

Spring 2024
BIOS 477/877
Bioinformatics and Molecular Evolution
Lecture 12

BIOS477/877 L12 - 1

1

TODAY'S TOPICS

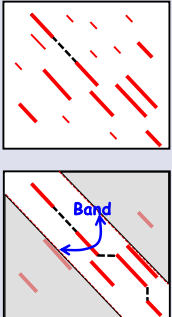
- Assignment 4 Review
- Similarity Search
 - FASTA & BLAST algorithm
 - BLAST website and options
- Assignment 6

BIOS477/877 L12 - 2

2

FASTA Algorithm

1. Find identities using k-tuples
2. Join diagonals without gaps
3. Choose top 10 diagonals using a scoring matrix (e.g., BLOSUM62)
init1: the top diagonal score
4. Join again with gaps
initn: score of the longer diagonal
5. A diagonal band is defined
(width: 32 if k=1, 16 if k=2 for protein)
6. Find optimal local alignment using dynamic programming algorithm within the band
opt: the final score



BIOS477/877 L12 - 3

3

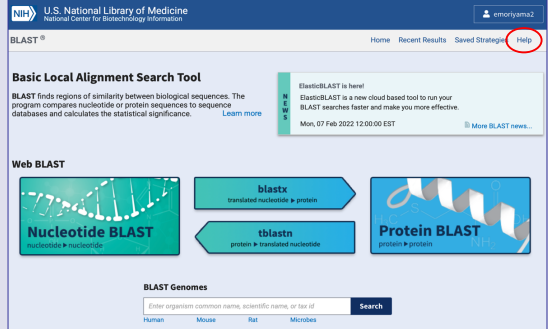
FASTA Algorithm

- Ranking
 - Database sequences are ranked based on z-values or OPT scores
 - z-value = the number of standard deviations from the mean (standardized score)
 - The high scored sequence pairs are aligned using the full Smith-Waterman dynamic programming algorithm (not just from the diagonal band; better alignment)
- FASTA/SSEARCH website
 - https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
 - <https://www.cbi.ac.uk/jdispatcher/sss/fasta>
 - [FASTA guide] https://fasta.bioch.virginia.edu/wrp_fasta/fasta_guide.pdf
 - [William R. Pearson's website] <https://fasta.bioch.virginia.edu/wrpearson/>

BIOS477/877 L12 - 4

4

BLAST Similarity Search



BIOS477/877 L12 - 5

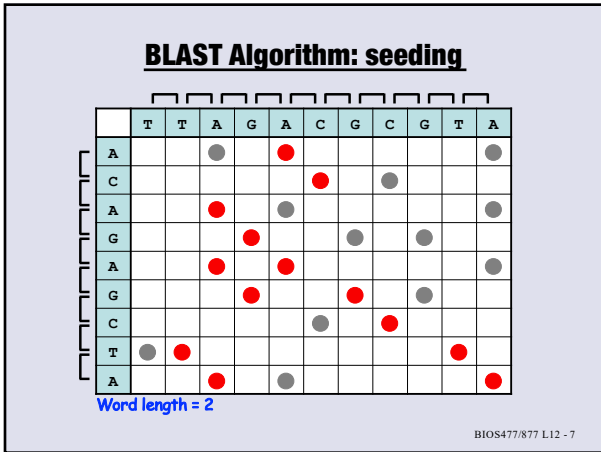
5

BLAST resources

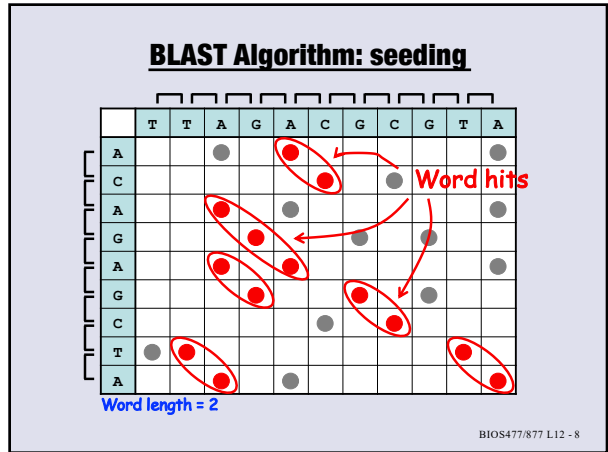
- BLAST
 - <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
 - [Guide to BLAST home and search pages] ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf (Also available on Canvas)
 - [BLAST Report Description] https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf (Also available on Canvas)
 - [BLAST Statistics] <https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>
 - [BLAST Command Line User Manual] <https://www.ncbi.nlm.nih.gov/books/NBK279690/>
 - [BLAST YouTube Tutorials] (Link is available from NCBI Help page or from Canvas)

BIOS477/877 L12 - 6

6



7



8

BLAST Algorithm: seeding

- Using **words** reduces the search space
- **Neighborhood** increases the sensitivity

Match = 2
 Mismatch (Ts) = -1
 Mismatch (Tv) = -5

$TC_{TC} = 2 + 2 = 4$

$TC_{TA} = 2 - 5 = -3$
 $TC_{TG} = 2 - 5 = -3$
 $TC_{AC} = -5 + 2 = -3$
 $TC_{AT} = -5 - 1 = -6$

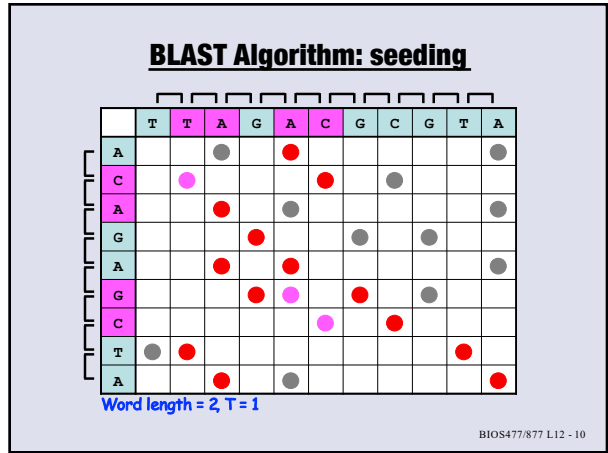
$TC_{TT} = 2 - 1 = 1$
 $TC_{CC} = -1 + 2 = 1$
 ...

Neighborhood

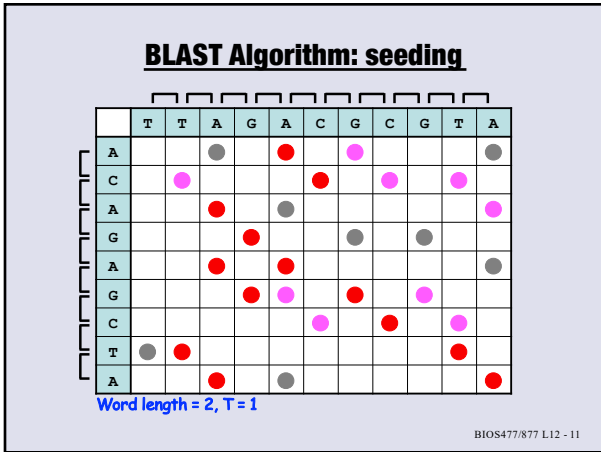
Neighborhood Threshold (T) = 1
 [minimum score allowed to be the neighborhood]

BIOS477/877 L12 - 9

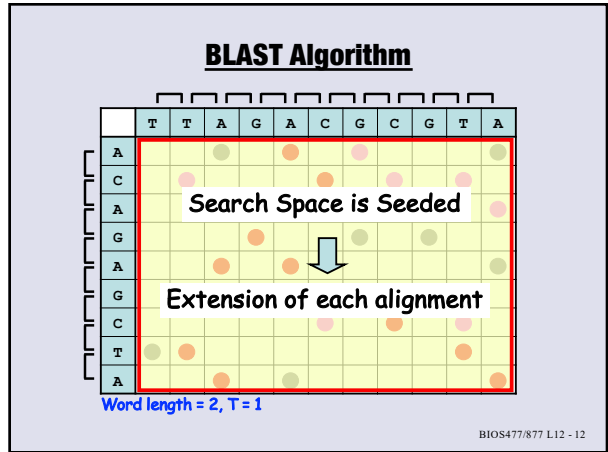
9



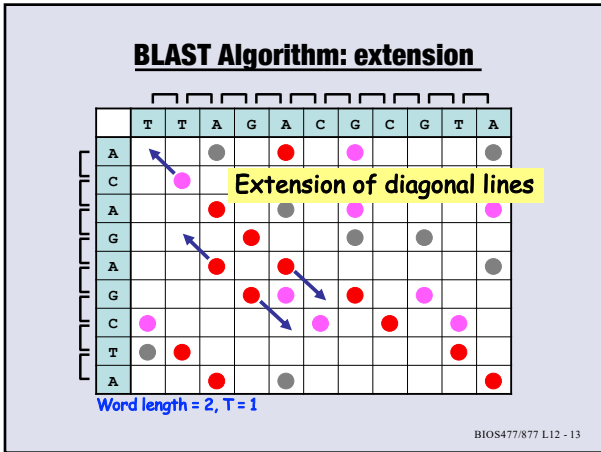
10



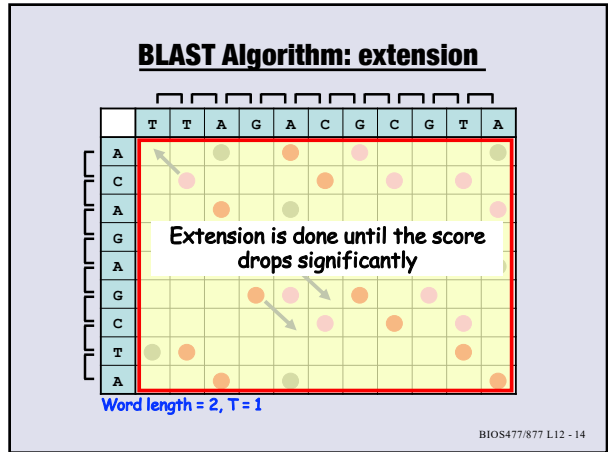
11



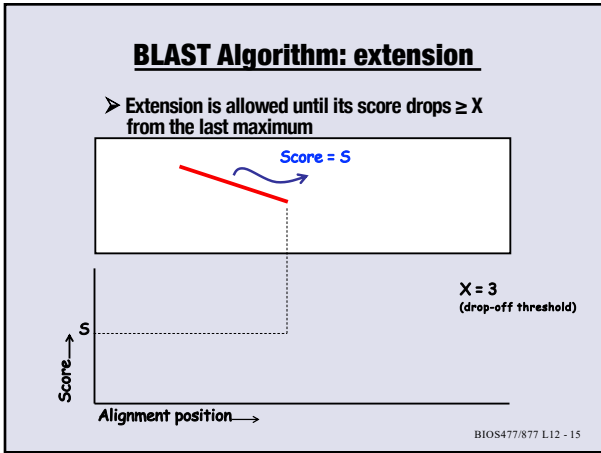
12



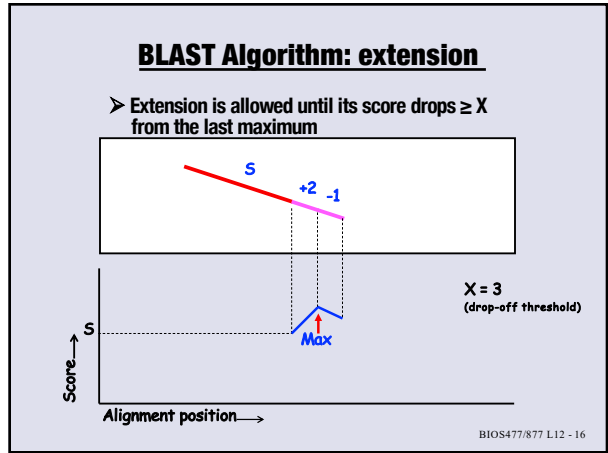
13



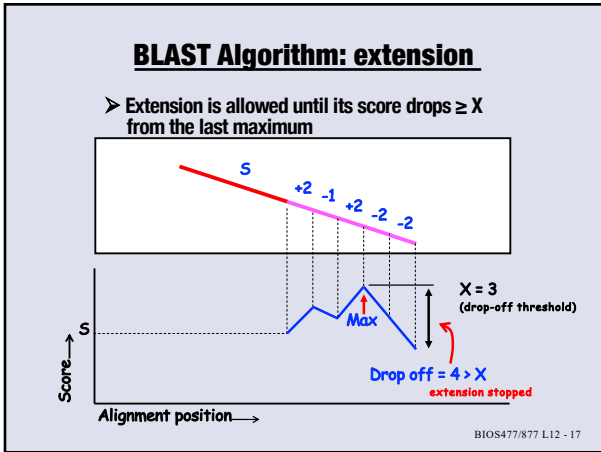
14



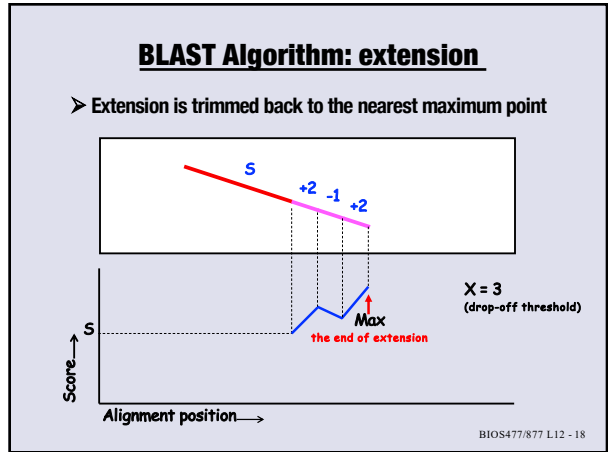
15



16



17



18

BLAST Algorithm: extension

➤ **Two-hit method:** Extension is invoked only when two hits are found within distance A and on the same diagonal

W=3, T=11, X=16, A=40

BIOS477/877 L12 - 19

19

BLAST Algorithm: evaluation

➤ **Alignment threshold (S1)** is used to choose only HSP (High-scoring Segment Pairs)

W=3, T=11, X=16, A=40, S1=41

BIOS477/877 L12 - 20

20

BLAST Algorithm: evaluation

➤ **Consistent alignments** are grouped for E-value calculation

W=3, T=11, X=16, A=40, S1=41

BIOS477/877 L12 - 21

21

BLAST Algorithm: Gapped extension

➤ **Gapped extension** is triggered after high score ungapped alignments are found

W=3, T=11, X=16, A=40, S1=41

BIOS477/877 L12 - 22

22

BLAST Algorithm: Gapped extension

➤ Another **threshold for gapped alignment (S2)** is used to choose the final set of HSPs

W=3, T=11, X=16, A=40, S1=41, S2=70

BIOS477/877 L12 - 23

23

BLAST Algorithm

- **Word-matching size (W)**
 - longer words: faster but less sensitive
- **Neighborhood threshold (T)**
 - lower T: detects weaker similarities
 - slower but more sensitive
- **Extension**
 - Drop-off score (X)
 - Two-hit method (A: distance b/w 2 hits)
- **HSP selection**
 - (ungapped alignment threshold: S1)
- **Gapped HSP extension**
 - (gapped alignment threshold: S2)

Word = 2

T=1

TC vs. = 4

TC = 1

CA = 1

CT = -2

TA = -3

TG = -3

BIOS477/877 L12 - 24

24

BLAST Similarity Search

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.

Web BLAST

Nucleotide BLAST
nucleotide → nucleotide

blastx
translated nucleotide → protein

tblastn
protein → translated nucleotide

Protein BLAST
protein → protein

BLAST Genomes

Enter organism common name, scientific name, or tax ID
Human Mouse Rat Microbes

<https://blast.ncbi.nlm.nih.gov/Blast.cgi> BIOS477/877 L12 - 25

25

blastp Protein Similarity Search

blastp Standard Protein BLAST

Enter Query Sequence

Click on "more" to see HowTo BLAST Guide

Choose Search Set

Database: Standard databases (nr, etc.) Experimental databases

Standard Database: Non-redundant protein sequences (nr)

Program Selection: Quick BLASTP (Accelerated protein-protein BLAST)

<https://blast.ncbi.nlm.nih.gov/Blast.cgi> BIOS477/877 L12 - 26

26

blastp Protein Similarity Search

blastp Standard Protein BLAST

Enter Query Sequence

Query sequence(s) in FASTA format or bare sequence(s) or accession number(s) (e.g., P01013) or gi number(s) (e.g., 129295) *Multiple sequences can be searched

Default database is NR

Database: Standard databases (nr, etc.) Experimental databases

Standard Database: Non-redundant protein sequences (nr)

Program Selection: Quick BLASTP (Accelerated protein-protein BLAST)

<https://blast.ncbi.nlm.nih.gov/Blast.cgi> BIOS477/877 L12 - 27

27

BLAST Databases (Protein)

Protein database	Description
Non-redundant (nr) default	Non-redundant protein sequences from GenPept, Swissprot, PIR, PDF, PDB, and RefSeq Select
RefSeq Select	NCBI RefSeq protein sequences from human, mouse, and prokaryotes
RefSeq proteins	NCBI Protein Reference Sequences
Model Organisms (landmark)	Proteome of 27 model organisms spanning a wide taxonomic range
UniProtKB/Swiss-Prot	Non-redundant UniProtKB/SwissProt sequences
Patented	Protein sequences derived from the Patent division of GenBank
Protein Data Bank (PDB)	Sequences from the Protein Data Bank
Metagenomic proteins	Proteins from WGS metagenomic projects
Transcriptome Shotgun Assembly proteins	CDS features on mRNA sequences in the Transcriptome Shotgun Assembly sequences

RefSeq: A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic DNA, transcripts, and proteins.
<https://www.ncbi.nlm.nih.gov/refseq/>

Landmark database: used by SmartBLAST
<https://ftp.ncbi.nlm.nih.gov/blast/dl/>

<https://blast.ncbi.nlm.nih.gov/Blast.cgi> BIOS477/877 L12 - 28

28

BLAST Landmark Database (Protein)

#	Organism	Assembly	Superkingdom
1	Bacillus subtilis	GDCE_0000030945.1	Bacteria
2	Danio rerio, GRCz11	GDCE_0000030985.1	Bacteria
3	Escherichia coli, str. K-12, subsp. MG1655	GDCE_0000030945.2	Bacteria
4	Micrococcus aeruginosus, NIES-843	GDCE_0000110625.1	Bacteria
5	Mycobacterium tuberculosis, H37Rv	GDCE_0001105935.2	Bacteria
6	Naissolus maritimus	GDCE_0000030945.1	Bacteria
7	Pasteurella multocida, 630	GDCE_0000030945.1	Bacteria
8	Pseudomonas aeruginosa	GDCE_0000030945.1	Bacteria
9	Stenotrophomonas maltophilia, MS1	GDCE_0001105935.2	Bacteria
10	Streptococcus pneumoniae, F55	GDCE_0000030945.1	Bacteria
11	Stenotrophomonas maltophilia	GDCE_0000030945.1	Bacteria
12	Stenotrophomonas, str. FGSG_1148	GDCE_0000030945.1	Bacteria
13	Thermotoga maritima, MS88	GDCE_0000030945.1	Bacteria
14	Methanobrevibacter smithii, ATCC 35061	GDCE_0000030945.1	Archaea
15	Staphylococcus aureus, DSMZ 6258	GDCE_0000030945.1	Archaea
16	Arabidopsis thaliana, (Drake, erosa)	GDCE_000001735.3	Eukaryota
17	Gemmatimonas viscum	GDCE_000002095.5	Eukaryota
18	Danio rerio, GRCz11	GDCE_000002095.5	Eukaryota
19	Drosophila melanogaster, A34	GDCE_000004085.1	Eukaryota
20	Drosophila melanogaster, (ruis, fv)	GDCE_000004085.1	Eukaryota
21	Caenorhabditis elegans	GDCE_000004085.1	Eukaryota
22	Homo sapiens, (human)	GDCE_000001405.30	Eukaryota
23	Leishmania major	GDCE_0000112285.1	Eukaryota
24	Mus musculus, (house mouse)	GDCE_0000011630.04	Eukaryota
25	Plasmodium falciparum, 3D7	GDCE_000002765.31	Eukaryota
26	Saccharomyces cerevisiae, (baker's yeast)	GDCE_000114045.2	Eukaryota
27	Schistosoma mansoni, (blood fluke, yema)	GDCE_000002095.5	Eukaryota

Proteomes from 27 representative genomes: used for a quick search across all the kingdoms

<https://blast.ncbi.nlm.nih.gov/Blast.cgi> BIOS477/877 L12 - 29

29

blastp Protein Similarity Search

blastp Standard Protein BLAST

Enter Query Sequence

Search can be limited against (or excluding) specific organism(s). But remember how it affects the database size and E-values.

Choose Search Set

Database: Standard databases (nr, etc.) Experimental databases

Standard Database: Non-redundant protein sequences (nr)

Organism: (Use mouse (tag:1000))

Program Selection: Quick BLASTP (Accelerated protein-protein BLAST)

<https://blast.ncbi.nlm.nih.gov/Blast.cgi> BIOS477/877 L12 - 30

30

blastp Protein Similarity Search

[Refseq Acc#]
 XM_ (for mRNA), XR_ (for non-coding RNA), XP_ (for protein)
 → model (prediction) entries from genome annotations
 NM_ (for mRNA), NR_ (for non-coding RNA), NP_ (for protein)
 → curated RefSeq records (high quality)

WP_ (for non-redundant RefSeq proteins)
 → a single non-redundant entry represents identical prokaryotic proteins (no specific taxonomical information)

BIOS477/877 L12 - 31

31

blastp Protein Similarity Search

Click to change the algorithm parameters

BIOS477/877 L12 - 32

32

blastp Protein Similarity Search

Algorithm parameters

General Parameters

- Max target sequences: 100
- Short queries: Automatically adjust parameters for short input seq
- Expect threshold: 0.05 → Expect (E) value
- 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

Annotations:

- Number of hit sequences to be shown in the output. For highly conserved sequences, this number needs to be increased to see more divergent hits.
- Statistical significance threshold for reporting matches. Lower E-value → more stringent
- Word size (shorter → more sensitive search). Limit the number of matches to a query range. Useful for finding domains.
- Scoring matrix
- Gap penalty: Existence (opening) & Extension

BIOS477/877 L12 - 33

33

blastp Protein Similarity Search

Algorithm parameters

General Parameters

- Max target sequences: 100
- Short queries: Automatic
- 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 complex
- Mask: Mask for lookup table only

Annotations:

- Only limited sets of combinations are available (BLAST uses a lookup table for K and λ)
- Gap penalty combinations are different depending on the scoring matrix

BIOS477/877 L12 - 34

34

Significance of Alignment Scores

→ K and λ → from Altschul & Gish (1996) table.

a	b	A	K	H (ends)	a	b	A	K	H (ends)
0-6	0.225	0.11	0.34		11	6-11	0.107	0.05	0.21
16	4-16	0.225	0.08	0.31	11	6-7	0.100	0.04	0.19
16	3	0.213	0.06	0.27	11	5	0.104	0.04	0.17
16	2	0.207	0.05	0.24	11	4	0.177	0.031	0.15
16	1	0.189	0.024	0.15	11	2	0.130	0.009	0.06
15	8-15	0.222	0.09	0.31	10	8-10	0.183	0.04	0.17
15	6-7	0.219	0.08	0.29	10	6-7	0.178	0.035	0.16
15	4-5	0.216	0.07	0.26	10	5	0.168	0.026	0.13
15	3	0.210	0.06	0.25	10	4	0.156	0.020	0.10
15	2	0.202	0.05	0.22	10	3	0.139	0.013	0.07
15	1	0.166	0.018	0.11	10	2	0.099	0.007	0.03
14	8-14	0.218	0.08	0.29	10	1	0.099	Linear	
14	5-7	0.214	0.07	0.27	9	7-9	0.164	0.029	0.13
14	4	0.205	0.05	0.24	9	5-6	0.152	0.021	0.10
14	3	0.201	0.05	0.22	9	4	0.134	0.014	0.07
14	2	0.188	0.034	0.17	9	3	0.107	0.006	0.04
14	1	0.140	0.009	0.07	9	1-2	Linear		
13	8-13	0.211	0.06	0.27	8	8	0.139	0.017	0.08
13	5-7	0.205	0.05	0.24	8	7	0.124	0.015	0.07
13	4	0.202	0.05	0.22	8	6	0.127	0.013	0.06
13	3	0.198	0.034	0.18	8	5	0.117	0.011	0.05
13	2	0.174	0.025	0.13	8	4	0.101	0.009	0.03
13	1	0.134	0.006	0.04	8	1-3	Borderline or linear		
12	7-12	0.205	0.06	0.24	7	7	0.100	0.010	0.04
12	5-6	0.197	0.05	0.21	7	6	0.094	0.010	0.03
12	4	0.192	0.04	0.18	7	5	0.094	0.010	0.03
12	3	0.178	0.028	0.15	7	1-5	Borderline or linear		
12	2	0.158	0.019	0.10					
12	1				1-6	1-6	Linear		

Annotations:

- For BLOSUM50: a : gap opening penalty, b : gap extension penalty (affine gap penalty)
- Based on 10,000 random sequence pairs, → calculated optimal local alignment scores, S .
- K and λ are estimated by fitting the distribution of S with $P(S \geq x) \approx Kmne^{-\lambda x}$
- This is what BLAST does!

BIOS477/877 L12 - 35

35

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

Annotations:

- Scoring matrix is adjusted based on amino acid composition, yielding more accurate E-values
- To mask off segments of low compositional complexity

BIOS477/877 L12 - 36

36

blastp Protein Similarity Search

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

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

Job Title: spQ58746.1
 RID: 02C1W3301N
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1
 Description: RecName: Full-Archaeal glutamate synthase [NADPH]; Al...

Q58746.1 → **Accession number as an input**

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

37

blastp Similarity Search: Result Page

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

Job Title: spQ58746.1
 RID: 02C1W3301N
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1
 Description: RecName: Full-Archaeal glutamate synthase [NADPH]; Al...

How to read this report? →

How to read BLAST output (available also from Canvas/References page)

Sequences producing significant alignments

Description	Scientific Name	Max Score	Total Score	Query Cov	E value	Per Idnt	Acc Len	Accession
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1043	1043	100%	0.0	100.00%	910	WP_010239869.1
glutamate synthase-related protein [Methanocaldococcus sp. F3405-22]	Methanocaldococcus sp. F3405-22	1032	1032	100%	0.0	98.43%	910	WP_012073860.1
glutamate synthase-related protein [Methanocaldococcus vulnificus]	Methanocaldococcus vulnificus	1031	1031	100%	0.0	98.43%	910	WP_012723273.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1029	1029	100%	0.0	97.84%	910	WP_012469262.1

38

blastp Similarity Search: Result Page

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

Job Title: spQ58746.1
 RID: 02C1W3301N
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1
 Description: RecName: Full-Archaeal glutamate syn...

Title: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule Type: Protein
Update date: 2023/01/12
Number of sequences: 522605238

Sequences producing significant alignments

Description	Scientific Name	Max Score	Total Score	Query Cov	E value	Per Idnt	Acc Len	Accession
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1043	1043	100%	0.0	100.00%	910	WP_010239869.1
glutamate synthase-related protein [Methanocaldococcus sp. F3405-22]	Methanocaldococcus sp. F3405-22	1032	1032	100%	0.0	98.43%	910	WP_012073860.1
glutamate synthase-related protein [Methanocaldococcus vulnificus]	Methanocaldococcus vulnificus	1031	1031	100%	0.0	98.43%	910	WP_012723273.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1029	1029	100%	0.0	97.84%	910	WP_012469262.1

39

blastp Similarity Search: Result Page

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

Job Title: spQ58746.1
 RID: 02C1W3301N
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1
 Description: RecName: Full-Archaeal glutamate synthase [NADPH]; Al...

Query name
Query length

Sequences producing significant alignments

Description	Scientific Name	Max Score	Total Score	Query Cov	E value	Per Idnt	Acc Len	Accession
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1043	1043	100%	0.0	100.00%	910	WP_010239869.1
glutamate synthase-related protein [Methanocaldococcus sp. F3405-22]	Methanocaldococcus sp. F3405-22	1032	1032	100%	0.0	98.43%	910	WP_012073860.1
glutamate synthase-related protein [Methanocaldococcus vulnificus]	Methanocaldococcus vulnificus	1031	1031	100%	0.0	98.43%	910	WP_012723273.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1029	1029	100%	0.0	97.84%	910	WP_012469262.1

40

blastp Similarity Search: Result Page

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

Job Title: spQ58746.1
 RID: 02C1W3301N
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1
 Description: RecName: Full-Archaeal glutamate synthase [NADPH]; Al...

Query coverage: Proportion of the query aligned

Bit scores **E-value**

Sequences producing significant alignments

Description	Scientific Name	Max Score	Total Score	Query Cov	E value	Per Idnt	Acc Len	Accession
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1043	1043	100%	0.0	100.00%	910	WP_010239869.1
glutamate synthase-related protein [Methanocaldococcus sp. F3405-22]	Methanocaldococcus sp. F3405-22	1032	1032	100%	0.0	98.43%	910	WP_012073860.1
glutamate synthase-related protein [Methanocaldococcus vulnificus]	Methanocaldococcus vulnificus	1031	1031	100%	0.0	98.43%	910	WP_012723273.1
glutamate synthase-related protein [Methanocaldococcus jannaschii]	Methanocaldococcus jannaschii	1029	1029	100%	0.0	97.84%	910	WP_012469262.1

41

blastp Similarity Search: Result Page

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

Job Title: spQ58746.1
 RID: 02C1W3301N
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1
 Description: RecName: Full-Archaeal glutamate synthase [NADPH]; Al...

Conserved domains

Distribution of the top 100 Blast Hits on 100 subject sequences

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

100 sequences selected

Putative conserved domains

Conserved domains: glutamate synthase superfamily

42

blastp Similarity Search: Result Page

↓

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

Range 1: 6 to 502 GenPept Graphics

Score: 227 bits (1877) 0.0 Compositional matrix adjust. 352/504(81%) 413/504(81%) 7/504(1%)

Query 6 VPPKYVEVDFRHLCEKTCESWVYRREGRIISYRRCGACRCVYVCRDALT 65
 Sbjct 6 +HVE-D +D C RC CS+GVVREGRI S C ACIR C CRDALT 65

Query 66 KENATSMRSHPLWVDARVDVNDKATGCLLSGNAKHPYFEMKLVLDACVNP 12
 Sbjct 66 YEKFTYRSHVYRVEKRVNKAQKIKVAGHGVLPYRFRDALLDACVNP 12

Query 126 DPLREPELRYTGKPKOLEFEEVEIDGKKIKAKLTKIAPLKLDTPEIAHSH 18
 Sbjct 126 DPLREPELRYTGKPKOLEFEEVEIDGKKIKAKLTKIAPLKLDTPEIAHSH 18

Query 186 GALSNAHLSFAKAVKCTGPTGEGGLPKALVYPAHITTVAGSAGFQNEVYKGS 24
 Sbjct 186 GALSNAHLSFAKAVKCTGPTGEGGLPKALVYPAHITTVAGSAGFQNEVYKGS 24

Query 246 ALIEIKGGAGKPGISGKPGKRYVETSAITWYFGGDAISPAHPIYSTEKAGLV 38
 Sbjct 246 ALIEIKGGAGKPGISGKPGKRYVETSAITWYFGGDAISPAHPIYSTEKAGLV 38

BIOS477/877 L12 - 43

43

blastp Similarity Search: Result Page

↓

BLAST® » blastp suite » results for RID-Q58746.1

Job Title: spQ58746.1
 RID: Q58746.1
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1

Sequences prof

select all 100

Scientific Name	Max Score	Total Score	Query E	Per. Ident	Accession
Methanoregula boonei	1043	1043	0.0	100.0%	WP_012107254.1
Methanoregula boonei	1032	1032	0.0	98.4%	WP_012073860.1
Methanoregula boonei	1031	1031	0.0	98.4%	WP_012073871.1
Methanoregula boonei	1029	1029	0.0	97.8%	WP_012468922.1

BIOS477/877 L12 - 44

44

blastp Similarity Search: Result Page

↓

BLAST® » blastp suite » results for RID-Q58746.1

Job Title: spQ58746.1
 RID: Q58746.1
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1

Blast Tree View

select all 100

BIOS477/877 L12 - 45

45

blastp Similarity Search: Result Page

↓

BLAST® » blastp suite » results for RID-Q58746.1

Job Title: spQ58746.1
 RID: Q58746.1
 Program: BLASTP
 Database: nr
 Query ID: Q58746.1

Sequences prof

select all 100

Scientific Name	Max Score	Total Score	Query E	Per. Ident	Accession
Methanoregula boonei	1043	1043	0.0	100.0%	WP_012073860.1
Methanoregula boonei	1032	1032	0.0	98.4%	WP_012073871.1
Methanoregula boonei	1031	1031	0.0	97.8%	WP_012468922.1

BIOS477/877 L12 - 46

46

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: 4

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)

show results in a new window

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

BIOS477/877 L12 - 47

47

blastp Protein Similarity Search

↓

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

Sequences producing significant alignments

select all 100 sequences selected

Description	Scientific Name	Max Score	Total Score	Query E	Per. Ident	Accession
glutamate synthase-related protein	Methanocaldococcus jannaschii	1043	1043	0.0	100.0%	WP_012073860.1
glutamate synthase-related protein	Methanocaldococcus jannaschii	1032	1032	0.0	98.4%	WP_012073871.1
glutamate synthase-related protein	Methanocaldococcus jannaschii	1029	1029	0.0	97.8%	WP_012468922.1

BIOS477/877 L12 - 48

48

blastp Protein Similarity Search

Algorithm parameters

General Parameters

Max target sequences: 100 Can be increased up to 5000

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 - 49

49

blastp Protein Similarity Search

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

Sequences producing significant alignments

Description	Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
glutamate synthase related protein [Methanocaldococcus jannaschii]	1043	1043	100%	0.0	100.00%	910	W61_010070860.1
glutamate synthase related protein [Methanocaldococcus jannaschii]	1042	1032	100%	0.0	98.43%	910	W61_010070863.1
glutamate synthase related protein [Methanocaldococcus jannaschii]	307	297	92%	8e-06	79.87%	499	W61_010070814.1
glutamate synthase related protein [Methanocaldococcus jannaschii]	297	297	42%	9e-05	60.00%	219	W61_010070830.1
glutamate synthase related protein [Methanocaldococcus jannaschii]	311	311	93%	1e-04	38.54%	421	W61_010070813.1
glutamate synthase related protein [Methanocaldococcus jannaschii]	313	380	71%	1e-04	48.32%	711	W61_010070821.1
glutamate synthase family protein [Methanocaldococcus jannaschii]	313	383	66%	1e-04	48.86%	709	W61_010070824.1
glutamate synthase domain-containing protein 1 [Thermoplasma volcanium]	313	313	98%	1e-04	37.18%	714	W61_010070827.1

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 L12 - 50

50

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?

Limit the search against eukaryotes

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

BIOS477/877 L12 - 51

51

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?

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 L12 - 52

52

blastp Similarity Search: a Case Study

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

Query seq. Specific hits

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 L12 - 53

53

blastp Similarity Search: a Case Study

[Search using the "Query subrange" option]

Specify the query range for the search

E-value increased (for short sequences)

BIOS477/877 L12 - 54

54

blastp Similarity Search: a Case Study

[Search using the "Query subrange" option]

Distribution of the top 46 Blast Hits on 46 subject sequences

Dihydroxylate dehydrogenase family protein [Tritrichomonas foetus]
Sequence ID: [GI1700364.1](#) Length: 924 Number of Matches: 1

BIOS477/877 L12 - 55

55

blastp Similarity Search: a Case Study

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

Distribution of the top 9 Blast Hits on 9 subject sequences

Dihydroxylate dehydrogenase family protein [Tritrichomonas foetus]
Sequence ID: [GI1700364.1](#) Length: 924 Number of Matches: 1

BIOS477/877 L12 - 55

56

blastp Similarity Search: a Case Study

[Max Matches = 3]

Distribution of the top 9 Blast Hits on 9 subject sequences

Dihydroxylate dehydrogenase family protein [Tritrichomonas foetus]
Sequence ID: [GI1700364.1](#) Length: 924 Number of Matches: 1

BIOS477/877 L12 - 57

57

blastp Similarity Search: a Case Study

Without Max Matches option (Downloaded in "Hit Table (csv)", imported to Excel)

query	subject	% identity	alignment length	mis-matches	gap	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	MCL4149161.1	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	3.88E-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	

Distribution of the top 9 Blast Hits on 9 subject sequences

Dihydroxylate dehydrogenase family protein [Tritrichomonas foetus]
Sequence ID: [GI1700364.1](#) Length: 924 Number of Matches: 1

BIOS477/877 L12 - 58

58