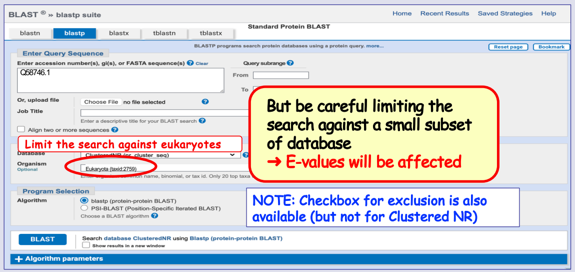
3

---

blastp similarity search:
[A case study] How to find more divergent hits

Query: Q58746.1 (AGLUS_METJA) Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]
→All similar sequences found so far are from bacteria and archaea

- Can we find similar sequences in eukaryotes?



Enter accession number(s), gi(s), or FASTA sequence(s): Q58746.1
Or, upload file: no file selected
Database: ClusteredNR (nr_cluster_seq)
Organism (Optional): Eukaryota (taxid:2759)
Algorithm: ● blastp (protein-protein BLAST) ○ PSI-BLAST (Position-Specific Iterated BLAST)
Search database ClusteredNR using Blastp (protein-protein BLAST)

Limit the search against eukaryotes

But be careful limiting the search against a small subset of database
→ E-values will be affected

NOTE: Checkbox for exclusion is also available (but not for Clustered NR)

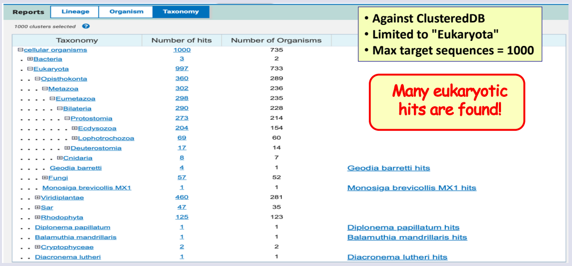
4

---

blastp similarity search:
[A case study] How to find more divergent hits

Query: Q58746.1 (AGLUS_METJA) Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]
→All similar sequences found so far are from bacteria and archaea

- Can we find similar sequences in eukaryotes?



Taxonomy	Number of hits	Number of Organisms	
⊟cellular organisms	1000	735	
. ⊞Bacteria	3	2	
. ⊟Eukaryota	997	733	
. . ⊟Opisthokonta	360	289	
. . . ⊟Metazoa	302	236	
. . . . ⊟Eumetazoa	298	235	
. . . . . ⊟Bilateria	290	228	
. . . . . . ⊟Protostomia	273	214	
. . . . ⊞Ecdysozoa	204	154	
. . . . ⊞Lophotrochozoa	69	60	
. . . . . . ⊞Deuterostomia	17	14	
. . . . . ⊞Cnidaria	8	7	
. . . . Geodia barretti	4	1	Geodia barretti hits
. . . ⊞Fungi	57	52	
. . . Monosiga brevicollis MX1	1	1	Monosiga brevicollis MX1 hits
. . ⊞Viridiplantae	460	281	
. . ⊞Sar	47	35	
. . ⊞Rhodophyta	125	123	
. . Diplonema papillatum	1	1	Diplonema papillatum hits
. . Balamuthia mandrillaris	1	1	Balamuthia mandrillaris hits
. . ⊞Cryptophyceae	2	2	
. . Diacronema lutheri	1	1	Diacronema lutheri hits

- Against ClusteredDB
- Limited to "Eukaryota"
- Max target sequences = 1000

Many eukaryotic hits are found!

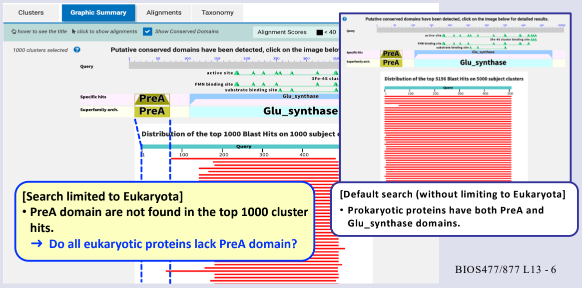
5

---

blastp similarity search:
[A case study] How to find more divergent hits

Query: Q58746.1 (AGLUS_METJA) Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]
→All similar sequences found so far are from bacteria and archaea

- Can we find similar sequences in eukaryotes?



[Search limited to Eukaryota]
- PreA domain are not found in the top 1000 cluster hits.
  → Do all eukaryotic proteins lack PreA domain?

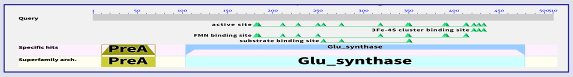
[Default search (without limiting to Eukaryota]
- Prokaryotic proteins have both PreA and Glu_synthase domains.

6

blastp similarity search:
[A case study] region-specific search

Query: Q58746.1 (AGLUS_METJA) Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]
→All eukaryotic sequences found so far are missing the N-terminal domain "PreA"

- Can we find eukaryotic sequences similar to PreA?

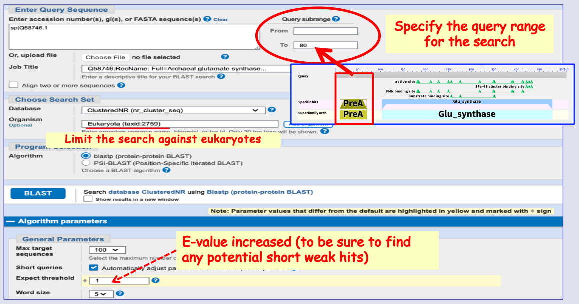


- **Options to perform region specific search:**
  - **"Query subrange" option**
  - **"Max matches in a query range" option**

---

blastp similarity search:
[A case study] region-specific search

Search using the "Query subrange" option



Enter accession number(s), gi(s), or FASTA sequence(s): sp|Q58746.1

Query subrange From: To: 80

Or, upload file: no file selected

Job Title: Q58746:RecName: Full=Archaeal glutamate synthase...

☐ Align two or more sequences

Database: ClusteredNR (nr_cluster_seq)

Organism (Optional): Eukaryota (taxid:2759)

Algorithm: ● blastp (protein-protein BLAST) ○ PSI-BLAST (Position-Specific Iterated BLAST)

Search database ClusteredNR 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

Max target sequences: 100

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 1

Word size: 5

Specify the query range for the search

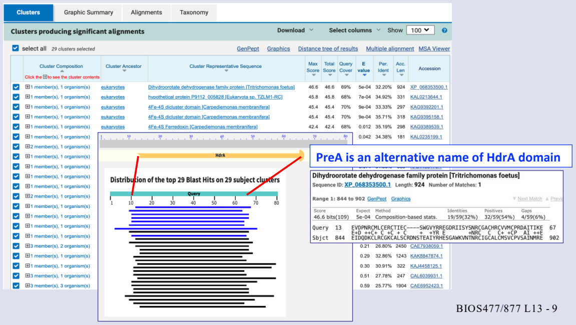
Limit the search against eukaryotes

E-value increased (to be sure to find any potential short weak hits)

---

blastp similarity search:
[A case study] region-specific search

Search using the "Query subrange" option



Clusters producing significant alignments

☑ select all 29 clusters selected

	Cluster Composition	Cluster Ancestor	Cluster Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	⊞1 member(s), 1 organism(s)	eukaryotes	Dihydroorotate dehydrogenase family protein [Tritrichomonas foetus]	46.6	46.6	69%	5e-04	32.20%	924	XP_068353500.1
☑	⊞1 member(s), 1 organism(s)	eukaryotes	hypothetical protein P9112_005828 [Eukaryota sp. TZLM1-RC]	45.8	45.8	68%	7e-04	34.92%	331	KAL0213644.1
☑	⊞1 member(s), 1 organism(s)	eukaryotes	4Fe-4S dicluster domain [Carpediemonas membranifera]	45.4	45.4	70%	9e-04	33.33%	297	KAG9392201.1
☑	⊞1 member(s), 1 organism(s)	eukaryotes	4Fe-4S dicluster domain [Carpediemonas membranifera]	45.4	45.4	70%	9e-04	35.71%	318	KAG9395158.1
☑	⊞1 member(s), 1 organism(s)	eukaryotes	4Fe-4S Ferredoxin [Carpediemonas membranifera]	42.4	42.4	68%	0.012	35.19%	298	KAG9389539.1
☑	⊞1 member(s), 1 organism(s)						0.042	34.38%	181	KAL0235199.1

Distribution of the top 29 Blast Hits on 29 subject clusters

PreA is an alternative name of HdrA domain

Dihydroorotate dehydrogenase family protein [Tritrichomonas foetus]

Sequence ID: XP_068353500.1 Length: 924 Number of Matches: 1

Range 1: 844 to 902 GenPept Graphics

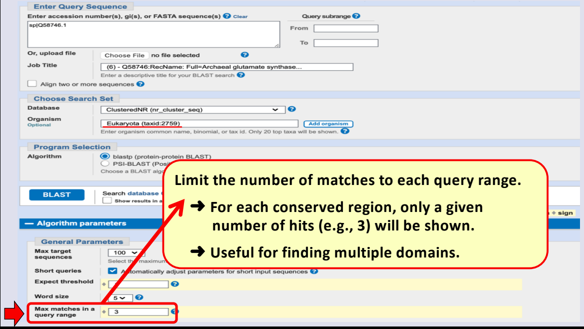
Score	Expect	Method	Identities	Positives	Gaps
46.6 bits(109)	5e-04	Composition-based stats.	19/59(32%)	32/59(54%)	4/59(6%)

```
Query  13   EVDPNRCMLCERCTIEC----SWGVYRREGDRIISYSNRCGACHRCVVMCPRDAITIKE  67
            E+D ++C+ C +C + C        +  +YR E          +NRC  C  C+ +CP  AI ++E
Sbjct  844  EIDQDKCLRCGKCALSCRDNSTEAIYRHESGAWKVNTNRCIGCALCMSVCPVSAINMRE  902
```

---

blastp similarity search:
[A case study] region-specific search

Search using the "Max matches in a query range" option



Enter accession number(s), gi(s), or FASTA sequence(s): sp|Q58746.1

Query subrange From: To:

Or, upload file: no file selected

Job Title: (6) - Q58746:RecName: Full=Archaeal glutamate synthase...

☐ Align two or more sequences

Database: ClusteredNR (nr_cluster_seq)

Organism (Optional): Eukaryota (taxid:2759)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Algorithm: ● blastp (protein-protein BLAST) ○ PSI-BLAST

Max target sequences: 100

Short queries: ☑ Automatically adjust parameters for short input sequences

Word size: 5

Max matches in a query range: 3

Limit the number of matches to each query range.

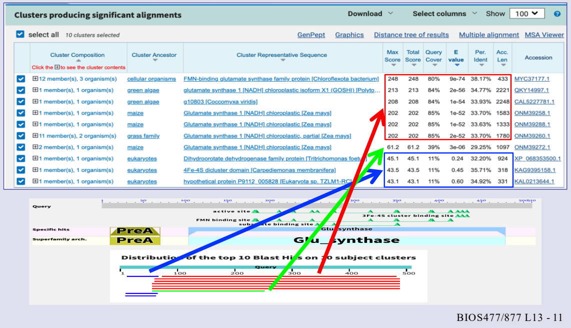
→ For each conserved region, only a given number of hits (e.g., 3) will be shown.

→ Useful for finding multiple domains.

---

blastp similarity search:
[A case study] region-specific search

Search using the "Max matches in a query range" option



Clusters producing significant alignments

☑ select all 10 clusters selected

	Cluster Composition	Cluster Ancestor	Cluster Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	⊞12 member(s), 3 organism(s)	cellular organisms	FMN-binding glutamate synthase family protein [Chloroflexota bacterium]	248	248	80%	9e-74	38.17%	433	MYC37177.1
☑	⊞1 member(s), 1 organism(s)	green algae	glutamate synthase 1 [NADH] chloroplastic isoform X1 (GOSHI) [Polyto...	213	213	84%	2e-56	34.77%	2221	QKY14997.1
☑	⊞1 member(s), 1 organism(s)	green algae	g10803 [Coccomyxa viridis]	208	208	84%	1e-54	33.93%	2248	CAL5227781.1
☑	⊞1 member(s), 1 organism(s)	maize	Glutamate synthase 1 [NADH] chloroplastic [Zea mays]	202	202	85%	1e-52	33.70%	1583	ONM39258.1
☑	⊞1 member(s), 1 organism(s)	maize	Glutamate synthase 1 [NADH] chloroplastic [Zea mays]	202	202	85%	1e-52	33.63%	1333	ONM39288.1
☑	⊞11 member(s), 2 organism(s)	grass family	Glutamate synthase 1 [NADH] chloroplastic, partial [Zea mays]	202	202	85%	2e-52	33.70%	1780	ONM39260.1
☑	⊞2 member(s), 1 organism(s)	maize	Glutamate synthase 1 [NADH] chloroplastic [Zea mays]	61.2	61.2	39%	3e-06	29.25%	1097	ONM39272.1
☑	⊞1 member(s), 1 organism(s)	eukaryotes	Dihydroorotate dehydrogenase family protein [Tritrichomonas foetus]	45.1	45.1	11%	0.24	32.20%	924	XP_068353500.1
☑	⊞1 member(s), 1 organism(s)	eukaryotes	4Fe-4S dicluster domain [Carpediemonas membranifera]	43.5	43.5	11%	0.45	35.71%	318	KAG9395158.1
☑	⊞1 member(s), 1 organism(s)	eukaryotes	hypothetical protein P9112_005828 [Eukaryota sp. TZLM1-RC]	43.1	43.1	11%	0.60	34.92%	331	KAL0213644.1

Distribution of the top 10 Blast Hits on 10 subject clusters

---

blastp similarity search:
[A case study] region-specific search

(Downloaded in "Hit Table (csv)", imported to Excel)

Without Max Matches

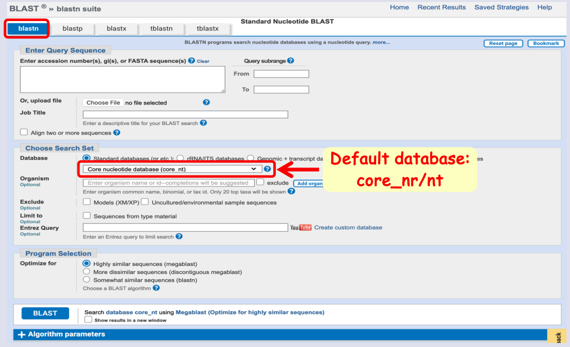
query	subject	% identity	alignment length	mis-matches	gap opens	query start	query end	subject start	subject end	e-value	bit score	% positives
Q58746.1	MYC37177.1	38.173	427	220	4	74	483	5	404	8.59E-74	248	51.99
Q58746.1	MCL4104575.1	40.938	320	180	4	178	491	869	1185	7.60E-60	223	54.37
Q58746.1	MCL4143097.1	41.009	317	174	4	183	491	217	528	2.13E-59	219	54.26
Q58746.1	XP_072036904.1	37.978	366	202	6	144	498	865	1216	3.72E-59	221	52.73
Q58746.1	GHP11700.1	40.171	351	179	5	167	491	924	1269	6.86E-59	221	52.14
…						…						
Q58746.1	CAL5227781.1	33.933	445	261	9	61	490	918	1344	1.15E-54	208	48.09

(This is 78th hit!)

query	subject	% identity	alignment length	mis-matches	gap opens	query start	query end	subject start	subject end	e-value	bit score	% positives
Q58746.1	MYC37177.1	38.173	427	220	4	74	483	5	404	8.59E-74	248	51.99
Q58746.1	QKY14997.1	34.773	440	262	7	61	490	881	1305	2.21E-56	213	48.41
Q58746.1	CAL5227781.1	33.933	445	261	9	61	490	918	1344	1.15E-54	208	48.09
Q58746.1	ONM39258.1	33.703	451	258	10	57	490	299	725	1.13E-52	202	48.12
Q58746.1	ONM39288.1	33.63	449	261	9	57	490	50	476	1.31E-52	202	47.88
Q58746.1	ONM39260.1	33.703	451	258	10	57	490	881	1307	1.52E-52	202	48.12
Q58746.1	ONM39272.1	29.245	212	116	7	57	253	881	1073	2.62E-06	61.2	43.4
Q58746.1	XP_068353500.1	32.203	59	36	1	13	67	844	902	0.24	45.1	54.24
Q58746.1	KAG9395158.1	35.714	56	36	0	11	66	215	270	0.45	43.5	50
		34.921	63	32	2	10	63	267	329	0.6	43.1	50.79

With Max Matches = 3

blastn similarity search



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

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 da…

Core nucleotide database (core_nt)

Organism (Optional): Enter organism name or id--completions will be suggested ☐ exclude Add organ…

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): 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 core_nt using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

Default database: core_nr/nt

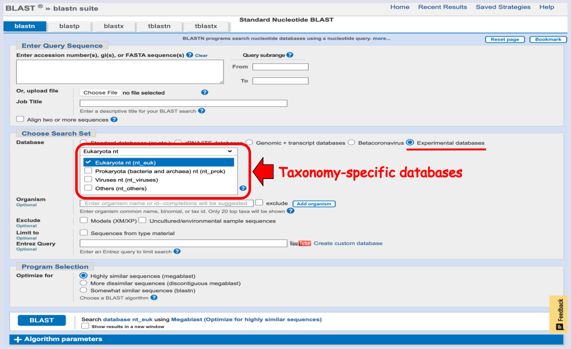
BLAST databases (nucleotide)

Nucleotide database	Description
Core nucleotide database (core_nt) default	Non-redundant nucleotide sequences from GenBank+EMBL+DDBJ+PDB+RefSeq, excluding ETS, STS, GSS, WGS, TSA, patent sequences, HTGS, and most eukaryote chromosome sequences; smaller than nr/nt
Nucleotide Collection (nr/nt)	Partially non-redundant nucleotide sequences from GenBank+EMBL+DDBJ+PDB+RefSeq, excluding ETS, STS, GSS, WGS, TSA, patent sequences, HTGS, and sequences > 100Mb
RefSeq databases	RefSeq RNA, RefSeq Select (human and mouse), RefSeq Genome, Human RefSeqGene, *etc.*
Whole-Genome-Shotgun contigs (WGS), Expressed sequence tags (EST), Sequence Read Archive (SRA), Transcriptome Shotgun Assembly (TSA), High Throughput Genomic Sequences (HTGS), Genomic survey sequences (GSS), Sequence tagged sites (STS)	
PDB nucleotide sequences	Sequences from the Protein Data Bank (PDB)
Patent sequences	Nucleotide sequences derived from the Patent division of GenBank
16S, 18S, 28S ribosomal RNA databases	Ribosomal RNA sequences including 16S (from bacteria and archaea type strains) and 18S (SSU)/28S (LS&) from fungi type and reference material)
Internal transcribed spacer region	Internal transcribed spacer region (ITS) from Fungi type and reference material

and more…

NCBI BLAST FTP site

BLAST databases (nucleotide)



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

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 ● Experimental databases

Eukaryota nt

- ☑ Eukaryota nt (nt_euk)
- ☐ Prokaryota (bacteria and archaea) nt (nt_prok)
- ☐ Viruses nt (nt_viruses)
- ☐ Others (nt_others)

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): 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)

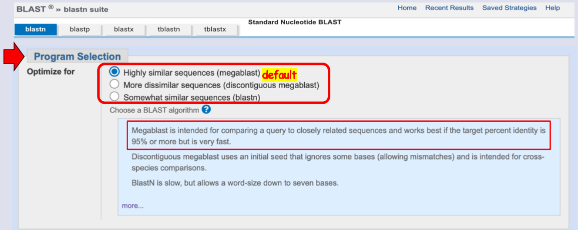
Choose a BLAST algorithm

Search database nt_euk using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

Taxonomy-specific databases

Nucleotide search algorithms



Standard Nucleotide BLAST

Program Selection

Optimize for:

- ● Highly similar sequences (megablast) default
- ○ More dissimilar sequences (discontiguous megablast)
- ○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast.

Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.

BlastN is slow, but allows a word-size down to seven bases.

more...

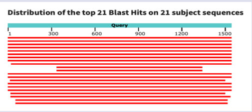
- megablast: w=28 (16~256) This is the default search method
- discontiguous megablast: allows some mismatches w=11 (or 12)
- blastn: w=11 (7~15) or w=7 for a short sequence

---

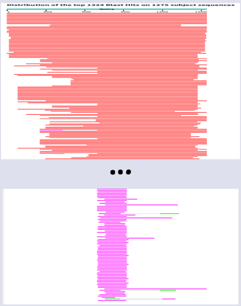
megablast *vs.* blastn

Query: DQ018115.1 M. jannaschii glutamate synthase gene, complete cds
Max Target sequences: 5000
Expect threshold: 0.0001

[megablast: only 21 hits, E=0 for 15 hits]



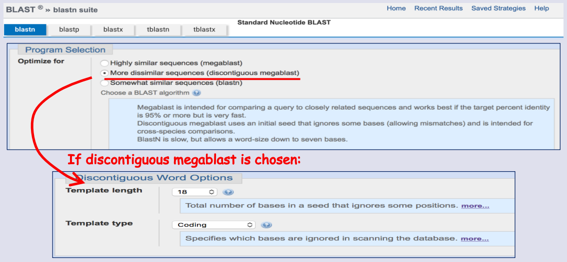
[blastn: 1275 hits, E<0.0001]



megablast:
for searching closely related sequences (>95%), very fast
word size = 16~64 (default: 28)

blastn:
more sensitive, slow,
word size = 7~15 (default: 11)

Discontiguous megablast



Standard Nucleotide BLAST

Program Selection

Optimize for:

- ○ Highly similar sequences (megablast)
- ● More dissimilar sequences (discontiguous megablast)
- ○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast.
Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.
BlastN is slow, but allows a word-size down to seven bases.

If discontiguous megablast is chosen:

Discontiguous Word Options

Template length: 18

Total number of bases in a seed that ignores some positions. more...

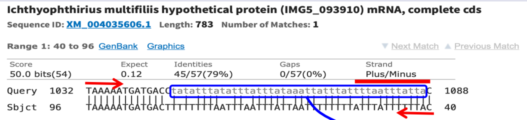
Template type: Coding

Specifies which bases are ignored in scanning the database. more...

Word matching based on discontiguous pattern (template):
e.g., for coding: 110110110110110 (w=11, t=16)
➜ mismatches are allowed for positions marked '0'

blastn/blastx output: examples

[blastn]



Ichthyophthirius multifiliis hypothetical protein (IMG5_093910) mRNA, complete cds
Sequence ID: XM_004035606.1 Length: 783 Number of Matches: 1

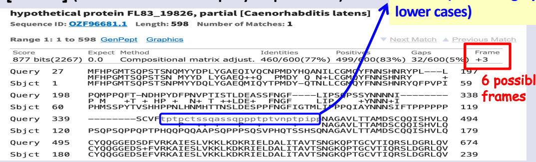
```
Range 1: 40 to 96  GenBank  Graphics                        ▼ Next Match  ▲ Previous Match

Score            Expect  Identities     Gaps        Strand
50.0 bits(54)    0.12    45/57(79%)     0/57(0%)    Plus/Minus

Query  1032  TAAAAATGATGACCtatatttatatttatttataaattatttattttaatttattaC  1088
             ||||||||||||| | | ||||  || |||||| |||| ||| | || |||| ||||
Sbjct  96    TAAAAATGATGACTTTTTTTTAATTTAATTTATTAATTTTTTTATTTATTTTTTAC  40
```

Low complexity region is masked (shown in gray lower cases)

[blastx] (translated nucleotide query vs. protein db)



hypothetical protein FL83_19826, partial [Caenorhabditis latens]
Sequence ID: OZF96681.1 Length: 598 Number of Matches: 1

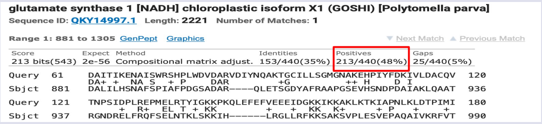
```
Range 1: 1 to 598  GenPept  Graphics                       ▼ Next Match  ▲ Previous Match

Score           Expect  Method                          Identities      Positives       Gaps          Frame
877 bits(2267)  0.0     Compositional matrix adjust.    460/600(77%)    499/600(83%)    32/600(5%)    +3

Query  27    MFHPGMTSQPSTSNQMYYDPLYGAEQIVQCNPMDYHQANILCGMQYFNNSHNRYPL---L  197
             MFHPGMTSQPSTSN MYYD LYGAEQ++Q  PMDY Q N+LCGMQYFNNSHNRY     +
Sbjct  1     MFHPGMTSQPSTSNPMYYDQLYGAEQMIQYTPMDY-QTNLLCGMQYFNNSHNRYQFPVPI  59
Query  198   PQMPPQFT-NDHPYDFPNVPTISTLDEASSFNGF----LIPSQPSSYNNNNI--------  338
             P M   +T + HP +  N+ T ++LDE+  FNGF    LIP    +YNNN+I
Sbjct  60    PHMSSPYTVSHHPPNLHNMHTTNSLDESPPFNGFIGTMLIPPPQIAYNNNSIFTPPPPPP  119
Query  339   --------SCVFtptpctssqassqppptptvnptpippNAGAVLTTAMDSCQQISHVLQ  494
                            P  ++ +   P  +     +    NAGAVLTTAMDSCQQISHVLQ
Sbjct  120   PSQPSQPPQPTPHQQPQQAAPSQPPPSQSVPHQTSSHSQNAGAVLTTAMDSCQQISHVLQ  179
Query  495   CYQQGGEDSDFVRKAIESLVKKLKDKRIELDALITAVTSNGKQPTGCVTIQRSLDGRLQV  674
             CYQQGGEDS+FVRKAIESLVKKLKDKRIELDALITAVTSNGKQPTGCVTIQRSLDGRLQV
Sbjct  180   CYQQGGEDSEFVRKAIESLVKKLKDKRIELDALITAVTSNGKQPTGCVTIQRSLDGRLQV  239
```

6 possible frames

blastp output: examples



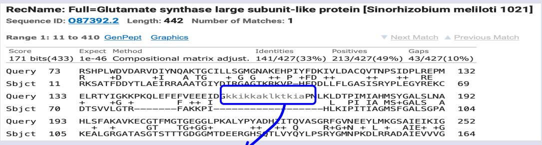
glutamate synthase 1 [NADH] chloroplastic isoform X1 (GOSHI) [Polytomella parva]
Sequence ID: QKY14997.1 Length: 2221 Number of Matches: 1

```
Range 1: 881 to 1305  GenPept  Graphics                     ▼ Next Match  ▲ Previous Match

Score           Expect  Method                          Identities      Positives       Gaps
213 bits(543)   2e-56   Compositional matrix adjust.    153/440(35%)    213/440(48%)    25/440(5%)

Query  61    DAITIKENAISWRSHPLWDVDARVDIYNQAKTGCILLSGMGNAKEHPIYFDKIVLDACQV  120
             DA+ +  NA S  + P    DAR        +G         ++ H    D I
Sbjct  881   DALILHSNAFSPIAFPDGSADAR----QLETSGDYAFRAAPGSEVHSNDPDAIAKLQAAT  936
Query  121   TNPSIDPLREPMELRTYIGKKPKQLEFEFVEEEIDGKKIKKAKLKTKIAPNLKLDTPIMI  180
                  +  R+  EL T + KK        +   +  KK  K+     + P   +    +
Sbjct  937   RGNDRELFRQFSELNTKLSKKIH------LRGLLRFKKSAKSVPLESVEPAQAIVKRFVT  990
```

Positives (+): Similar amino acid pairs
These AA pairs have positive scores in the scoring matrix used



RecName: Full=Glutamate synthase large subunit-like protein [Sinorhizobium meliloti 1021]
Sequence ID: O87392.2 Length: 442 Number of Matches: 1

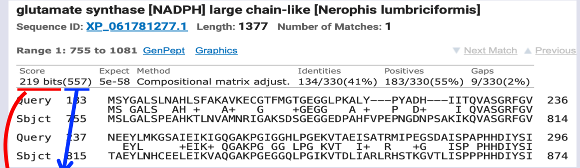
```
Range 1: 11 to 410  GenPept  Graphics                       ▼ Next Match  ▲ Previous Match

Score           Expect  Method                          Identities      Positives       Gaps
171 bits(433)   1e-46   Compositional matrix adjust.    141/427(33%)    213/427(49%)    43/427(10%)

Query  73    RSHPLWDVDARVDIYNQAKTGCILLSGMGNAKEHPIYFDKIVLDACQVTNPSIDPLREPM  132
             R    +D     +I   A TG   + G G  ++ P +FD ++     ++   ++  RE
Sbjct  11    RKSATFDDYTLAEIRRAAATGIYDIRGAGTKRKVP-HFDDLLFLGASISRYPLEGYREKC  69
Query  133   ELRTYIGKKPKQLEFEFVEEEIDGkkikkaklktkiaPNLKLDTPIMIAHMSYGALSLNA  192
             +    +G +       F ++ I                   L  PI IA MS+GALS  A
Sbjct  70    DTSVVLGTR-------FAKKPI------------------HLKIPITIAGMSFGALSGPA  104
Query  193   HLSFAKAVKECGTFMGTGEGGLPKALYPYADHIITQVASGRFGVNEEYLMKGSAIEIKIG  252
               +  +     GT   TG+GG+      ++  ++ Q    R+G+N + L +  AIE+ +G
Sbjct  105   KEALGRGATASGTSTTTGDGGMTDEERGHSQTLVYQYLPSRYGMNPKDLRRADAIEVVVG  164
```

Low complexity region is masked

blastp statistics: bit score



glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

```
Range 1: 755 to 1081  GenPept  Graphics                     ▼ Next Match  ▲ Previous

Score           Expect  Method                          Identities      Positives       Gaps
219 bits(557)   5e-58   Compositional matrix adjust.    134/330(41%)    183/330(55%)    9/330(2%)

Query  183   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV  236
             MS GALS  AH +   A+   G    +GEGG   A +   P  D+    I QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814
Query  237   NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL     +EIK+ QGAKPG GG LPG KVT  I+  R   +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

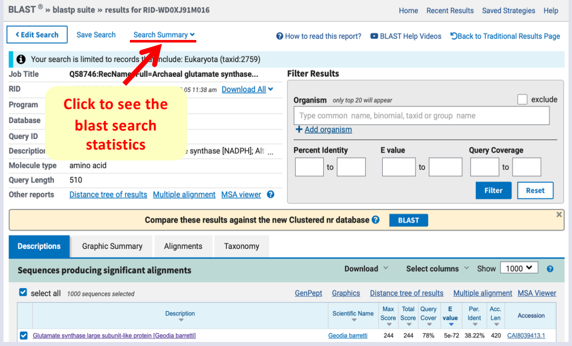
Raw Score (S): simply based on pairwise scores & gap penalties

Normalized Score or **Bit Score (S'_{bit}):**

$$S'_{bit} = (\lambda S - \log_e K) / \log_e 2, \quad [S'_{nat} = \lambda S - \log_e K]$$

λ and *K* are specific to the scoring system used (scoring matrix, gap penalties, etc.)
→ Where can we find the values for λ and *K* ?

blastp search summary



BLAST ® » blastp suite » results for RID-WD0XJ91M016

Your search is limited to records that include: Eukaryota (taxid:2759)

Job Title: Q58746:RecName: Full=Archaeal glutamate synthase...
RID: …05 11:38 am Download All ⌄
Program
Database
Query ID
Description: … synthase [NADPH]; Alt ...
Molecule type: amino acid
Query Length: 510
Other reports: Distance tree of results Multiple alignment MSA viewer

Filter Results
Organism only top 20 will appear ☐ exclude
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity ☐ to ☐ E value ☐ to ☐ Query Coverage ☐ to ☐
Filter Reset

Compare these results against the new Clustered nr database ? BLAST

Sequences producing significant alignments Download ⌄ Select columns ⌄ Show 1000 ⌄

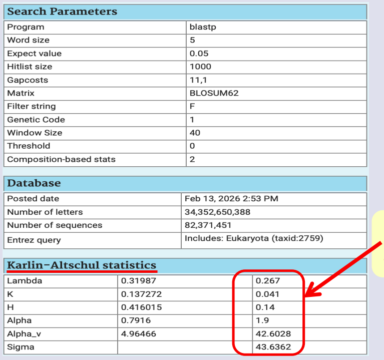
☑ select all 1000 sequences selected GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Glutamate synthase large subunit-like protein [Geodia barretti]	Geodia barretti	244	244	78%	5e-72	38.22%	420	CAI8039413.1

Click to see the blast search statistics

(Top portion of any blast output)

blastp search summary



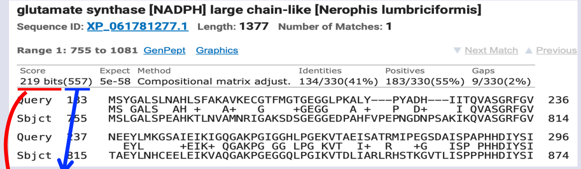
Search Parameters	
Program	blastp
Word size	5
Expect value	0.05
Hitlist size	1000
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	0
Composition-based stats	2

Database	
Posted date	Feb 13, 2026 2:53 PM
Number of letters	34,352,650,388
Number of sequences	82,371,451
Entrez query	Includes: Eukaryota (taxid:2759)

Karlin–Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.416015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Used to calculate bit scores and E-values for the alignments with gaps

blastp statistics: bit score



glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

```
Range 1: 755 to 1081  GenPept  Graphics                     ▼ Next Match  ▲ Previous

Score           Expect  Method                          Identities      Positives       Gaps
219 bits(557)   5e-58   Compositional matrix adjust.    134/330(41%)    183/330(55%)    9/330(2%)

Query  183   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV  236
             MS GALS  AH +   A+   G    +GEGG   A +   P  D+    I QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814
Query  237   NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL     +EIK+ QGAKPG GG LPG KVT  I+  R   +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

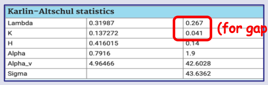
Raw Score (S): simply based on pairwise scores & gap penalties

Normalized Score or **Bit Score (S'_{bit}):**

$$S'_{bit} = (\lambda S - \log_e K) / \log_e 2, \quad [S'_{nat} = \lambda S - \log_e K]$$

$$\lambda = 0.267,\ K = 0.041,\ S'_{bit} = \{0.267 \times 557 - \log_e(0.041)\} / \log_e 2 = 219.2$$

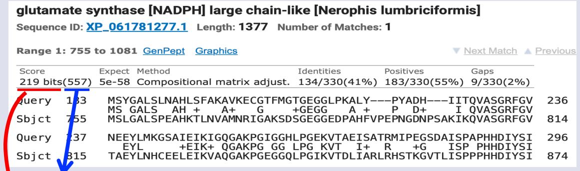
λ and *K* are scoring system specific



Karlin–Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.416015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

(for gapped alignments)

blastp statistics: bit score



glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]

Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

Range 1: 755 to 1081 GenPept Graphics ▼ Next Match ▲ Previous

```
 Score            Expect  Method                        Identities      Positives       Gaps
 219 bits(557)    5e-58   Compositional matrix adjust.  134/330(41%)    183/330(55%)    9/330(2%)

Query  183    MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV  236
              MS GALS  AH +   A+   G     +GEGG    A +    P  D+     I QVASGRFGV
Sbjct  755    MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814

Query  237    NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
                EYL      +EIK+ QGAKPG GG LPG KVT   I+  R   +G   ISP PHHDIYSI
Sbjct  815    TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

Raw Score (S): simply based on pairwise scores & gap penalties

Normalized Score or **Bit Score (S'_{bit}):**

$$S'_{bit} = (\lambda S - \log_e K) / \log_e 2, \quad [S'_{nat} = \lambda S - \log_e K]$$

$$\lambda = 0.267,\ K = 0.041,\ S'_{bit} = \{0.267 \times 557 - \log_e(0.041)\} / \log_e 2 = 219.2$$

Raw scores (S) depend on the scoring system; cannot be compared

Bit scores (S'_{bit}) are normalized using λ and K;

→ independent of scoring system; can be compared

Pairwise alignment vs. database searching

[For a pairwise alignment]

- Probability of getting the alignment score $S \geq x$ by chance

  $P(S \geq x) = 1 - \exp[-Kmne^{-\lambda x}] \approx Kmne^{-\lambda x}$

  Karlin-Altschul equation (Karlin & Altschul, 1990)

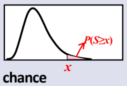
$e^a \approx 1+a$, where $a = -Kmne^{-\lambda x}$

Taylor series approximation

➜ works if a is small (<<1)

[For database searching]

- **Multiple** pairwise alignments ➜ multiple testing problem
  - $P(S \geq x)$: Probability of getting the alignment score (S) larger than x by chance from one pairwise alignment



  - ○ If $P(S \geq x) = 0.05$, $P(S<x) = 1 - P(S \geq x) = 0.95$

    ➜ 0.95 is the prob. to have one pairwise alignment with $S<x$ by chance
  - ○ For 10 alignments, $0.95^{10} \approx 0.60$ is the prob. to have all 10 alignments with $S<x$

    ➜ 1-0.60 = 0.40 is the prob. to have at least one alignment with $S \geq x$ by chance
  - ○ For 100 alignments, $0.95^{100} \approx 0.006$ is the prob. to have all 100- alignments with $S<x$

    ➜ 1-0.006 ≈ 0.99 is the prob. to have at least one alignment with $S \geq x$ by chance

Pr=0.05 as the significance level is not good enough
if many alignments need to be tested!

Bonferroni correction

- Multiple comparison correction
  - Instead of using Prob = α as the threshold, use Prob = α/N (for N comparisons) as the threshold
    - ○ For 10 alignments, use α' = 0.05/10 = 0.005 (instead of 0.05) as the threshold
      - → For $P(S \geq x) = 0.005$, $(1-0.005)^{10} \approx 0.95$ is the probability to have all 10 alignments with $S<x$ by chance
      - → 1-0.95 = 0.05 is the probability to have at least one alignment with $S \geq x$ by chance
    - ○ For 100 alignments, use α' = 0.05/100 = 0.0005 (instead of 0.05) as the threshold
      - → For $P(S \geq x) = 0.0005$, $(1-0.0005)^{100} \approx 0.95$ is the probability to have all 100 alignments with $S<x$ by chance
      - → 1-0.95 = 0.05 is the probability to have at least one alignment with $S \geq x$ by chance

Bonferroni correction in database searching

- Multiple comparison correction
  - Threshold without correction: $P = \alpha$
  - Threshold with correction: $P = \alpha' = \alpha/N$

    (for N comparisons)
  - $E = N \times P$

    ➜ For E-value, using $E = \alpha$ as the threshold is equivalent to using the threshold corrected for multiple comparisons

    ➜ For database searching:

    N = the database size = the number of entries
    = the number of alignments

blastp statistics: E-value

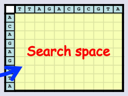
- Karlin-Altschul equation (Karlin & Altschul, 1990)

  [For a pairwise alignment]

  $P = Kmne^{-\lambda S}$ (Lec 11 slide 3)

  m, n: lengths of the sequences compared

  ➜ $m \times n$: search space



  [For database similarity searching]

  Used by FASTA

  $E = Kmne^{-\lambda S}$ (used by BLAST instead of $E = N \times P$)

  → the expected number of HSPs with scores $\geq S$
  - m: length of the query
  - n: length of the database (total number of residues)

  Consider a database as a single very long sequence

  $P = 1 - e^{-E}$ ($P \approx E$ if $E < 0.01$)

  → the probability of having at least one HSP with its score $\geq S$

blastp statistics: E-value

- Karlin-Altschul equation (Karlin & Altschul, 1990)

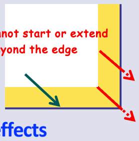
  $E = Km'n'e^{-\lambda S}$ (See also Altschul & Gish, 1996)
  - m': effective length of the query

    $m' = m - l$
  - n': effective length of the database

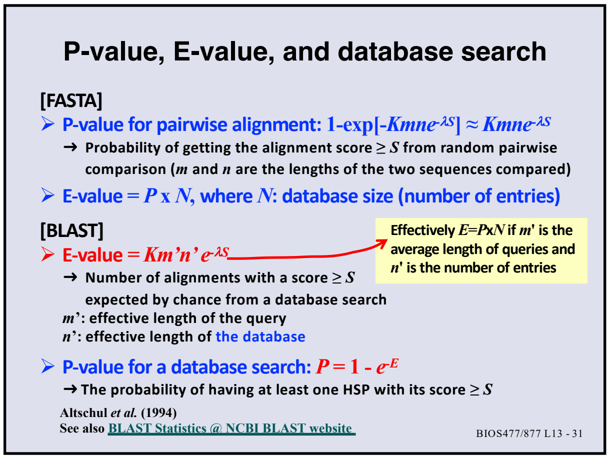
    $n' = n - l \times$ (number of sequences in the database)
  - l: length adjustment ➜ correction for edge effects

    → HSPs cannot occur too close to the search space edges.

    → Effective lengths of HSPs should be shorter than the actual lengths.



For blastn and tblastx: $l = \ln(Kmn)/H$ is used
For blastp, blastx, tblastn: adjusted m' and n' is calculated
using the finite-size correction (FSC) (Park et al., 2012)



P-value, E-value, and database search

[FASTA]

- P-value for pairwise alignment: $1\text{-}\exp[-Kmne^{-\lambda S}] \approx Kmne^{-\lambda S}$
  - → Probability of getting the alignment score ≥ S from random pairwise comparison (m and n are the lengths of the two sequences compared)
- E-value = P x N, where N: database size (number of entries)

[BLAST]

- E-value = $Km'n'e^{-\lambda S}$
  - → Number of alignments with a score ≥ S expected by chance from a database search

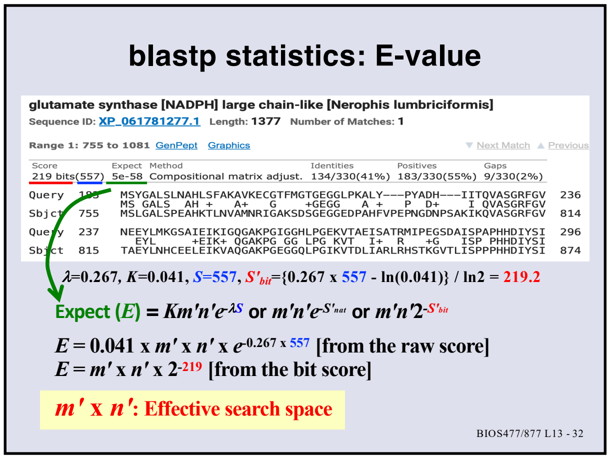
  m': effective length of the query
  n': effective length of the database

Effectively E=PxN if m' is the average length of queries and n' is the number of entries

- P-value for a database search: $P = 1 - e^{-E}$
  - → The probability of having at least one HSP with its score ≥ S

Altschul *et al.* (1994)
See also BLAST Statistics @ NCBI BLAST website

31



blastp statistics: E-value

glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

Range 1: 755 to 1081 GenPept Graphics

```
Score             Expect  Method                          Identities      Positives       Gaps
219 bits(557)     5e-58   Compositional matrix adjust.    134/330(41%)    183/330(55%)    9/330(2%)

Query  185   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV  236
             MS GALS  AH +   A+   G     +GEGG   A +   P  D+     I QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814

Query  237   NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL      +EIK+ QGAKPG GG LPG KVT   I+  R   +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

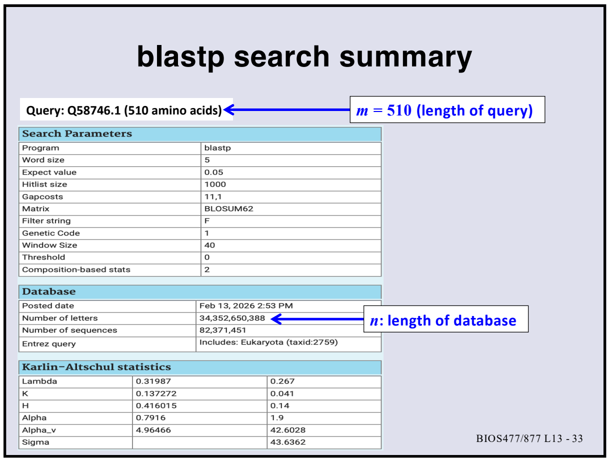
λ=0.267, K=0.041, S=557, S'_{bit}={0.267 x 557 - ln(0.041)} / ln2 = 219.2

Expect (E) = $Km'n'e^{-\lambda S}$ or $m'n'e^{-S'_{nat}}$ or $m'n'2^{-S'_{bit}}$

$E = 0.041 \text{ x } m' \text{ x } n' \text{ x } e^{-0.267 \text{ x } 557}$ [from the raw score]
$E = m' \text{ x } n' \text{ x } 2^{-219}$ [from the bit score]

m' x n': Effective search space

32



blastp search summary

Query: Q58746.1 (510 amino acids)

m = 510 (length of query)

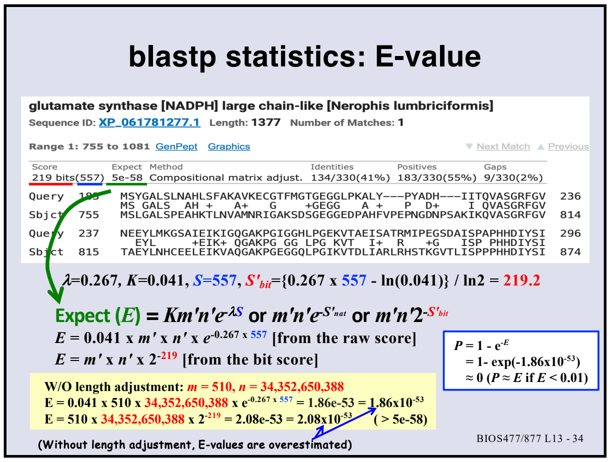
Search Parameters	
Program	blastp
Word size	5
Expect value	0.05
Hitlist size	1000
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	0
Composition-based stats	2

Database	
Posted date	Feb 13, 2026 2:53 PM
Number of letters	34,352,650,388
Number of sequences	82,371,451
Entrez query	Includes: Eukaryota (taxid:2759)

n: length of database

Karlin−Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.416015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

33



blastp statistics: E-value

glutamate synthase [NADPH] large chain-like [Nerophis lumbriciformis]
Sequence ID: XP_061781277.1 Length: 1377 Number of Matches: 1

Range 1: 755 to 1081 GenPept Graphics

```
Score             Expect  Method                          Identities      Positives       Gaps
219 bits(557)     5e-58   Compositional matrix adjust.    134/330(41%)    183/330(55%)    9/330(2%)

Query  185   MSYGALSLNAHLSFAKAVKECGTFMGTGEGGLPKALY---PYADH---IITQVASGRFGV  236
             MS GALS  AH +   A+   G     +GEGG   A +   P  D+     I QVASGRFGV
Sbjct  755   MSLGALSPEAHKTLNVAMNRIGAKSDSGEGGEDPAHFVPEPNGDNPSAKIKQVASGRFGV  814

Query  237   NEEYLMKGSAIEIKIGQGAKPGIGGHLPGEKVTAEISATRMIPEGSDAISPAPHHDIYSI  296
               EYL      +EIK+ QGAKPG GG LPG KVT   I+  R   +G   ISP PHHDIYSI
Sbjct  815   TAEYLNHCEELEIKVAQGAKPGEGGQLPGIKVTDLIARLRHSTKGVTLISPPPHHDIYSI  874
```

λ=0.267, K=0.041, S=557, S'_{bit}={0.267 x 557 - ln(0.041)} / ln2 = 219.2

Expect (E) = $Km'n'e^{-\lambda S}$ or $m'n'e^{-S'_{nat}}$ or $m'n'2^{-S'_{bit}}$

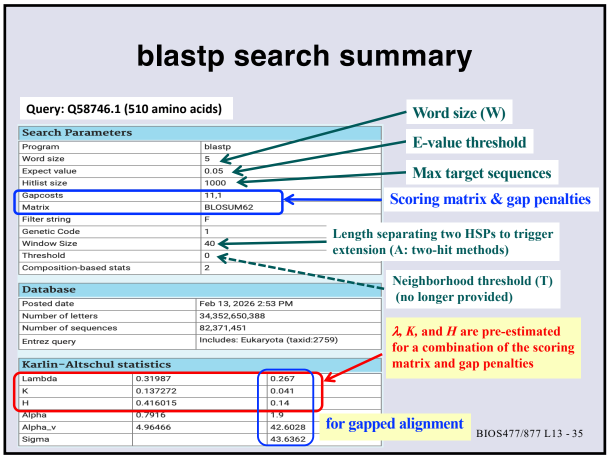
$E = 0.041 \text{ x } m' \text{ x } n' \text{ x } e^{-0.267 \text{ x } 557}$ [from the raw score]
$E = m' \text{ x } n' \text{ x } 2^{-219}$ [from the bit score]

$P = 1 - e^{-E}$
$= 1 - \exp(-1.86\text{x}10^{-53})$
≈ 0 ($P \approx E$ if $E < 0.01$)

W/O length adjustment: m = 510, n = 34,352,650,388
E = 0.041 x 510 x 34,352,650,388 x $e^{-0.267 \text{ x } 557}$ = 1.86e-53 = $1.86\text{x}10^{-53}$
E = 510 x 34,352,650,388 x 2^{-219} = 2.08e-53 = $2.08\text{x}10^{-53}$ (> 5e-58)

(Without length adjustment, E-values are overestimated)

34



blastp search summary

Query: Q58746.1 (510 amino acids)

Search Parameters	
Program	blastp
Word size	5
Expect value	0.05
Hitlist size	1000
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	0
Composition-based stats	2

Database	
Posted date	Feb 13, 2026 2:53 PM
Number of letters	34,352,650,388
Number of sequences	82,371,451
Entrez query	Includes: Eukaryota (taxid:2759)

Karlin−Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.416015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Word size (W)

E-value threshold

Max target sequences

Scoring matrix & gap penalties

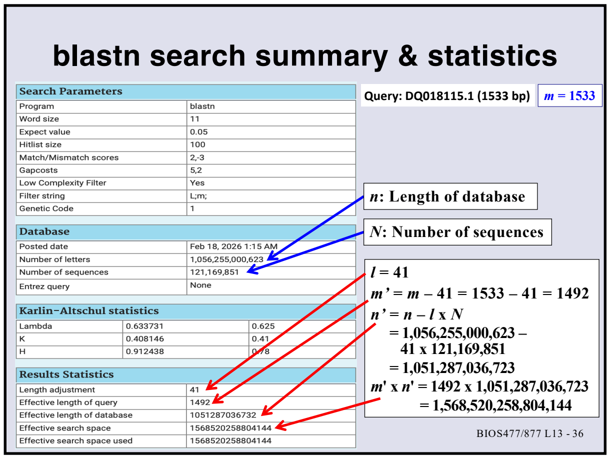
Length separating two HSPs to trigger extension (A: two-hit methods)

Neighborhood threshold (T) (no longer provided)

λ, K, and H are pre-estimated for a combination of the scoring matrix and gap penalties

for gapped alignment

35



blastn search summary & statistics

Query: DQ018115.1 (1533 bp)

m = 1533

Search Parameters	
Program	blastn
Word size	11
Expect value	0.05
Hitlist size	100
Match/Mismatch scores	2,-3
Gapcosts	5,2
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Database	
Posted date	Feb 18, 2026 1:15 AM
Number of letters	1,056,255,000,623
Number of sequences	121,169,851
Entrez query	None

Karlin−Altschul statistics		
Lambda	0.633731	0.625
K	0.408146	0.41
H	0.912438	0.78

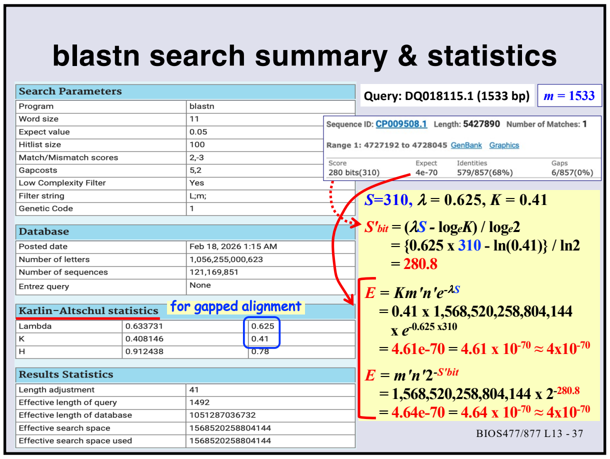
Results Statistics	
Length adjustment	41
Effective length of query	1492
Effective length of database	1051287036732
Effective search space	1568520258804144
Effective search space used	1568520258804144

n: Length of database

N: Number of sequences

l = 41
m' = m – 41 = 1533 – 41 = 1492
n' = n – l x N
= 1,056,255,000,623 – 41 x 121,169,851
= 1,051,287,036,723
m' x n' = 1492 x 1,051,287,036,723
= 1,568,520,258,804,144

36



blastn search summary & statistics

Search Parameters	
Program	blastn
Word size	11
Expect value	0.05
Hitlist size	100
Match/Mismatch scores	2,-3
Gapcosts	5,2
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Database	
Posted date	Feb 18, 2026 1:15 AM
Number of letters	1,056,255,000,623
Number of sequences	121,169,851
Entrez query	None

Karlin−Altschul statistics		for gapped alignment
Lambda	0.633731	0.625
K	0.408146	0.41
H	0.912438	0.78

Results Statistics	
Length adjustment	41
Effective length of query	1492
Effective length of database	1051287036732
Effective search space	1568520258804144
Effective search space used	1568520258804144

Query: DQ018115.1 (1533 bp)

$m = 1533$

Sequence ID: CP009508.1 Length: 5427890 Number of Matches: 1

Range 1: 4727192 to 4728045 GenBank Graphics

Score	Expect	Identities	Gaps
280 bits(310)	4e-70	579/857(68%)	6/857(0%)

$S=310,\ \lambda = 0.625,\ K = 0.41$

$$S'_{bit} = (\lambda S - \log_e K) / \log_e 2$$
$$= \{0.625 \text{ x } 310 - \ln(0.41)\} / \ln 2$$
$$= 280.8$$

$$E = Km'n'e^{-\lambda S}$$
$$= 0.41 \text{ x } 1{,}568{,}520{,}258{,}804{,}144 \text{ x } e^{-0.625 \text{ x} 310}$$
$$= 4.61\text{e-}70 = 4.61 \text{ x } 10^{-70} \approx 4\text{x}10^{-70}$$

$$E = m'n'2^{-S'bit}$$
$$= 1{,}568{,}520{,}258{,}804{,}144 \text{ x } 2^{-280.8}$$
$$= 4.64\text{e-}70 = 4.64 \text{ x } 10^{-70} \approx 4\text{x}10^{-70}$$