Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 12

1

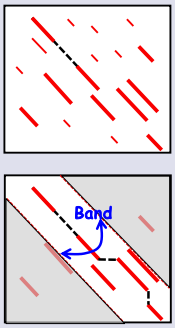
TODAY'S TOPICS

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

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



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.ebi.ac.uk/jdispatcher/sss/fasta

[FASTA guide]

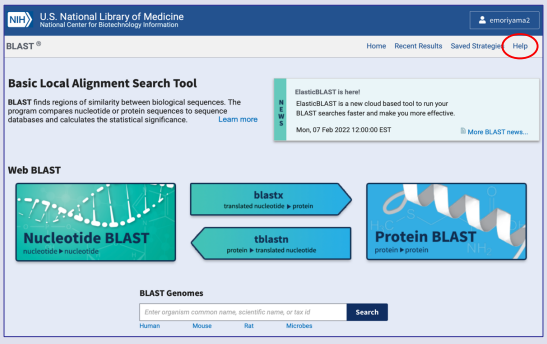
https://fasta.bioch.virginia.edu/wrp_fasta/fasta_guide.pdf

[William R. Pearson's website]

https://fasta.bioch.virginia.edu/wrpearson/

4

BLAST Similarity Search



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

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]

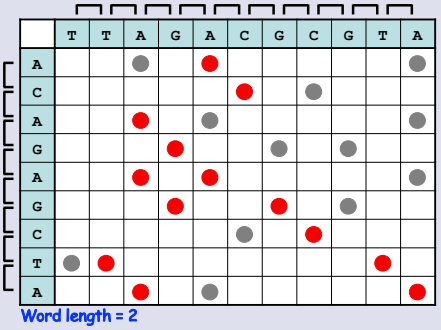
https://www.ncbi.nlm.nih.gov/books/NBK279690/

[BLAST YouTube Tutorials]

(Link is available from NCBI Help page or from Canvas)

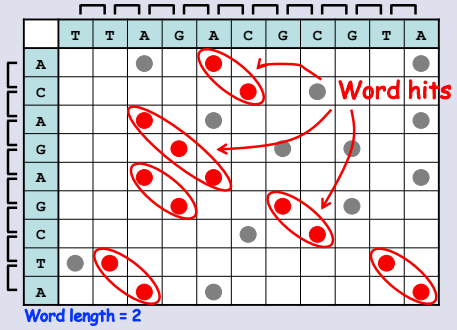
6

BLAST Algorithm: seeding



7

BLAST Algorithm: seeding



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 / TT = 2 - 1 = 1

TC / CC = -1 + 2 = 1

...

Neighborhood

TC / TA = 2 - 5 = -3

TC / TG = 2 - 5 = -3

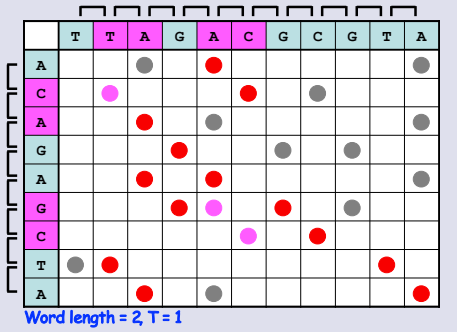
TC / AC = -5 + 2 = -3

TC / AT = -5 - 1 = -6

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

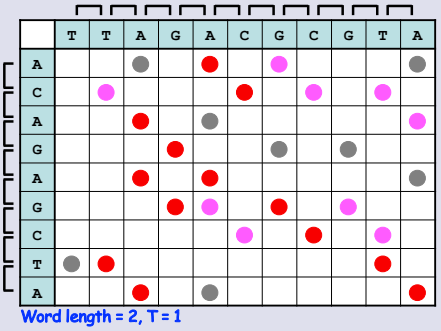
9

BLAST Algorithm: seeding



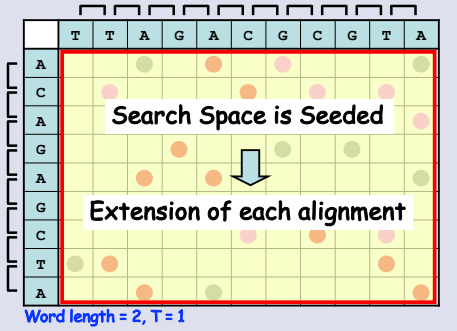
10

BLAST Algorithm: seeding

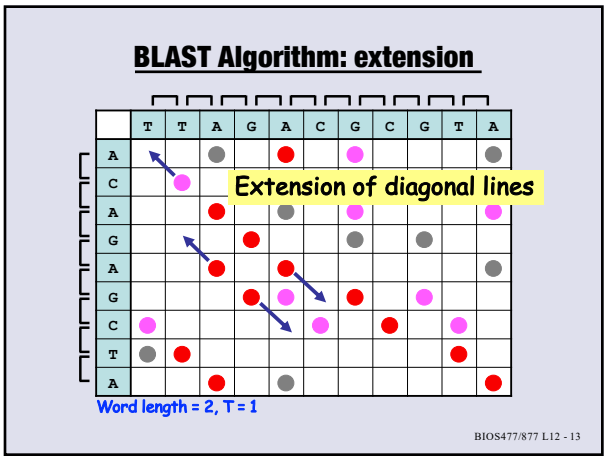


11

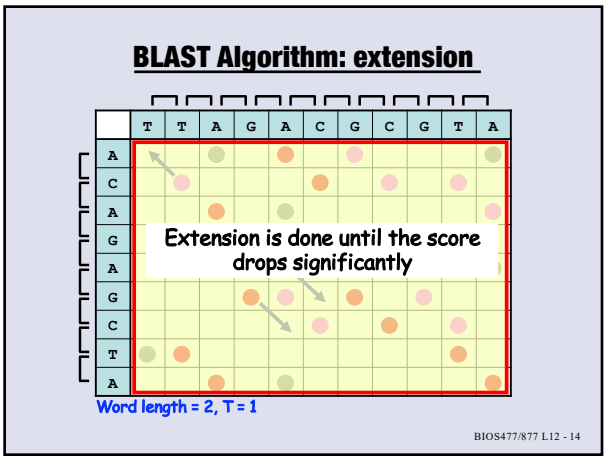
BLAST Algorithm



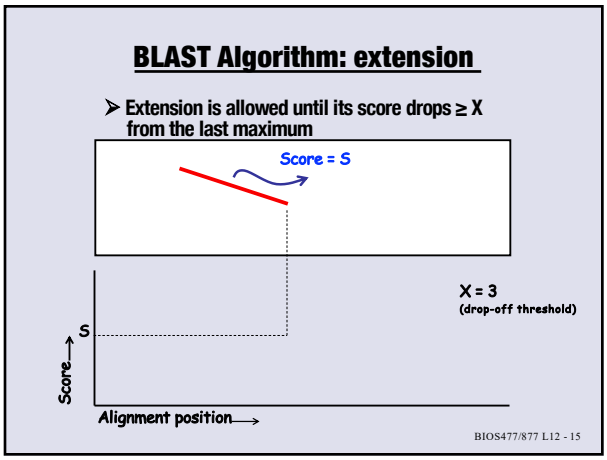
12



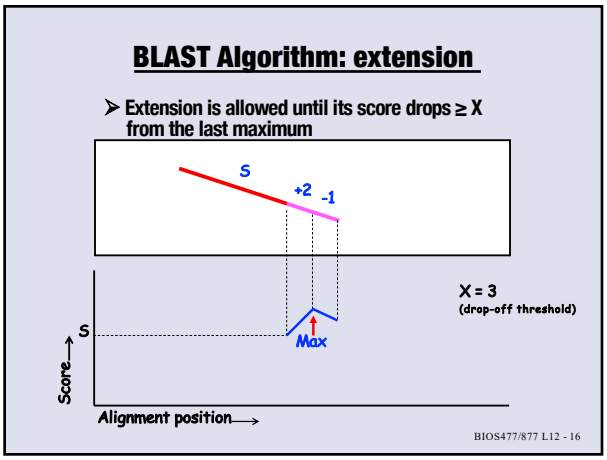
13



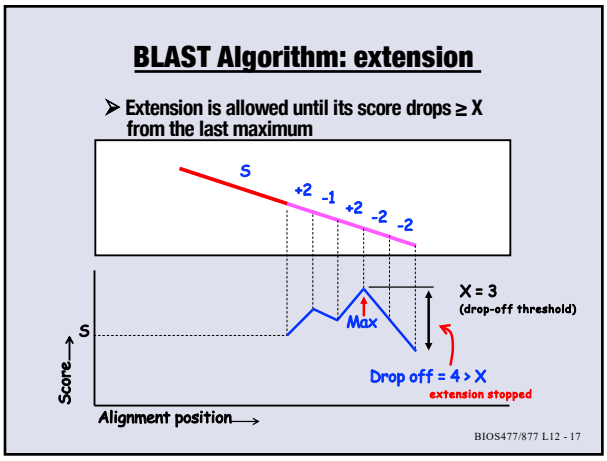
14



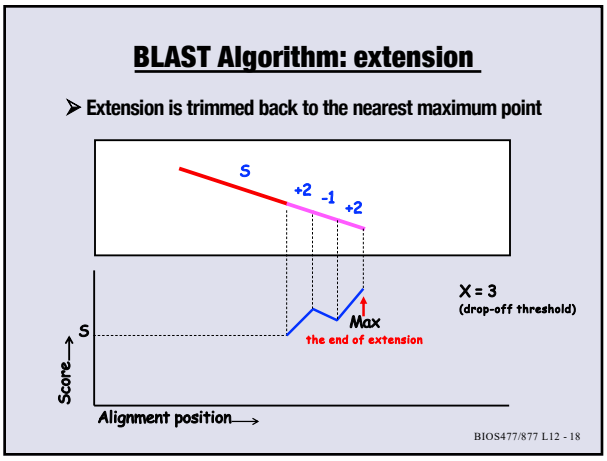
15



16



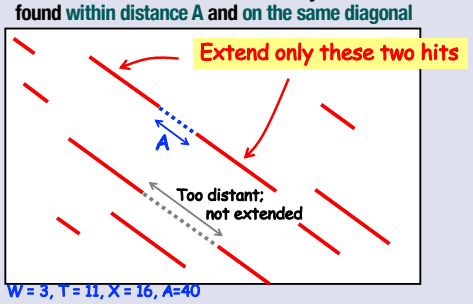
17



18

BLAST Algorithm: extension

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

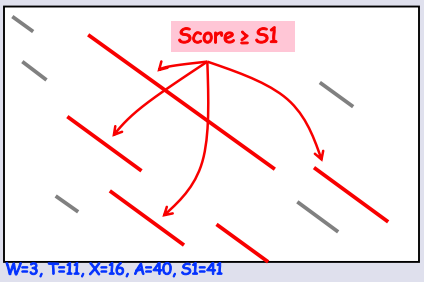


W = 3, T = 11, X = 16, A=40

19

BLAST Algorithm: evaluation

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

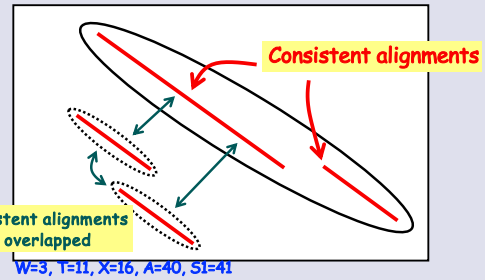


W=3, T=11, X=16, A=40, S1=41

20

BLAST Algorithm: evaluation

➢ Consistent alignments are grouped for E-value calculation

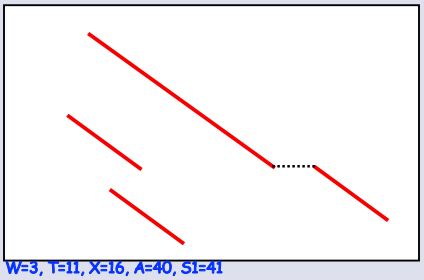


W=3, T=11, X=16, A=40, S1=41

21

BLAST Algorithm: Gapped extension

➢ Gapped extension is triggered after high score ungapped alignments are found

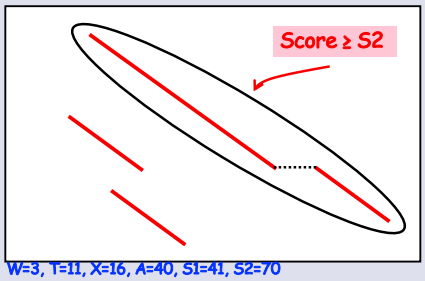


W=3, T=11, X=16, A=40, S1=41

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

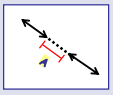
23

BLAST Algorithm

➢ Word-matching size (W)
→ longer words: faster but less sensitive

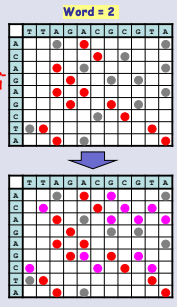
➢ Neighborhood threshold (T)
→ lower T: detects weaker similarities
→ slower but more sensitive

➢ Extension
• Drop-off score (X)
• Two-hit method
(A: distance b/w 2 hits)



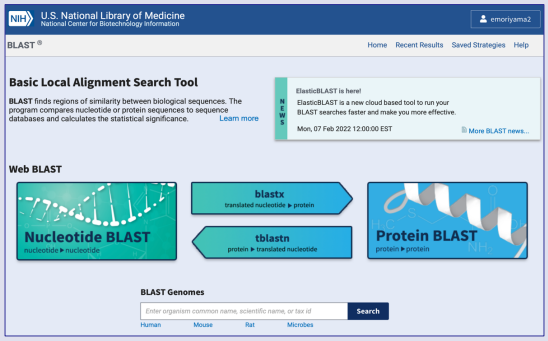
➢ HSP selection
(ungapped alignment threshold: S1)

➢ Gapped HSP extension
(gapped alignment threshold: S2)



24

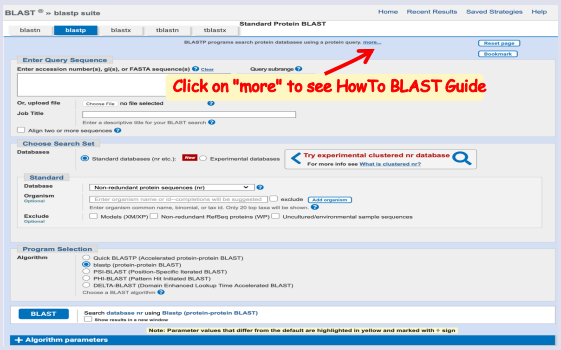
BLAST Similarity Search



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

25

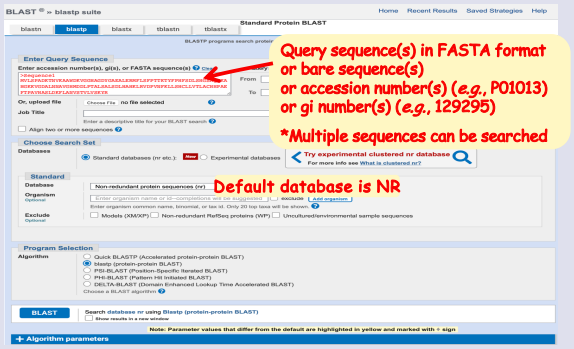
blastp Protein Similarity Search



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

26

blastp Protein Similarity Search



27

BLAST Databases (Protein)

Protein database	Description
Non-redundant (nr) default	Non-redundant protein sequences from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq
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/db/

28

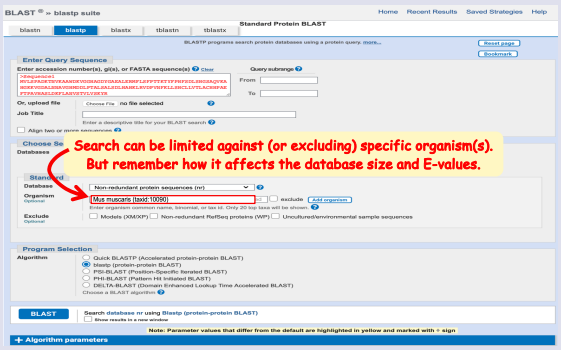
BLAST Landmark Database (Protein)

#	Organism	Assembly	Superkingdom
1	Bacillus subtilis	GCF_000009045.1	Bacteria
2	Deinococcus radiodurans R1	GCF_000008565.1	Bacteria
3	Escherichia coli str. K-12 substr. MG1655	GCF_000005845.2	Bacteria
4	Microcystis aeruginosa NIES-843	GCF_000010625.1	Bacteria
5	Mycobacterium tuberculosis H37Rv	GCF_000195955.2	Bacteria
6	Neisseria meningitidis	GCF_000008805.1	Bacteria
7	Peptoclostridium difficile 630	GCF_000009205.1	Bacteria
8	Pseudomonas aeruginosa	GCF_000006765.1	Bacteria
9	Shewanella oneidensis MR-1	GCF_000146165.2	Bacteria
10	Streptococcus pneumoniae R6	GCF_000007045.1	Bacteria
11	Streptomyces coelicolor	GCF_000203835.1	Bacteria
12	Synechocystis sp. PCC 1148	GCF_000340785.1	Bacteria
13	Thermotoga maritima MSB8	GCF_000008545.1	Bacteria
14	Methanothermobacter thermautotrophicus str. Delta H	GCF_000008645.1	Archaea
15	Sulfolobus acidocaldarius DSM 639	GCF_000012285.1	Archaea
16	Arabidopsis thaliana (thale cress)	GCF_000001735.3	Eukaryota
17	Caenorhabditis elegans	GCF_000002985.6	Eukaryota
18	Danio rerio (zebrafish)	GCF_000002035.6	Eukaryota
19	Dictyostelium discoideum AX4	GCF_000004695.1	Eukaryota
20	Drosophila melanogaster (fruit fly)	GCF_000001215.4	Eukaryota
21	Glycine max (soybean)	GCF_000004515.3	Eukaryota
22	Homo sapiens (human)	GCF_000001405.30	Eukaryota
23	Leishmania donovani	GCF_000002285.1	Eukaryota
24	Mus musculus (house mouse)	GCF_000001635.24	Eukaryota
25	Plasmodium falciparum 3D7	GCF_000002765.31	Eukaryota
26	Saccharomyces cerevisiae (baker's yeast)	GCF_000146045.2	Eukaryota
27	Schizosaccharomyces pombe (fission yeast)	GCF_000002945.1	Eukaryota

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

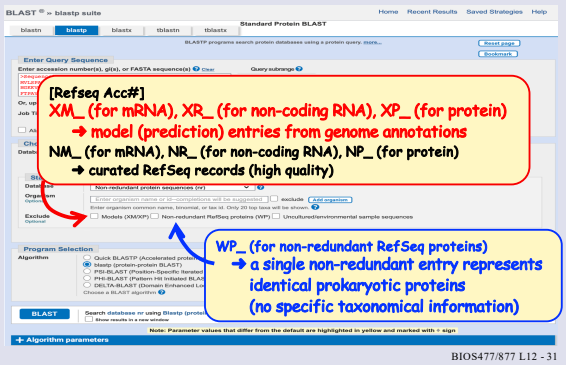
29

blastp Protein Similarity Search



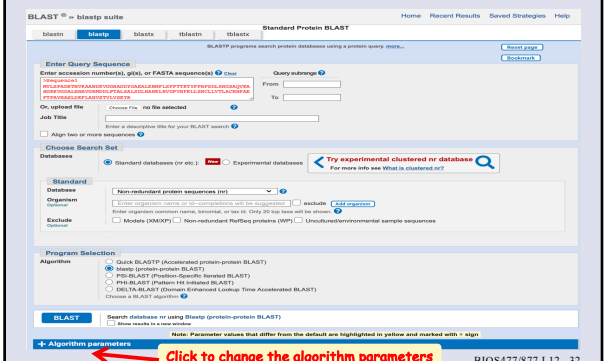
30

blastp Protein Similarity Search



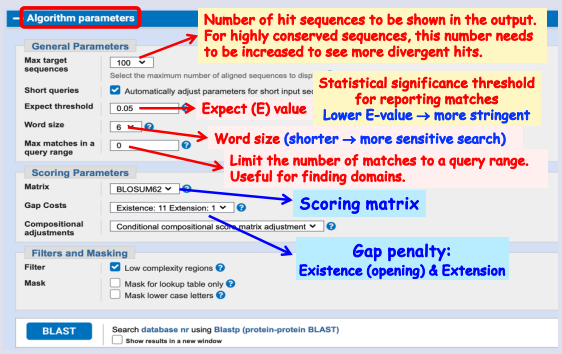
31

blastp Protein Similarity Search



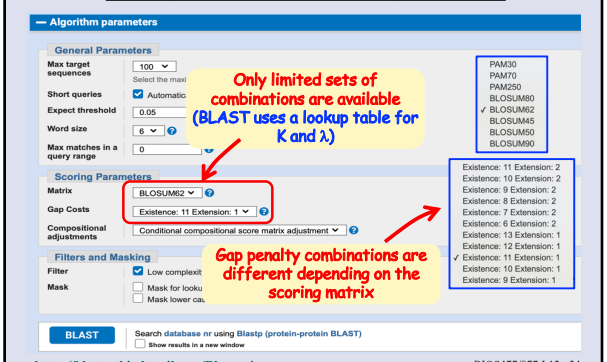
32

blastp Protein Similarity Search



33

blastp Protein Similarity Search



34

Significance of Alignment Scores

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

a	b	λ	K	H (nats)	a	b	λ	K	H (nats)
∞	0–∞	0.232	0.11	0.34	11	8–11	0.197	0.05	0.21
					11	6, 7	0.190	0.04	0.19
16	4–16	0.222	0.08	0.31	11	5	0.184	0.04	0.17
16	3	0.213	0.06	0.27	11	4	0.177	0.031	0.15
16	2	0.207	0.05	0.24	11	3	0.167	0.028	0.11
16	1	0.180	0.024	0.15	11	2	0.130	0.009	0.06
					11	1		Linear	
15	8–15	0.222	0.09	0.31					
15	6, 7	0.219	0.08	0.29	10	8–10	0.183	0.04	0.17
15	4, 5	0.216	0.07	0.28	10	6, 7	0.178	0.035	0.16
15	3	0.210	0.06	0.25	10	5	0.168	0.026	0.13
15	2	0.202	0.05	0.22	10	4	0.156	0.020	0.10
15	1	0.166	0.018	0.11	10	3	0.139	0.013	0.07
					10	2	0.099	0.007	0.03
14	8–14	0.218	0.08	0.29	10	1		Linear	
14	5–7	0.214	0.07	0.27					
14	4	0.205	0.05	0.24	9	7–9	0.164	0.029	0.13
14	3	0.201	0.05	0.22	9	5, 6	0.152	0.021	0.10
14	2	0.188	0.034	0.17	9	4	0.134	0.014	0.07
14	1	0.140	0.009	0.07	9	3	0.107	0.008	0.04
					9	1, 2		Linear	
13	8–13	0.211	0.06	0.27					
13	5–7	0.205	0.05	0.24	8	8	0.139	0.017	0.08
13	4	0.202	0.05	0.22	8	7	0.134	0.015	0.07
13	3	0.188	0.034	0.18	8	6	0.127	0.013	0.06
13	2	0.174	0.023	0.13	8	5	0.117	0.011	0.05
13	1	0.114	0.006	0.04	8	4	0.101	0.009	0.03
					8	1–3		Borderline or linear	
12	7–12	0.205	0.06	0.24					
12	5, 6	0.197	0.05	0.21	7	7	0.100	0.010	0.04
12	4	0.192	0.04	0.18	7	6	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		Borderline		1–6	1–6		Linear	

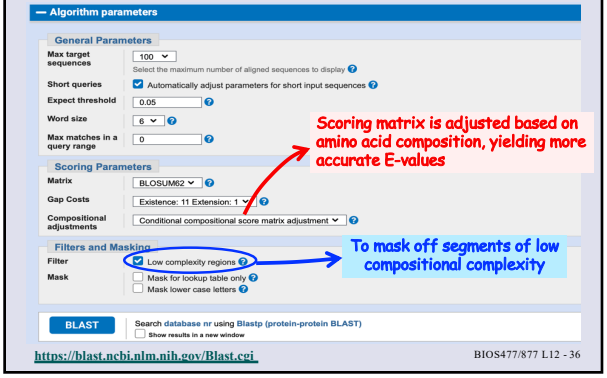
← For BLOSUM50:
a: gap opening penalty
b: gap extension penalty
(affine gap penalty)

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 $P(S \geq x) \approx Kmne^{-\lambda x}$

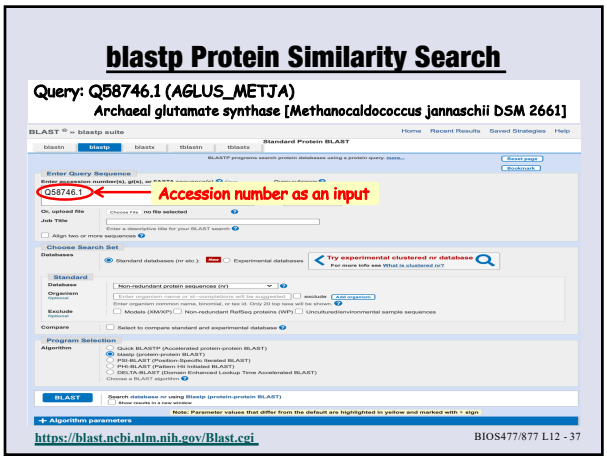
This is what BLAST does!

35

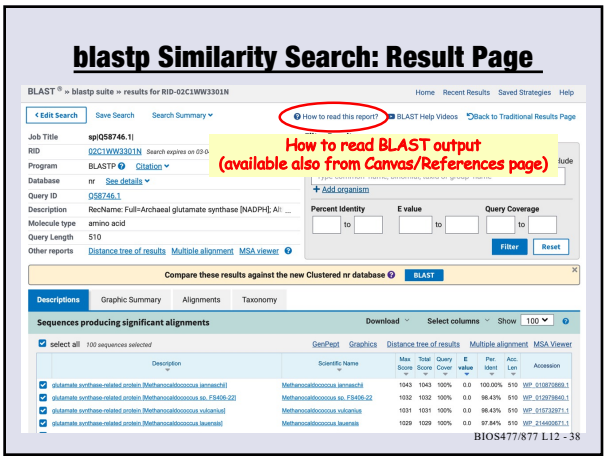
blastp Protein Similarity Search



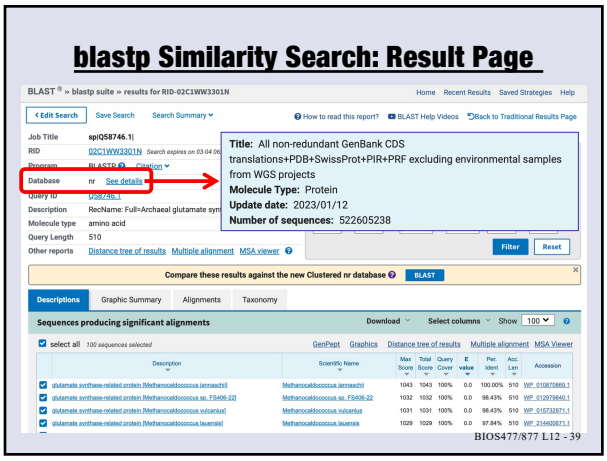
36



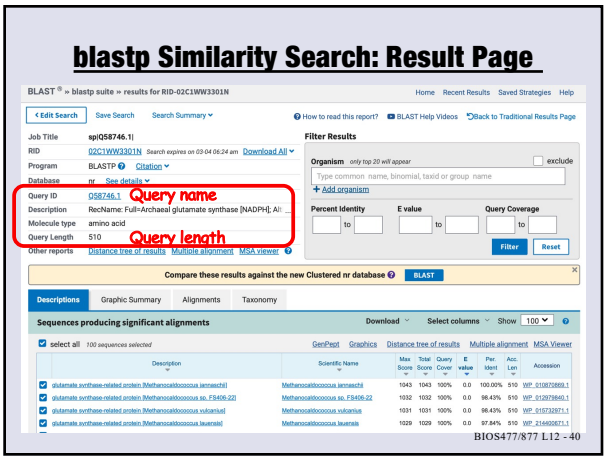
37



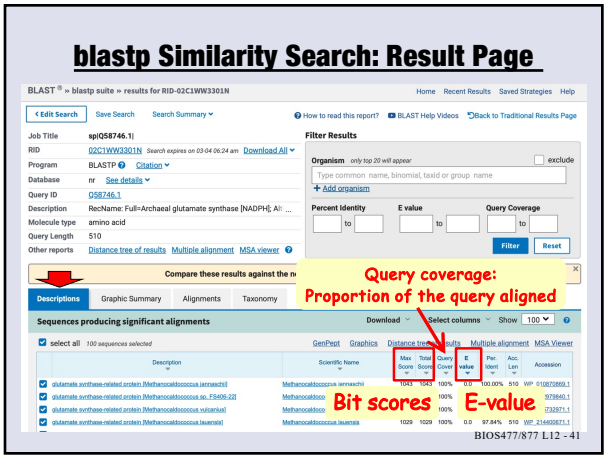
38



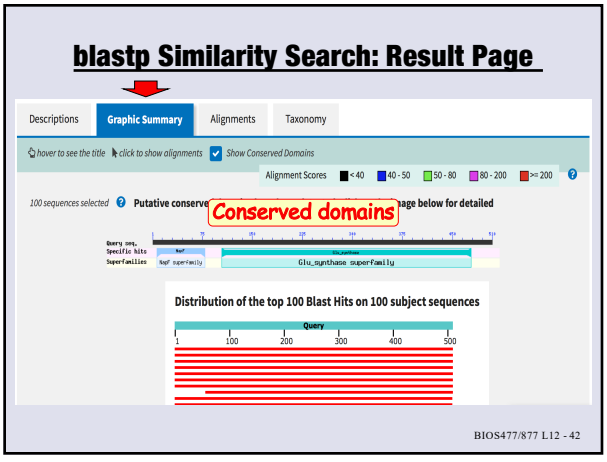
39



40

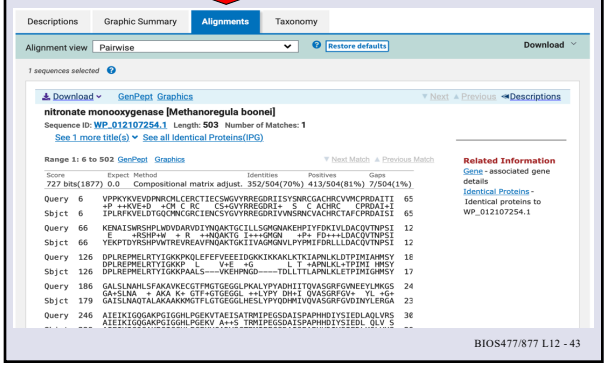


41



42

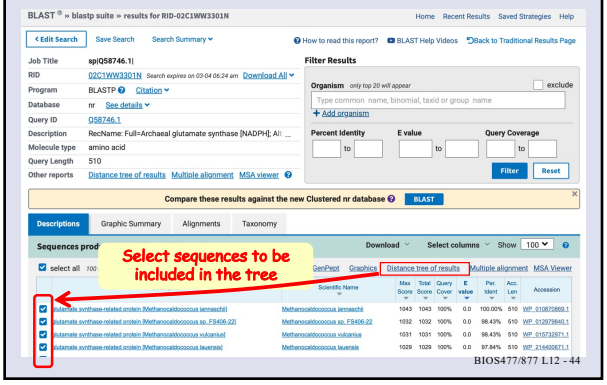
blastp Similarity Search: Result Page



BIOS477/877 L12 - 43

43

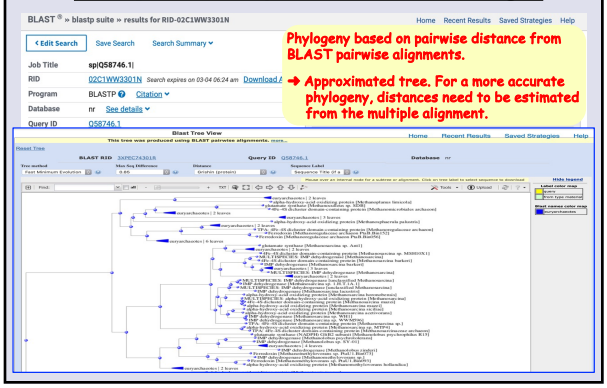
blastp Similarity Search: Result Page



BIOS477/877 L12 - 44

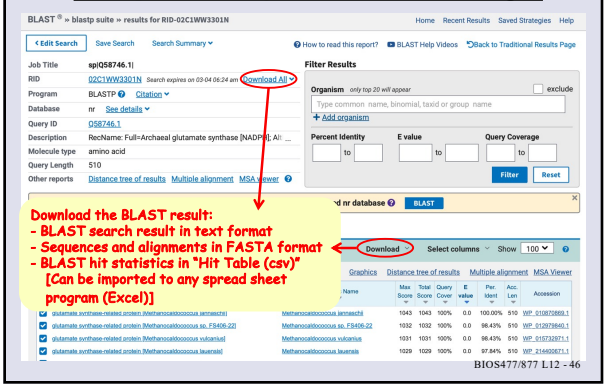
44

blastp Similarity Search: Result Page



45

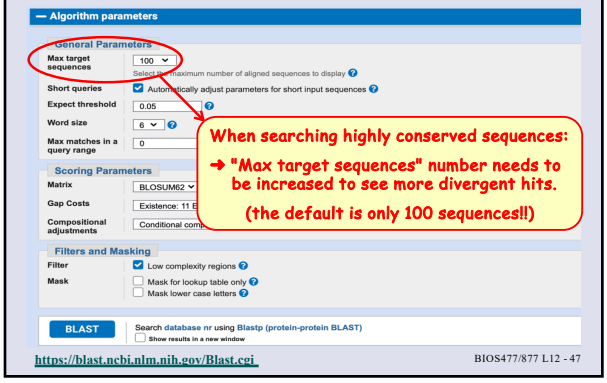
blastp Similarity Search: Result Page



BIOS477/877 L12 - 46

46

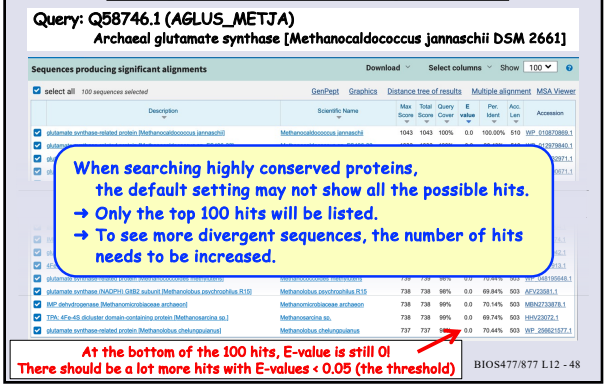
blastp Protein Similarity Search



BIOS477/877 L12 - 47

47

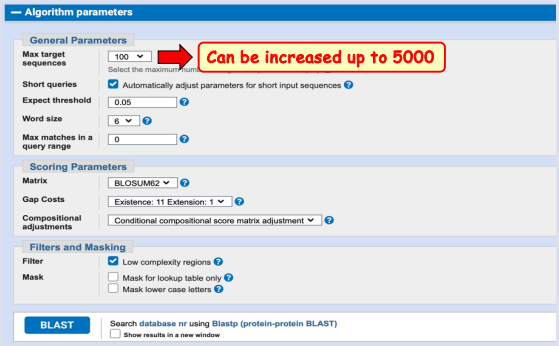
blastp Protein Similarity Search



BIOS477/877 L12 - 48

48

blastp Protein Similarity Search

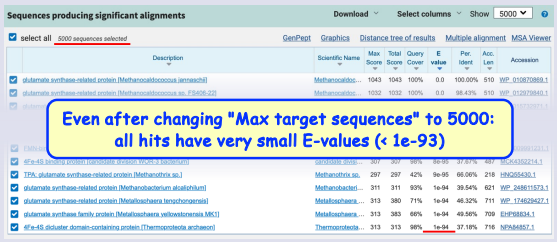


https://blast.ncbi.nlm.nih.gov/Blast.cgi

49

blastp Protein Similarity Search

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



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

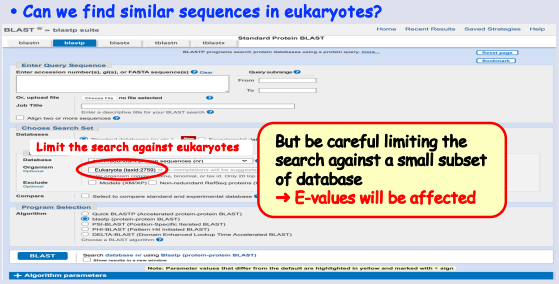
all hits are from bacterial/archaeal proteins

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?

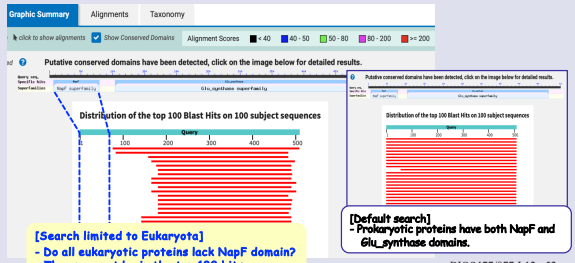


51

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?



52

blastp Similarity Search: a Case Study

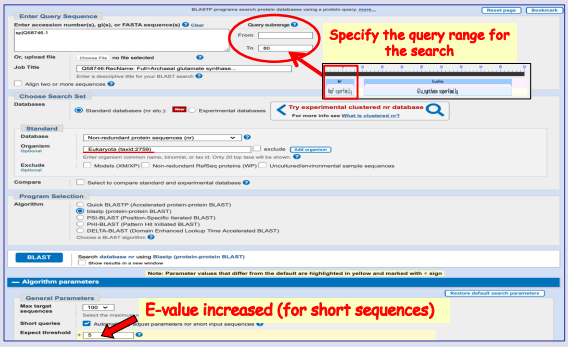
Query: Q58746.1 (AGLUS_METJA)
Archaeal glutamate synthase [*Methanocaldococcus jannaschii* DSM 2661]

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

53

blastp Similarity Search: a Case Study

[Search using the "Query subrange" option]



54

blastp Similarity Search: a Case Study

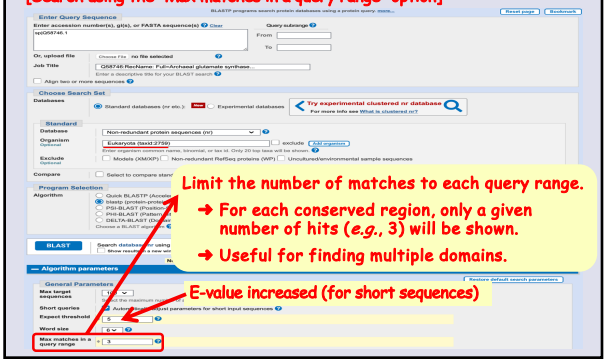
[Search using the "Query subrange" option]

BIOS477/877 L12 - 55

55

blastp Similarity Search: a Case Study

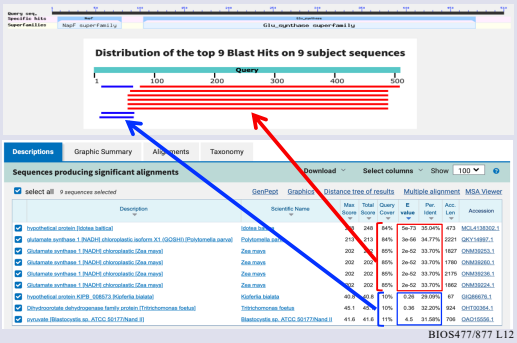
[Search using the "Max matches in a query range" option]



56

blastp Similarity Search: a Case Study

[Max Matches = 3]



BIOS477/877 L12 - 57

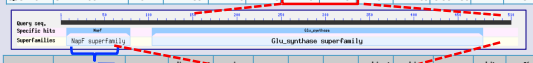
57

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 opens	query start	query end	subject start	subject end	e-value	bit score	% 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	3.83E-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



query	subject	% identity	alignment length	mis-matches	gap opens	query start	query end	subject start	subject end	e-value	bit score	% positives
Q58746.1	MCL4138302.1	35.045	448	247	6	78	509	47	466	5.42E-73	248	51.79
Q58746.1	QKY14997.1	34.773	440	262	7	61	490	881	1305	3.22E-56	213	48.41
Q58746.1	ONM39253.1	33.703	451	258	10	57	490	881	1307	2.09E-52	202	48.12
Q58746.1	ONM39260.1	33.703	451	258	10	57	490	881	1307	2.21E-52	202	48.12
Q58746.1	ONM39236.1	33.703	451	258	10	57	490	891	1317	2.29E-52	202	48.12
Q58746.1	ONM39224.1	33.703	451	258	10	57	490	578	1004	2.40E-52	202	48.12
Q58746.1	GIQ86676.1	29.091	55	37	1	13	65	7	61	0.26	40.8	54.55
Q58746.1	OHT00364.1	32.203	59	36	1	13	67	844	902	0.36	45.1	54.24
Q58746.1	OAO15556.1	31.579	57	38	1	10	66	225	280	4.5	41.6	52.63

With Max Matches = 3

BIOS477/877 L12 - 58

58