

Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 13

BIOS477/877 L13 - 1

1

TODAY'S TOPICS

- BLAST website and options (blastp & blastn)
- BLAST & FASTA Statistics

BIOS477/877 L13 - 2

2

blastp Protein Similarity Search

Query: Q58746.1 (AGLUS_METJA)
Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]

BIOS477/877 L13 - 3

3

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-Q2C1W3301N

BIOS477/877 L13 - 4

4

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-Q2C1W3301N

BIOS477/877 L13 - 5

5

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-Q2C1W3301N

BIOS477/877 L13 - 6

6

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-02C1W3301N

Job Title: spj058746.1
RID: 02C1W3301N
Program: BLASTP
Database: nr
Query ID: 058746.1
Description: RecName: Full=Archaeal glutamate synthase [NADPH]; Alt=...
Molecule type: amino acid
Query Length: 510

Filter Results
Organism: only top 20 will appear
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity: [] to []
E value: [] to []
Query Coverage: [] to []

Compare these results against the new nr database

Query coverage: Proportion of the query aligned

Bit scores E-value

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Methanocaldococcus jannaschii	Methanocaldococcus jannaschii	1033	1033	100%	0.0	100.00%	510	WP_012673869.1
Methanocaldococcus sp. F3406-22	Methanocaldococcus sp. F3406-22	1032	1032	100%	0.0	99.99%	510	WP_012733860.1
Methanocaldococcus subterraneus	Methanocaldococcus subterraneus	1031	1031	100%	0.0	99.99%	510	WP_012733871.1
Methanocaldococcus burtonii	Methanocaldococcus burtonii	1029	1029	100%	0.0	97.84%	510	WP_012669827.1

7

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-02C1W3301N

Job Title: spj058746.1
RID: 02C1W3301N
Program: BLASTP
Database: nr
Query ID: 058746.1
Description: RecName: Full=Archaeal glutamate synthase [NADPH]; Alt=...
Molecule type: amino acid
Query Length: 510

Filter Results
Organism: only top 20 will appear
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity: [] to []
E value: [] to []
Query Coverage: [] to []

Compare these results against the new Clustured nr database

Conserved domains

Distribution of the top 100 Blast Hits on 100 subject sequences

BIOS477/877 L13 - 8

8

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-02C1W3301N

Job Title: spj058746.1
RID: 02C1W3301N
Program: BLASTP
Database: nr
Query ID: 058746.1
Description: RecName: Full=Archaeal glutamate synthase [NADPH]; Alt=...
Molecule type: amino acid
Query Length: 510

Filter Results
Organism: only top 20 will appear
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity: [] to []
E value: [] to []
Query Coverage: [] to []

Compare these results against the new Clustured nr database

Select sequences to be included in the tree

nitronate monoxygenase [Methanoregula boonei]
Sequence ID: WP_012107254.1 Length: 503 Number of Matches: 1
See 1 more title(s) See all Identical Proteins (PSS)

Range 1: 6 to 502

Score	Expect	Method	Identical	Positives	Gaps	
722	ans(1877)	0.0	Compositional matrix adjust.	352/504(81%)	413/504(81%)	7/504(1%)
65	6	YPPKYVVEVDPNOLCERCIECSWGVYREGRRISYNSRCGACRVCRCVMDIIT	65			
65	6	SP++KVE+D++CH+C+RE++S+GVYREGRRISYNS+C+ACRRC+CPDIAI+1	65			
65	6	DLRPFVLEFDGQNRGRCLENSYNSREGVIVNSGKACRRCVCFDPAI+1	65			
12	66	KENATSMRSHPLVQDARDVQDLYNQAKTLLSNGAKKHPIYDFKVLVDCVNPIS	12			
18	126	DLRPFVLEFDGQNRGRCLENSYNSREGVIVNSGKACRRCVCFDPAI+1	18			
18	126	DLRPFVLEFDGQNRGRCLENSYNSREGVIVNSGKACRRCVCFDPAI+1	18			
24	186	GALSNAHLSPFAKAKVECTPTGTEGGGLPKALVYPADHITVAVASRFGVNEVYKGS	24			
24	186	GASLNA++AKA+K+GTH+GTEGGGL++L+YV+DRIH+QVASSRFGV++Y++H+	24			
23	179	GATSLNAQALAKAAKQMGITLFGEGELHSLYVQDMDIVQVAVSRFGVDNYLGERA	23			
36	246	ATEIKIGGAGKPIGGHPLGKVKVIAEISATRIPEGGDASPAHHPIYSDENLAQVRS	36			
36	246	ALTEKIGGAGKPIGGHPLGKVKVIAEISATRIPEGGDASPAHHPIYSDENLAQVRS	36			

BIOS477/877 L13 - 9

9

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-02C1W3301N

Job Title: spj058746.1
RID: 02C1W3301N
Program: BLASTP
Database: nr
Query ID: 058746.1
Description: RecName: Full=Archaeal glutamate synthase [NADPH]; Alt=...
Molecule type: amino acid
Query Length: 510

Filter Results
Organism: only top 20 will appear
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity: [] to []
E value: [] to []
Query Coverage: [] to []

Compare these results against the new Clustured nr database

Select sequences to be included in the tree

nitronate monoxygenase [Methanoregula boonei]
Sequence ID: WP_012107254.1 Length: 503 Number of Matches: 1
See 1 more title(s) See all Identical Proteins (PSS)

Range 1: 6 to 502

Score	Expect	Method	Identical	Positives	Gaps	
722	ans(1877)	0.0	Compositional matrix adjust.	352/504(81%)	413/504(81%)	7/504(1%)
65	6	YPPKYVVEVDPNOLCERCIECSWGVYREGRRISYNSRCGACRVCRCVMDIIT	65			
65	6	SP++KVE+D++CH+C+RE++S+GVYREGRRISYNS+C+ACRRC+CPDIAI+1	65			
65	6	DLRPFVLEFDGQNRGRCLENSYNSREGVIVNSGKACRRCVCFDPAI+1	65			
12	66	KENATSMRSHPLVQDARDVQDLYNQAKTLLSNGAKKHPIYDFKVLVDCVNPIS	12			
18	126	DLRPFVLEFDGQNRGRCLENSYNSREGVIVNSGKACRRCVCFDPAI+1	18			
18	126	DLRPFVLEFDGQNRGRCLENSYNSREGVIVNSGKACRRCVCFDPAI+1	18			
24	186	GALSNAHLSPFAKAKVECTPTGTEGGGLPKALVYPADHITVAVASRFGVNEVYKGS	24			
24	186	GASLNA++AKA+K+GTH+GTEGGGL++L+YV+DRIH+QVASSRFGV++Y++H+	24			
23	179	GATSLNAQALAKAAKQMGITLFGEGELHSLYVQDMDIVQVAVSRFGVDNYLGERA	23			
36	246	ATEIKIGGAGKPIGGHPLGKVKVIAEISATRIPEGGDASPAHHPIYSDENLAQVRS	36			
36	246	ALTEKIGGAGKPIGGHPLGKVKVIAEISATRIPEGGDASPAHHPIYSDENLAQVRS	36			

BIOS477/877 L13 - 10

10

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-02C1W3301N

Job Title: spj058746.1
RID: 02C1W3301N
Program: BLASTP
Database: nr
Query ID: 058746.1
Description: RecName: Full=Archaeal glutamate synthase [NADPH]; Alt=...
Molecule type: amino acid
Query Length: 510

Filter Results
Organism: only top 20 will appear
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity: [] to []
E value: [] to []
Query Coverage: [] to []

Compare these results against the new nr database

Phylogeny based on pairwise distance from BLAST pairwise alignments.

Approximated tree. For a more accurate phylogeny, distances need to be estimated from the multiple alignment.

Download the BLAST result:

- BLAST search result in text format
- Sequences and alignments in FASTA format
- BLAST hit statistics in "Hit Table (csv)" [Can be imported to any spreadsheet program (Excel)]

BIOS477/877 L13 - 11

11

blastp Similarity Search: Result Page

BLAST® = blastp suite = results for RID-02C1W3301N

Job Title: spj058746.1
RID: 02C1W3301N
Program: BLASTP
Database: nr
Query ID: 058746.1
Description: RecName: Full=Archaeal glutamate synthase [NADPH]; Alt=...
Molecule type: amino acid
Query Length: 510

Filter Results
Organism: only top 20 will appear
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity: [] to []
E value: [] to []
Query Coverage: [] to []

Compare these results against the new nr database

Phylogeny based on pairwise distance from BLAST pairwise alignments.

Approximated tree. For a more accurate phylogeny, distances need to be estimated from the multiple alignment.

Download the BLAST result:

- BLAST search result in text format
- Sequences and alignments in FASTA format
- BLAST hit statistics in "Hit Table (csv)" [Can be imported to any spreadsheet program (Excel)]

BIOS477/877 L13 - 12

12

blastp Protein Similarity Search

Algorithm parameters

General Parameters

Max target sequences: 100

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 0.05

Word size: 6

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only Mask lower case letters

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

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

BIOS477/877 L13 - 13

13

blastp Protein Similarity Search

Query: Q58746.1 (AGLUS_METJA)
Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]

Sequences producing significant alignments

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1043	1043	100%	0.0	100.00%	519	WP_033073098.1
glutamate synthase [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1043	1043	100%	0.0	100.00%	519	WP_033073098.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	737	737	98%	0.0	100.00%	369	WP_033073098.1
glutamate synthase [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	737	737	98%	0.0	100.00%	369	WP_033073098.1
glutamate synthase [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	738	738	98%	0.0	99.94%	369	AF220961.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	738	738	98%	0.0	99.14%	369	AF220961.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	738	738	98%	0.0	99.14%	369	AF220961.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	737	737	97%	0.0	97.44%	369	WP_256621572.1

When searching highly conserved sequences:
→ "Max target sequences" number needs to be increased to see more divergent hits. (the default is only 100 sequences!!)

When searching highly conserved proteins, the default setting may not show all the hits.
→ Only the top 100 hits will be listed.
→ To see more divergent sequences, the number of hits needs to be increased.

At the bottom of the 100 hits, E-value is still 0!
There should be a lot more hits with E-values < 0.05 (the threshold)

BIOS477/877 L13 - 14

14

blastp Protein Similarity Search

Algorithm parameters

General Parameters

Max target sequences: 100

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 0.05

Word size: 6

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only Mask lower case letters

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

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

BIOS477/877 L12 - 15

15

blastp Protein Similarity Search

Query: Q58746.1 (AGLUS_METJA)
Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]

Sequences producing significant alignments

select all 5000 sequences selected

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1043	1043	100%	0.0	100.00%	519	WP_033073098.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1043	1043	100%	0.0	100.00%	519	WP_033073098.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	737	737	98%	0.0	100.00%	369	WP_033073098.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	737	737	98%	0.0	100.00%	369	WP_033073098.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	738	738	98%	0.0	99.14%	369	AF220961.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	738	738	98%	0.0	99.14%	369	AF220961.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	738	738	98%	0.0	99.14%	369	AF220961.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	737	737	97%	0.0	97.44%	369	WP_256621572.1

Even after changing "Max target sequences" to 5000: all hits have very small E-values (< 1e-93)

All hits are from bacterial/archaeal proteins

BIOS477/877 L13 - 16

16

blastp Similarity Search: a Case Study

Query: Q58746.1 (AGLUS_METJA)
Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]

• Can we find similar sequences in eukaryotes?

BLAST - blastp results

Standard Protein BLAST

Database: nr

Algorithm: Blastp

Expect threshold: 0.05

Word size: 6

Max matches in a query range: 0

Filters and Masking: Low complexity regions

BLAST Search database nr 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

BIOS477/877 L13 - 17

17

blastp Similarity Search: a Case Study

Query: Q58746.1 (AGLUS_METJA)
Archaeal glutamate synthase [Methanocaldococcus jannaschii DSM 2661]

• Can we find similar sequences in eukaryotes?

BLAST Summary

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 100 Blast Hits on 100 subject sequences

Distribution of the top 100 Blast Hits on 100 subject sequences

[Search limited to Eukaryota]
- Do all eukaryotic proteins lack NapF domain?
- They may not be in the top 100 hits.

[Default search]
- Prokaryotic proteins have both NapF and Glu_synthase domains.

BIOS477/877 L13 - 18

18

blastp Similarity Search: a Case Study

Query: Q58746.1 (A6LUS_METJA)
Archaeal glutamate synthase [*Methanocaldococcus jannaschii* DSM 2661]

Query seq. Specific hits Superfamilies

• Can we find eukaryotic sequences similar to the NapF domain?
- Use "Query subrange" option, or
- Use "Max matches in a query range" option

BIOS477/877 L13 - 19

19

blastp Similarity Search: a Case Study

[Search using the "Query subrange" option]

Enter Query Sequence: Q58746.1

On: uploaded file

Job Title: Q58746.1: Archaeal glutamate synthase

Choose Search Set: Standard databases (nr, etc.)

Algorithm Selection: Basic (protein-protein BLAST)

General Parameters: Max target sequences: 100

E-value increased (for short sequences)

BIOS477/877 L13 - 20

20

blastp Similarity Search: a Case Study

[Search using the "Query subrange" option]

Enter Query Sequence: Q58746.1

On: uploaded file

Job Title: Q58746.1: Archaeal glutamate synthase

Choose Search Set: Standard databases (nr, etc.)

Algorithm Selection: Basic (protein-protein BLAST)

General Parameters: Max target sequences: 100

E-value increased (for short sequences)

BIOS477/877 L13 - 21

21

blastp Similarity Search: a Case Study

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

Enter Query Sequence: Q58746.1

On: uploaded file

Job Title: Q58746.1: Archaeal glutamate synthase

Choose Search Set: Standard databases (nr, etc.)

Algorithm Selection: Basic (protein-protein BLAST)

General Parameters: Max target sequences: 100

E-value increased (for short sequences)

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.

BIOS477/877 L13 - 22

22

blastp Similarity Search: a Case Study

[Max Matches = 3]

Query seq. Specific hits Superfamilies

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments

BIOS477/877 L13 - 23

23

blastp Similarity Search: a Case Study

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

Without Max Matches option

query	subject	% identity	alignment length	mis-matches	gaps	query start	query end	subject start	subject end	e-value	score	bit	% positives
Q58746.1	MCL4138302.1	35.045	448	247	6	78	509	47	466	5.42E-73	248	51.79	
Q58746.1	MCL4141611.1	34.856	416	226	4	85	483	3	390	1.11E-63	222	49.76	
Q58746.1	MCL4141667.1	37.681	345	193	3	160	483	33	376	3.83E-62	218	53.33	
Q58746.1	MCL4115492.1	41.003	339	188	6	167	499	872	1204	3.95E-60	224	54.28	
Q58746.1	MCL4104575.1	40.938	320	180	4	178	491	869	1185	1.11E-59	223	54.37	
Q58746.1	MCL4143097.1	41.009	317	174	4	183	491	217	528	3.10E-59	219	54.26	
Q58746.1	MCL4123473.1	41.956	317	175	4	183	493	125	438	9.24E-59	216	55.84	
Q58746.1	GHP11700.1	40.171	351	179	5	167	491	924	1269	9.99E-59	221	52.14	

With Max Matches = 3

BIOS477/877 L13 - 24

24

Nucleotide Similarity Search

BLAST® = blastn suite

Standard Nucleotide BLAST

Enter Query Sequence: [] Query subrange: []

Or, upload file: []

Choose Search Set:

- Standard databases (nr/nt)
- Experimental databases
- Try experimental taxonomic nt databases

Database: **Nucleotide collection (nr/nt)**

Program Selection:

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

BIOS477/877 L13 - 25

25

BLAST Databases (Nucleotide)

Nucleotide database	Description
Nucleotide Collection (nr/nt) default	Partially non-redundant nucleotide sequences from GenBank, EMBL, DDBJ, PDB, and 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 ribosomal RNA	16S ribosomal RNA (Bacteria and Archaea type strains)
18S ribosomal RNA	18S ribosomal RNA sequences (SSU) from Fungi type and reference material
28S ribosomal RNA	28S ribosomal RNA sequences (LSU) from Fungi type and reference material
Internal transcribed spacer region	Internal transcribed spacer region (ITS) from Fungi type and reference material

and more...

<https://ftp.ncbi.nlm.nih.gov/blast/db/>

BIOS477/877 L13 - 26

26

Nucleotide Similarity Search

BLAST® = blastn suite

Standard Nucleotide BLAST

Enter Query Sequence: [] Query subrange: []

Choose Search Set:

- Standard databases (nr/nt)
- Experimental databases
- Try experimental taxonomic nt databases

Database: **Substudies nt**

Program Selection:

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

BIOS477/877 L13 - 27

27

Nucleotide Similarity Search

BLAST® = blastn suite

Standard Nucleotide BLAST

Enter Query Sequence: [] Query subrange: []

Choose Search Set:

- Standard databases (nr/nt)
- Experimental databases
- Try experimental taxonomic nt databases

Database: **Nucleotide collection (nr/nt)**

Program Selection:

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

BIOS477/877 L13 - 28

28

megablast vs. blastn

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

[blastn: 1028 hits, E<0.0001]

[megablast: only 14 hits, all E=0]

Distribution of the top 13 Blast Hits on 13 subject sequences

megablast:
 for searching closely related sequences (>95%), very fast
 word size = 28 bases (16~256)

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

BIOS477/877 L13 - 29

29

Discontiguous megablast

Program Selection:

- 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 90% or more but is very fast. Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is translated 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

Template type: Coding

Word matching based on discontiguous pattern (template):
 e.g., for coding: 1101101101101 (w=11, t=16)
 → mismatches are allowed for '0' positions

BIOS477/877 L13 - 30

30

BLASTN/BLASTX Results

[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

Score	Expect	Method	Identities	Positives	Gaps
50.0 bits(54)	0.12	45/57(79%)	Q/57(0%)		

Query 1032 TAAAGATGATGAC... 1088
 Sbjct 96 TAAAGATGATGAC... 40

[blastx] (translated query vs. protein db)
hypothetical protein FL83_19625, partial [Caenorhabditis latens]
 Sequence ID: **OZ58681.1** Length: 598 Number of Matches: 1

Range 1: 1 to 598 GenBank Graphics

Score	Expect	Method	Identities	Positives	Gaps
877 bits(2267)	0.0	Compositional matrix adjust.	450/600(75%)	499/600(83%)	32/600(5%)

Query 27 MFHFGHTSPSTSN... 338
 Sbjct 1 MFHFGHTSPSTSN... 338

Query 198 POMPPOIT... 338
 Sbjct 60 POMPPOIT... 338

Query 339 SCV... 494
 Sbjct 120 PSQSPQPP... 179

Query 495 CV... 274
 Sbjct 188 CV... 239

BIOS477/877 L13 - 31

Annotations: Low complexity region is masked (shown in gray lower cases); 6 possible frames.

31

BLASTP Results

[blastp]
glutamate synthase 1 [NADH] chloroplastic isoform X1 (GOSHI) [Polytomella parva]
 Sequence ID: **Q9Y14997.1** Length: 2221 Number of Matches: 1

Range 1: 881 to 1305 GenBank Graphics

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 DAITIKENAT... 120
 Sbjct 881 DALIHSNAP... 936

Query 121 TNPSIDP... 188
 Sbjct 937 RGNDR... 998

Rec-Name: Full-Glutamate synthase large subunit-like protein [Sinorhizobium meliloti 10211]
 Sequence ID: **O87392.2** Length: 442 Number of Matches: 1

Range 1: 11 to 410 GenBank Graphics

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

Query 73 RSHPLVD... 132
 Sbjct 11 RKSAT... 69

Query 133 ELRTYG... 192
 Sbjct 70 DTSV... 154

Query 193 H... 252
 Sbjct 105 KEAL... 164

BIOS477/877 L13 - 32

Annotations: Positives (+): Similar amino acid pairs. These AA pairs have positive scores in the scoring matrix used; Low complexity region is masked.

32

BLASTP Statistics

[blastp]
4Fe-4S dicluster domain-containing protein [Thermoproteota archaeon]
 Sequence ID: **NP84857.1** Length: 716 Number of Matches: 1

Range 1: 231 to 703 GenBank Graphics

Score	Expect	Method	Identities	Positives	Gaps
313 bits(802)	1e-94	Compositional matrix adjust.	187/503(37%)	267/503(53%)	31/503(6%)

Query 8 PKYK... 66
 Sbjct 731 PKYK... 290

Query 67 ENAT... 126
 Sbjct 291 RVAES... 350

Raw Score (S): simply based on pairwise scores & gap penalties
Normalized Score or Bit Score (S_{bit}):
 $S'_{bit} = (\lambda S - \log_e K) / \log_e 2$, $[S'_{nat} = \lambda S - \log_e K]$

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

BIOS477/877 L13 - 33

33

BLASTP Search Summary

BLAST® = blastp suite » results for RID-YBJ4MMX5013

Job Title: spIQ58746.1
 RID: YBJ4MMX5013 expires on 03-05-2011 pm

Click to see the blast search statistics

Filter Results: Organism, Database, Query ID, Description, Molecule type, Query Length

Sequences producing significant alignments

Scientific Name	Max Score	Total Score	Query Cover	E value	Hit	Len	Acc	Accession
glutamate synthase-related protein [Methanosarcina acetivorans]	1043	1043	100%	0.0	100.00%	910	WP	O103086.1

(Top portion of any blast output)

BIOS477/877 L13 - 34

34

BLASTP Search Summary

Search Parameters	Program	blastp
Word size	5	
Expect value	0.05	
Hitlist size	5000	
Gapcosts	11,1	
Matrix	BLOSUM62	
Filter string	F	
Genetic Code	1	
Window Size	40	
Threshold	0	
Composition-based stats	2	

Database	Posted date	Mar 3, 2024 2:30 AM
Number of letters	278,407,168,794	
Number of sequences	722,992,963	
Entrez query	None	

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

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

BIOS477/877 L13 - 35

35

BLASTP Statistics

[blastp]
4Fe-4S dicluster domain-containing protein [Thermoproteota archaeon]
 Sequence ID: **NP84857.1** Length: 716 Number of Matches: 1

Range 1: 231 to 703 GenBank Graphics

Score	Expect	Method	Identities	Positives	Gaps
313 bits(802)	1e-94	Compositional matrix adjust.	187/503(37%)	267/503(53%)	31/503(6%)

Query 8 PKYK... 66
 Sbjct 731 PKYK... 290

Query 67 ENAT... 126
 Sbjct 291 RVAES... 350

Raw Score (S): simply based on pairwise scores & gap penalties
Normalized Score or Bit Score (S_{bit}):
 $S'_{bit} = (\lambda S - \log_e K) / \log_e 2$, $[S'_{nat} = \lambda S - \log_e K]$
 $\lambda = 0.267$, $K = 0.041$, $S'_{bit} = \{0.267 \times 802 - \log_e(0.041)\} / \log_e 2 = 313.5$

λ and K are scoring system specific

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

(for gapped alignments)

BIOS477/877 L13 - 36

36

BLASTP Statistics

[blastp]
 4Fe-4S dicluster domain-containing protein [Thermoproteota archaeon]
 Sequence ID: [NPA84857.1](#) Length: 716 Number of Matches: 1

Range 1: 231 to 703 [GenPept](#) [Graphics](#) Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
313 bits(802)	1e-94	Compositional matrix adjust.	187/503(37%)	267/503(53%)	31/503(6%)

```

Query 8 PKYKVEVDPNRCMLCERCTIECSWGVYRREGDR-IISYSNRCGACHRCVVMCPDAITIK 66
Sbjct 231 PKYAV +E+ C+ C+ E+ GV + +G + + + C+ C+ CP DA+ + +
Query 57 ENAITSWRSHPLWVDARVDIYNQAKTGCLLSGGMNAKEHPYFDKIVLDAQVTNPSID 126
Sbjct 291 RVAESMTNRARIDSSTFNVMROMASIGHPPVIGGAENKVFPSLDLQTLTFPGQTSRPPID 350
  
```

Raw Score (S): simply based on pairwise scores & gap penalties
Normalized Score or Bit Score (S'bit):
 $S'_{bit} = (\lambda S - \log_e K) / \log_e 2$, $[S'_{nat} = \lambda S - \log_e K]$
 $\lambda = 0.267$, $K = 0.041$, $S'_{bit} = \{0.267 \times 802 - \log_e(0.041)\} / \log_e 2 = 313.5$

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

BIOS477/877 L13 - 37

37

Pairwise alignment vs. database searching

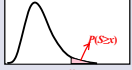
[For a pairwise alignment]
 > **Karlin-Altschul equation** (Karlin & Altschul, 1990)
 $P(S \geq x) = 1 - \exp[-Kmn e^{-\lambda S}] \approx Kmn e^{-\lambda x} \dots$
 Probability of getting the alignment score $S \geq x$ by chance

$e^a \approx 1+a$, where $a = -Kmn e^{-\lambda x}$
 Taylor series approximation
 → works if a is small ($\ll 1$)

[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 probability to have **one pairwise alignment** with $S < x$ by chance
- For 10 alignments, $0.95^{10} \approx 0.60$ is the probability to have **all 10 alignments** with $S < x$
 → $1 - 0.60 = 0.40$ is the probability to have **at least one alignment** with $S \geq x$ by chance
- For 100 alignments, $0.95^{100} \approx 0.006$ is the probability to have **all 100** with $S < x$
 → $1 - 0.006 \approx 0.99$ is the probability to have **at least one alignment** with $S \geq x$ by chance

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



BIOS477/877 L13 - 38

38

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 $\alpha' = 0.05/10 = 0.005$ (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 $\alpha' = 0.05/100 = 0.0005$ (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

BIOS477/877 L13 - 39

39

Bonferroni correction in database searching

> **Multiple comparison correction**
 → Threshold without correction: $P = \alpha$
 → Threshold with correction: $P = \alpha^2 = \alpha/N$ (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

BIOS477/877 L13 - 40


40

BLAST Statistics

> **Karlin-Altschul equation** (Karlin & Altschul, 1990)
 [For a pairwise alignment]
 $P = Kmn e^{-\lambda S}$ (Lec 11 slide 12)
 m, n : lengths of the sequences compared
 → $m \times n$: search space

[For database similarity searching]
 $E = Kmn e^{-\lambda S}$ (used by BLAST instead of $E = N \times P$)
 E-value: the expected number of HSPs with scores $\geq S$
 m : length of the query
 n : length of the database (total number of residues)
 $P = 1 - e^{-E}$ ($P \approx E$ if $E < 0.01$)
 → the probability of having at least one HSP with its score $\geq S$

Used by FASTA



BIOS477/877 L13 - 41

41

BLAST Statistics

> **Karlin-Altschul equation** (Karlin & Altschul, 1990)
 (See also Altschul & Gish, 1996)
 $E = Km' n' e^{-\lambda S}$

m' : effective length of the query
 n' : effective length of the database
 $m' = m - l$
 $n' = n - l \times$ (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.

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

BIOS477/877 L13 - 42

42

P-value, E-value, and database search

[FASTA]

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

[BLAST]

- **E-value** = $Km'n'e^{-\lambda S}$
 - Number of alignments with a score $\geq S$ expected by chance from a database search
 - m' : effective length of the query
 - n' : effective length of the database
- **P-value for a database search**: $P = 1 - e^{-E}$
 - The probability of having at least one HSP with its score $\geq S$

BLAST Statistics: <https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-L1.html>
Altschul et al. (1994)

43

BLASTP Statistics

[blastp]

4Fe-4S dcluster domain-containing protein [Thermoproteota archaeon]
Sequence ID: **NP484857.1** Length: 716 Number of Matches: 1

Range 1: 231 to 703 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
313 bits(802)	1e-94	Compositional matrix adjust.	187/503(37%)	267/503(53%)	31/503(6%)

Query: PKYKVEVDPNRCMLCERCITIECSWGVYRREGDR-IISYNSRCGACHRCVVMCPDRAITIK 66
 Sbjct: 231 PKYVVFVKYDILICGCGTCAMVCPGVIVKMKGKVPVAAAREADICGACMNYCPTDAVKVE 290

Query: 67 ENATSWRSHPLWVDARVDIYNQAKTGCILLSGMNAKEHPYFDKIVLDACQVTPSID 126
 Sbjct: 291 RVAESMTNRARIDSDTFNVMRQMASIGHPPVIGGAENKYPFSLDQLTFIPGQTSRPPID 350

$\lambda = 0.267, K = 0.041, S = 802, S'_{hit} = \{0.267 \times 802 - \ln(0.041)\} / \ln 2 = 313.5$

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

$E = 0.041 \times m' \times n' \times e^{-0.267 \times 802}$ [from the raw score]
 $E = m' \times n' \times 2^{-313}$ [from the bit score]

$m' \times n'$: Effective search space

44

BLASTP Search Summary

Query: **Q58746.1**
Query length: 510 amino acids

m = 510 (length of query)

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

Database	
Posted date	Mar 3, 2024 2:30 AM
Number of letters	278,407,168,794
Number of sequences	722,992,963
Entrez query	None

n: length of database

Karlin-Altschul statistics		
Lambda	0.31987	0.267
K	0.137272	0.041
H	0.476015	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

45

BLASTP Statistics

[blastp]

4Fe-4S dcluster domain-containing protein [Thermoproteota archaeon]
Sequence ID: **NP484857.1** Length: 716 Number of Matches: 1

Range 1: 231 to 703 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
313 bits(802)	1e-94	Compositional matrix adjust.	187/503(37%)	267/503(53%)	31/503(6%)

Query: PKYKVEVDPNRCMLCERCITIECSWGVYRREGDR-IISYNSRCGACHRCVVMCPDRAITIK 66
 Sbjct: 231 PKYVVFVKYDILICGCGTCAMVCPGVIVKMKGKVPVAAAREADICGACMNYCPTDAVKVE 290

Query: 67 ENATSWRSHPLWVDARVDIYNQAKTGCILLSGMNAKEHPYFDKIVLDACQVTPSID 126
 Sbjct: 291 RVAESMTNRARIDSDTFNVMRQMASIGHPPVIGGAENKYPFSLDQLTFIPGQTSRPPID 350

$\lambda = 0.267, K = 0.041, S = 802, S'_{hit} = \{0.267 \times 802 - \ln(0.041)\} / \ln 2 = 313.5$

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

$E = 0.041 \times m' \times n' \times e^{-0.267 \times 802}$ [from the raw score]
 $E = m' \times n' \times 2^{-313}$ [from the bit score]

**$P = 1 - e^{-E}$
 $= 1 - \exp(-5.89 \times 10^{-81})$
 ≈ 0 ($P \approx E$ if $E < 0.01$)**

W/O length adjustment: $m=510, n=278,407,168,794$
 $E = 0.041 \times 510 \times 278,407,168,794 \times e^{-0.267 \times 802} = 5.89E-81 = 5.89 \times 10^{-81}$
 $E = 510 \times 278,407,168,794 \times 2^{-313} = 8.51E-81 = 8.51 \times 10^{-81}$ ($> E-94$)

(Without length adjustment, E-values are overestimated)

46

BLASTN Search Summary & Statistics

Query: **DQ018115.1**
Query length: 1533 bp

m = 1533

Search Parameters	
Program	blastn
Word size	11
Expect value	0.05
Hitlist size	5000
Match/Mismatch scores	2,-3
Gaps	5,2
Low Complexity Filter	Yes
Filter string	Lm
Genetic Code	1

Database	
Posted date	Mar 1, 2024 12:40 PM
Number of letters	1,574,905,710,618
Number of sequences	103,965,835
Entrez query	None

n: Length of database

N: Number of sequences

Karlin-Altschul statistics for gapped alignment		
Lambda	0.63371	0.625
K	0.408146	0.41
H	0.912438	0.78

Results Statistics

Length adjustment: 42
 Effective length of query: 1491
 Effective length of database: 1570539145548
 Effective search space: 2341673866012068
 Effective search space used: 2341673866012068

$l = 42$
 $m' = m - l = 1533 - 42 = 1491$
 $n' = n - l \times N = 1,574,905,710,618 - 42 \times 103,965,835 = 1,570,539,145,548$

$m' \times n'$

47

BLASTN Search Summary & Statistics

Query: **DQ018115.1**
Query length: 1533 bp

m = 1533

Search Parameters	
Program	blastn
Word size	11
Expect value	0.05
Hitlist size	5000
Match/Mismatch scores	2,-3
Gaps	5,2
Low Complexity Filter	Yes
Filter string	Lm
Genetic Code	1

Database	
Posted date	Mar 1, 2024 12:40 PM
Number of letters	1,574,905,710,618
Number of sequences	103,965,835
Entrez query	None

n: Length of database

N: Number of sequences

Karlin-Altschul statistics for gapped alignment		
Lambda	0.63371	0.625
K	0.408146	0.41
H	0.912438	0.78

Results Statistics

Length adjustment: 42
 Effective length of query: 1491
 Effective length of database: 1570539145548
 Effective search space: 2341673866012068
 Effective search space used: 2341673866012068

$S = 255, \lambda = 0.625, K = 0.41$

$S'_{hit} = (\lambda S - \log_2 K) / \log_2 2$
 $= \{0.625 \times 255 - \ln(0.41)\} / \ln 2$
 $= 231$

$E = Km'n'e^{-\lambda S}$
 $= 0.41 \times 2,341,673,866,012,068$
 $\times e^{-0.625 \times 255}$
 $= 5.84e-55 = 5.8 \times 10^{-55} \approx 1 \times 10^{-54}$

$E = m'n'2^{-S'_{hit}}$
 $= 2,341,673,866,012,068 \times 2^{-231}$
 $= 6.8e-55 = 6.8 \times 10^{-55} \approx 1 \times 10^{-54}$

48

BLAST Search Set vs. Format Option

[Limit the search space BEFORE starting the search]

**Using limited search set reduces the search space:
 $E = Km'n'e^{-AS}$
→ E-values become smaller**

BIOS477/877 L13 - 49

49

BLAST Search Set vs. Format Option

[Limit the search result AFTER the search]

**Filtering the results does not affect the search space
→ E-values are not affected**

BIOS477/877 L13 - 50

50

BLAST Search Set vs. Format Option

Limiting the search for "Sulfolobus" sequences

Search space is reduced

BIOS477/877 L13 - 51

51

BLAST Search Set vs. Format Option

After the search, results are filtered for "Sulfolobus" sequences

Search space is NOT affected

BIOS477/877 L13 - 52

52

BLAST Search Set vs. Format Option

Search is NOT limited; results are filtered Search is limited

(Database size is ~20,000 times larger)

Score	Query cov	E-value	% ident	Score	Query cov	E-value	% ident
432	99%	2e-143	44.44%	432	99%	9e-148	44.44%
60.1	26%	7e-06	28.29%	60.1	26%	4e-10	28.29%

(E-values are ~10⁴ times larger)

**$E = Km'n'e^{-AS}$
E-value is affected by the database size!**

BIOS477/877 L13 - 53

53

BLASTP Search Summary

Query: Q58746.1
Query length: 510 amino acids

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

BIOS477/877 L13 - 54

54

FASTA

https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
(includes also SSEARCH)

<https://www.ebi.ac.uk/jdispatcher/sss/fasta>
(includes also SSEARCH)
With graphic output
Results can be obtained through email

<https://www.genome.jp/tools/fasta/>
(search can be done against KEGG databases)

BIOS477/877 L13 - 55

55

FASTA Similarity Search

This page provides search against comprehensive databases, like SwissProt and NCBI RefSeq. The FASTA Similarity Search can be used for email, anonymous searches. The NCBI database is the best first choice for searching in a genome database from a closely related organism (e.g. *Rabies* genome for variations).

The Individual Proteomes/Genomes page provides searches against selected proteomes.

FASTA Program information

Retrieve result HTML: Show recent searches

Choose: (A) Program, (B) Query (sequences/accidions), (C) Database and (D) Start Search:

(A) Program:

(B) Query:

Or upload query from file:

(C) Database:

(D) Start Search:

Annotations: include domain annotations Exclude low complexity (seg)

Comments (optional):

Other search options: Show Histogram

Search matrix: open ext: Klupp: Statistical estimates: E: Bit:

Alignment Options: Highlight similarities differences compact differences. Output format:

https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

BIOS477/877 L13 - 56

56

FASTA Statistics

Query: TMP-9
151-161 (GPFHGLSKAP2_XENLA Src kinase-associated phosphoprotein 2 - 328 aa Library: SwissProt (Uniprot) 20938021 residues in 565254 sequences)

<= 40	961	E11
41	152	E11
42	15	E11
43	16	E11
44	17	E11
45	18	E11
46	19	E11
47	20	E11
48	21	E11
49	22	E11
50	23	E11
51	24	E11
52	25	E11
53	26	E11
54	27	E11
55	28	E11
56	29	E11
57	30	E11
58	31	E11
59	32	E11
60	33	E11
61	34	E11
62	35	E11
63	36	E11
64	37	E11
65	38	E11
66	39	E11
67	40	E11
68	41	E11
69	42	E11
70	43	E11
71	44	E11
72	45	E11
73	46	E11
74	47	E11
75	48	E11
76	49	E11
77	50	E11
78	51	E11
79	52	E11
80	53	E11
81	54	E11
82	55	E11
83	56	E11
84	57	E11
85	58	E11
86	59	E11
87	60	E11
88	61	E11
89	62	E11
90	63	E11
91	64	E11
92	65	E11
93	66	E11
94	67	E11
95	68	E11
96	69	E11
97	70	E11
98	71	E11
99	72	E11
100	73	E11
101	74	E11
102	75	E11
103	76	E11
104	77	E11
105	78	E11
106	79	E11
107	80	E11
108	81	E11
109	82	E11
110	83	E11
111	84	E11
112	85	E11
113	86	E11
114	87	E11
115	88	E11
116	89	E11
117	90	E11
118	91	E11
119	92	E11
120	93	E11
121	94	E11
122	95	E11
123	96	E11
124	97	E11
125	98	E11
126	99	E11
127	100	E11
128	101	E11
129	102	E11
130	103	E11
131	104	E11
132	105	E11
133	106	E11
134	107	E11
135	108	E11
136	109	E11
137	110	E11
138	111	E11
139	112	E11
140	113	E11
141	114	E11
142	115	E11
143	116	E11
144	117	E11
145	118	E11
146	119	E11
147	120	E11

Statistics: Expectation_m[E]: 7.7578/-0.000147; num: 5.02886/-0.000
mean_var=4.70884/-12.2253; E%: 479.2/-115.01; E2: 0-trim: 131 in 2/62
Lambda= 0.159356
Statistics sampled from 60000 (62842) to 60602 sequences
Kolmogorov-Smirnov statistic: 0.0137 (N=19) at 52
Algorithm: FASTA (3.0 Nov 2011) (optimized)

Parameters: BLSB matrix (151-51x5), open/ext: -10/-2
Klupp: 2, E-point: 1 (0.429), E-opt: 0.2 (0.124), width: 16
Scan time: 11.510

The best scores are:

sp G5F9W1 SKAP2_XENLA Src kinase-associated phosphoprot	(328)	213	494.7	7.5e-129	1.000	328	all	all
sp G5K6P7 SKA2A_XENLA Src kinase-associated phosphoprot	(330)	1961	459.7	2.5e-120	0.927	0.967	330	all
sp G5U9P4 SKAP2_XENLA Src kinase-associated phosphoprot	(330)	1786	419.5	3.3e-110	0.986	0.952	330	all
sp G1K8Z1 SKAP2_TAKU1 Src kinase-associated phosphoprot	(329)	1891	259.7	4.1e-68	0.549	0.772	337	all
sp G75863 SKAP2_HUMAN Src kinase-associated phosphoprot	(359)	931	222.6	5.2e-57	0.611	0.887	347	all
sp G521P7 SKAP2_BOVIN Src kinase-associated phosphoprot	(358)	930	222.7	6.7e-57	0.595	0.821	346	all
sp G3U08 SKAP2_MOUSE Src kinase-associated phosphoprot	(358)	915	219.2	6.0e-56	0.604	0.812	346	all

NOTE: FASTA calculates P-value for each pairwise alignment

BIOS477/877 L13 - 57

57

FASTA Statistics

Observed distribution
Expected distribution

Smith-Waterman score
Query: human glucose transporter
Against: SwissProt (~84,000 seq)

Distribution based on database sequences can be used to simulate extreme value distribution

$P(S \geq x) = 1 - \exp[-Kmn e^{-\lambda x}]$
or
 $P(S \geq x) = mn 2^{-x}$
where m : location parameter β (or $1/\lambda$); scale parameter for Gumbel EVD

$S^*_{hit} = (\lambda S - \ln K) / \ln 2$

Pearson (1998, 2013)

BIOS477/877 L13 - 58

58

FASTA Statistics

$P(S \geq x) = 1 - \exp[-Kmn e^{-\lambda x}]$
or
 $P(S \geq x) = mn 2^{-x}$
where m and n : the lengths of query and each hit sequence

$E(565254) = P \times 565254$

NOTE: FASTA calculates P-value for each pairwise alignment

BIOS477/877 L13 - 59

59

FASTA Statistics

Optimum raw score, standardized score, Bit score, E-value

Region: 284-277740-278 score=206 bits=42.1; Id=0.521; Q=0.988
Region: 278-283740-283 score=254 bits=53.8; Id=0.816; Q=0.988
Initial: 1869 initial: 590; num: 79; Z-score: 993.0; bits: 192; E(565254): 9.4e-48

Smith-Waterman score: 1207.5476 identity: 19.638 similarity: 543 aa overlap: 12613-339

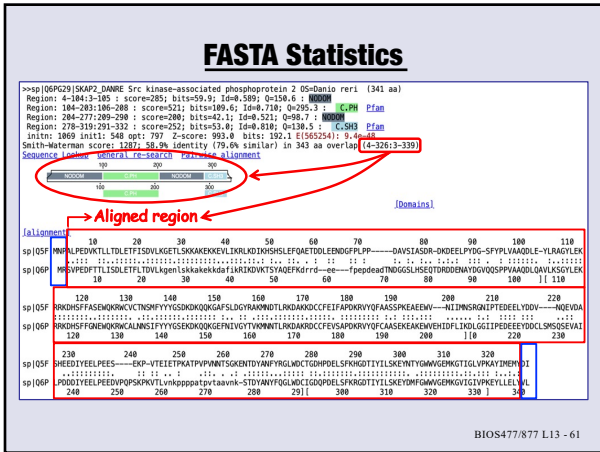
Sequence Lookup General nt-search Pairwise alignment

$P(S \geq x) = 1 - \exp[-e^{-\lambda(x-\beta)}]$
or
 $P(S \geq x) = mn 2^{-x}$
 $E = P \times 565254 = 9.40 \times 10^{-48}$

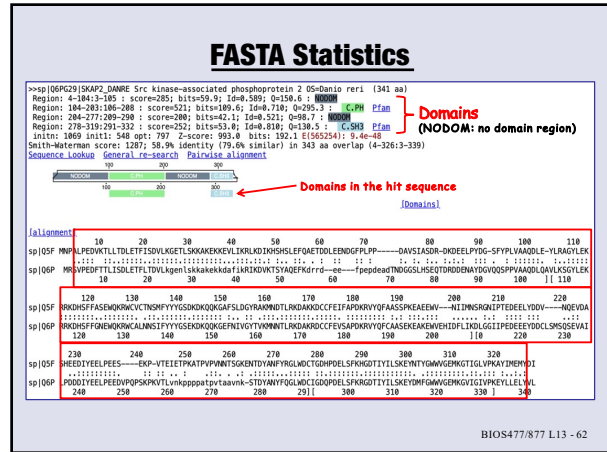
NOTE: FASTA calculates P-value for each pairwise alignment

BIOS477/877 L13 - 60

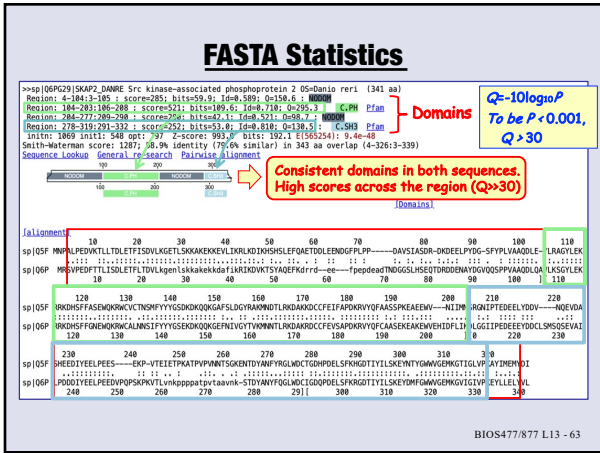
60



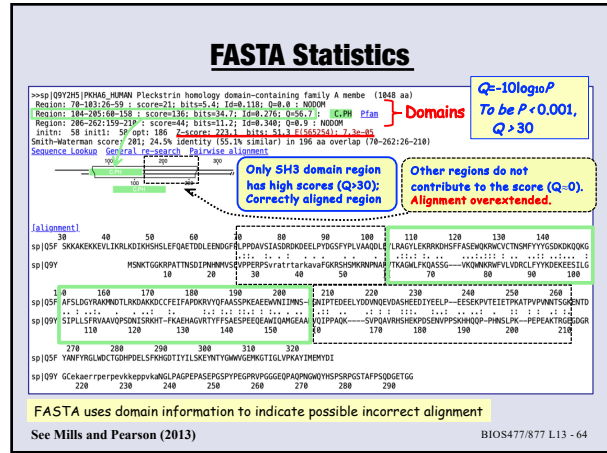
61



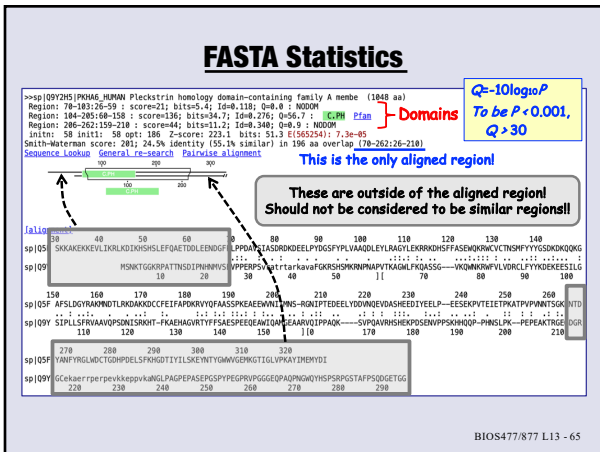
62



63



64



65