BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 **Lecture 12**

1

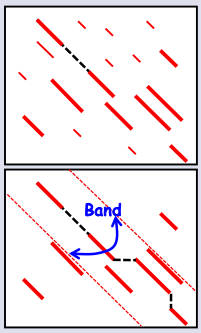
Today's topics

- Assignment 2 Review
- Similarity Search
  - BLAST algorithm
  - BLAST website and options
- Assignment 6

2

FASTA algorithm: summary

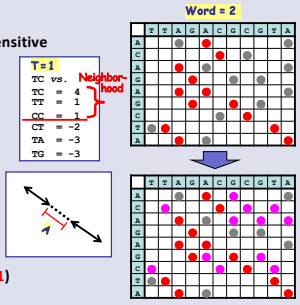
1. Find identities using k-tuples
2. Join diagonals without gaps
3. Choose top 10 diagonals using a scoring matrix (e.g., BLOSUM62) (init1: the top diagonal score)
4. Join again with gaps (initn: score of the longer diagonal)
5. A diagonal band is defined (width: 32 if k=1, 16 if k=2 for protein)
6.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



3

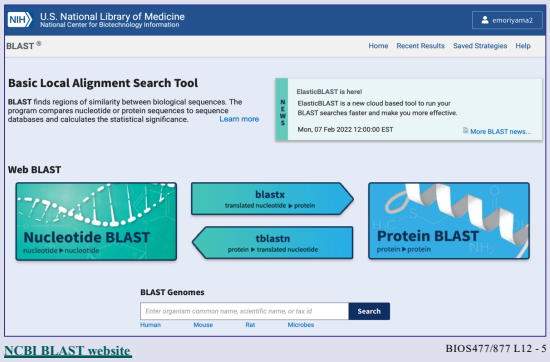
BLAST algorithm: summary

- 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 threshold (X)
  - Two-hit method (A: distance b/w 2 hits)
- HSP selection (ungapped alignment threshold: S1)
- Gapped HSP extension (gapped alignment threshold: S2)



4

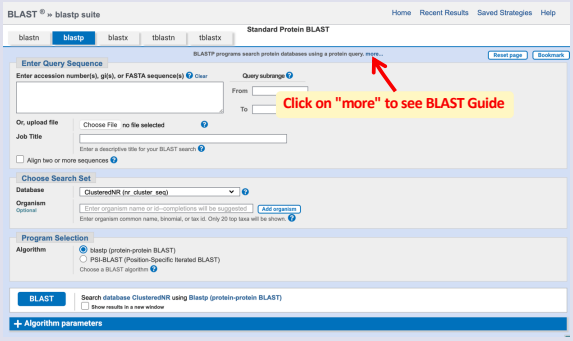
BLAST similarity search



NCBI BLAST website

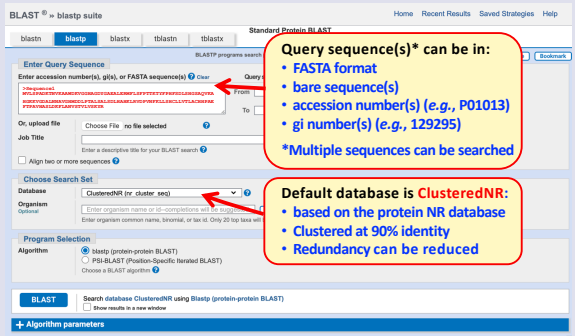
5

blastp protein similarity search



6

blastp protein similarity search



7

BLAST databases (protein)

Protein databas default	Description
Clustered non-redundant	Derived from the protein NR database by clustering sequences at 90% identity and 90% length
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
Non-redundant (nr)	Non-redundant protein sequences from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq
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.

Model Organisms (landmark) database: used by SmartBLAST

NCBI BLAST FTP site

8

BLAST landmark database (protein)

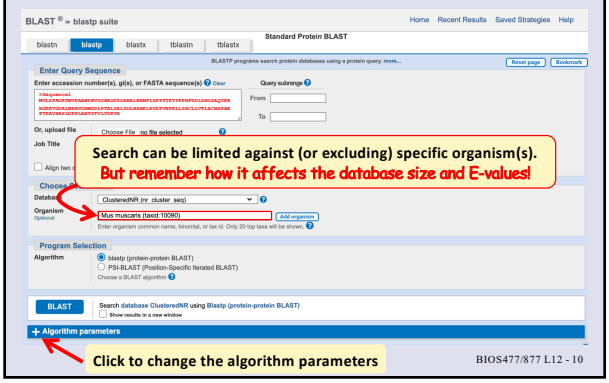
➢ Proteomes from 27 representative genomes

#	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_000002035.5	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_000122285.1	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

➜ Used for a quick search across all the kingdoms (Smart Blast)

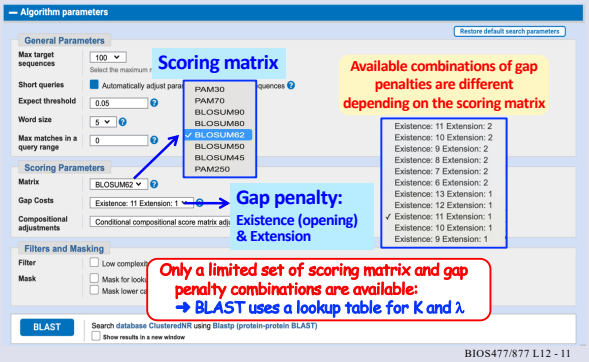
9

blastp protein similarity search



10

blastp search parameters



11

Estimation of K and λ in BLAST

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

a	b	λ	K	H (nats)
∞	0–∞	0.232	0.11	0.34
16	4–16	0.222	0.08	0.31
16	3	0.213	0.06	0.27
16	2	0.207	0.05	0.24
16	1	0.180	0.024	0.15
15	8–15	0.222	0.09	0.31
15	6, 7	0.219	0.08	0.29
15	4, 5	0.216	0.07	0.28
15	3	0.210	0.06	0.25
15	2	0.202	0.05	0.22
15	1	0.166	0.018	0.11
14	8–14	0.218	0.08	0.29
14	5–7	0.214	0.07	0.27
14	4	0.205	0.05	0.24
14	3	0.201	0.05	0.22
14	2	0.188	0.034	0.17
14	1	0.140	0.009	0.07
13	8–13	0.211	0.06	0.27
13	5–7	0.205	0.05	0.24
13	4	0.202	0.05	0.22
13	3	0.188	0.034	0.18
13	2	0.174	0.025	0.13
13	1	0.114	0.006	0.04
12	7–12	0.205	0.06	0.24
12	5, 6	0.197	0.05	0.21
12	4	0.192	0.04	0.18
12	3	0.178	0.028	0.15
12	2	0.158	0.019	0.10
12	1		Borderline	

a	b	λ	K	H (nats)
11	8–11	0.197	0.05	0.21
11	6, 7	0.190	0.04	0.19
11	5	0.184	0.04	0.17
11	4	0.177	0.031	0.15
11	3	0.167	0.028	0.11
11	2	0.130	0.009	0.06
11	1		Linear	
10	8–10	0.183	0.04	0.17
10	6, 7	0.178	0.035	0.16
10	5	0.168	0.026	0.13
10	4	0.156	0.020	0.10
10	3	0.139	0.013	0.07
10	2	0.099	0.007	0.03
10	1		Linear	
9	7–9	0.164	0.029	0.13
9	5, 6	0.152	0.021	0.10
9	4	0.134	0.014	0.07
9	3	0.107	0.008	0.04
9	1, 2		Linear	
8	8	0.139	0.017	0.08
8	7	0.134	0.015	0.07
8	6	0.127	0.013	0.06
8	5	0.117	0.011	0.05
8	4	0.101	0.009	0.03
8	1–3		Borderline or linear	
7	7	0.100	0.010	0.04
7	6	0.094	0.010	0.03
7	1–5		Borderline or linear	
1–6	1–6		Linear	

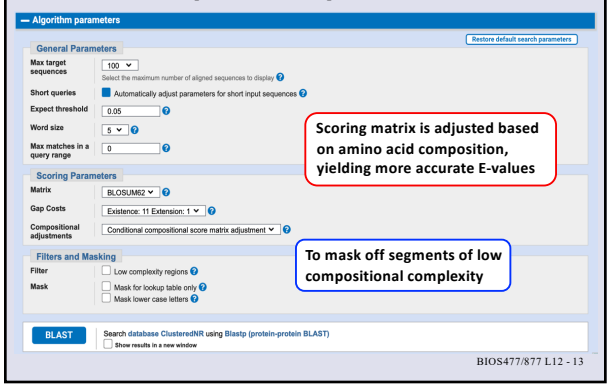
← For BLOSUM50;
a: gap opening penalty
b: gap extension penalty
(affine gap penalty)

Based on 10,000 random sequence pairs,
➜ calculate optimal local alignment scores, S.
➜ estimate K and λ by fitting the distribution of S with $P(S \geq x) \approx Kmne^{-\lambda x}$

This is what BLAST does

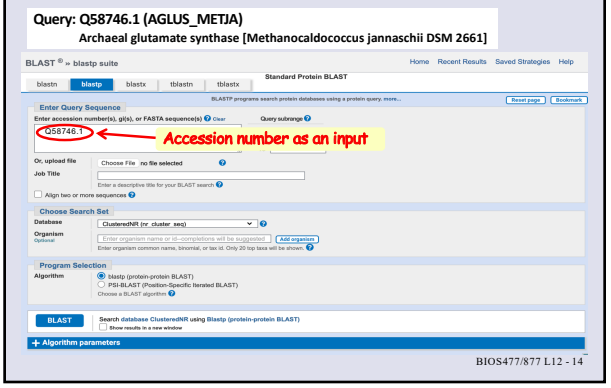
12

blastp search parameters



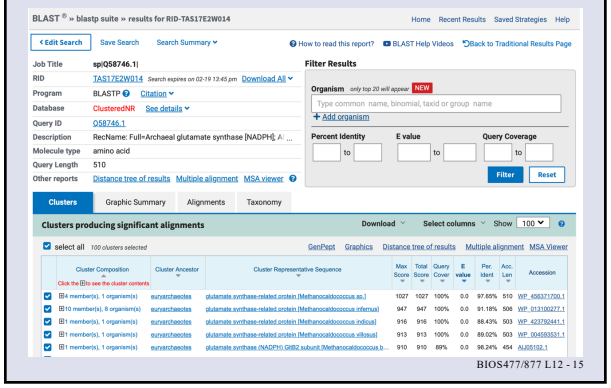
13

blastp similarity search (example)



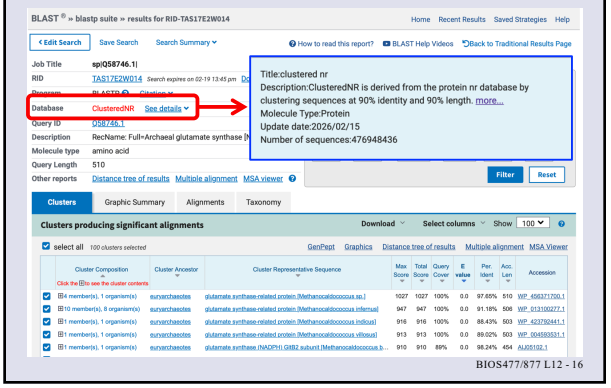
14

blastp similarity search: result page



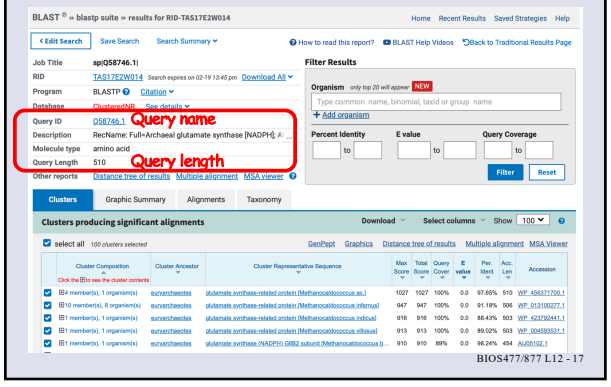
15

blastp similarity search: result page



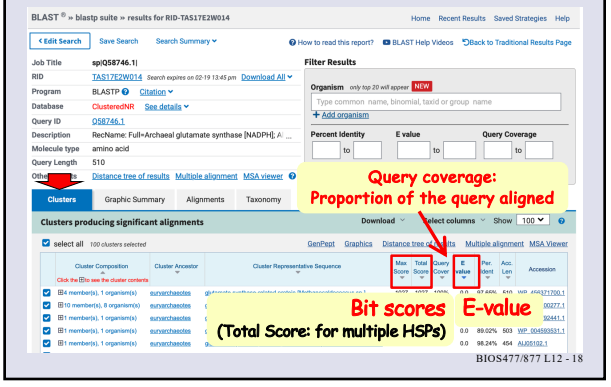
16

blastp similarity search: result page



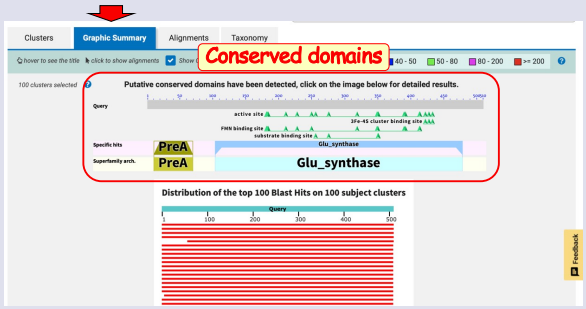
17

blastp similarity search: result page



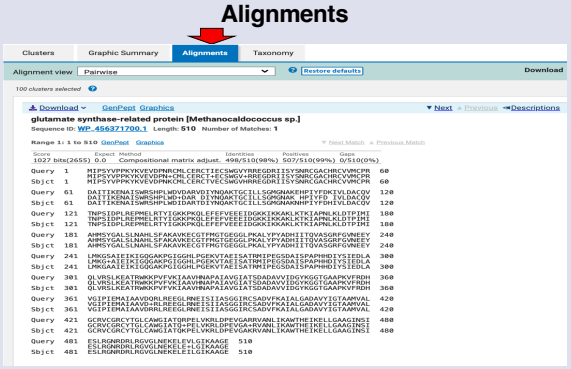
18

blastp similarity search result:
Graphic summary and conserved domains



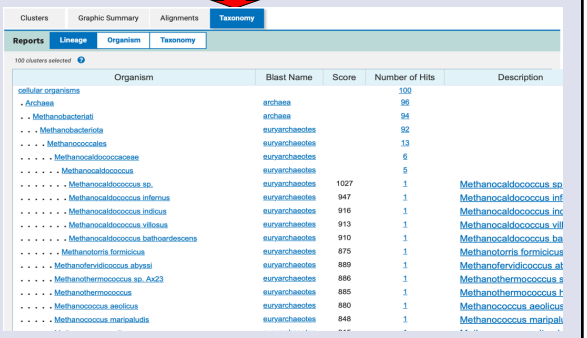
19

blastp similarity search result:
Alignments



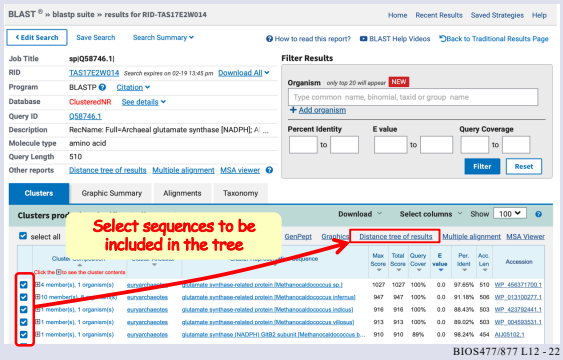
20

blastp similarity search result:
Taxonomic view



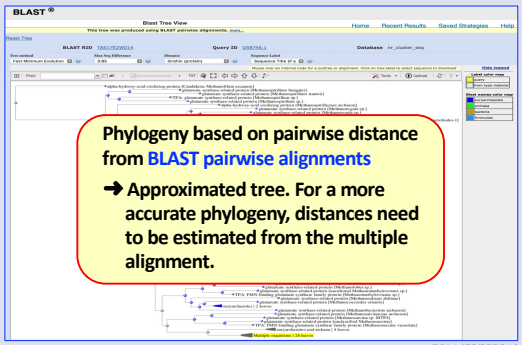
21

blastp similarity search result:
Tree view



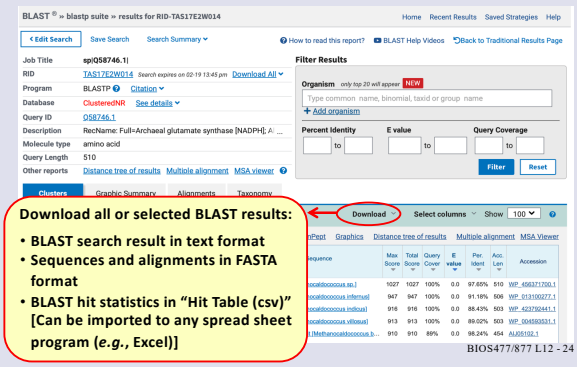
22

blastp similarity search result:
Tree view



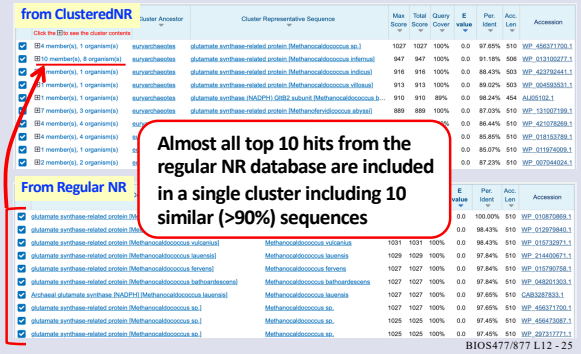
23

blastp similarity search result:
How to download the results



24

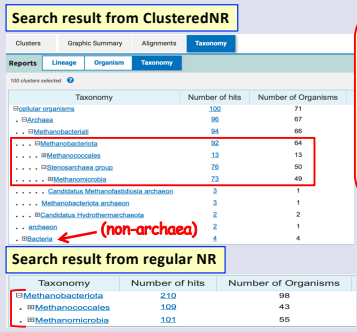
blastp similarity search:
ClusteredNR vs. regular NR databases



BIOS477/877 L12 - 25

25

blastp similarity search:
ClusteredNR vs. regular NR databases



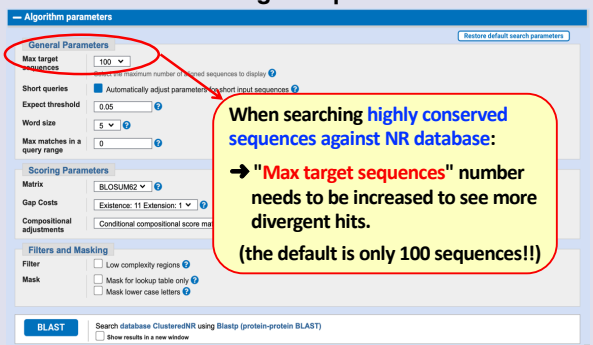
For highly conserved sequences, by using ClusteredNR:
- → Search can be faster
- → Results can be less redundant
- → Taxonomic coverage can be broader

Numbers of sequences:
- ClusteredNR: 476,948,436
- NR: 1,005,769,372
  - → NR is ~2 times larger
  - → E-values will be affected

BIOS477/877 L12 - 26

26

blastp search parameters:
Max target sequences



When searching highly conserved sequences against NR database:

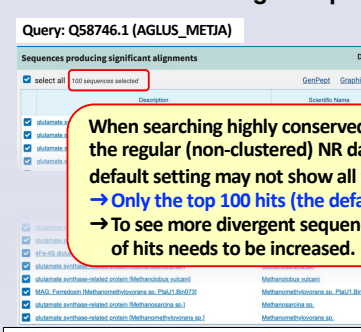
- → "Max target sequences" number needs to be increased to see more divergent hits.

(the default is only 100 sequences!!)

BIOS477/877 L12 - 27

27

blastp search parameters:
Max target sequences



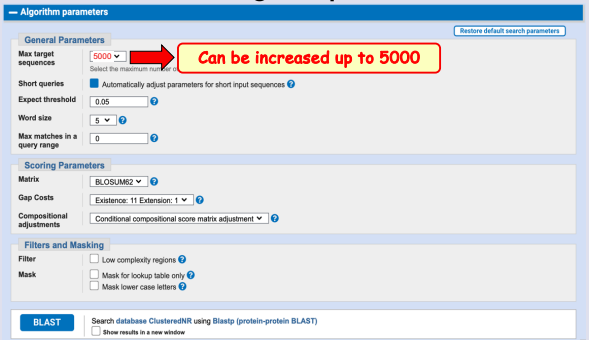
When searching highly conserved proteins using the regular (non-clustered) NR database, the default setting may not show all the possible hits:
- → Only the top 100 hits (the default) will be listed.
- → To see more divergent sequences, the number of hits needs to be increased.

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 L12 - 28

28

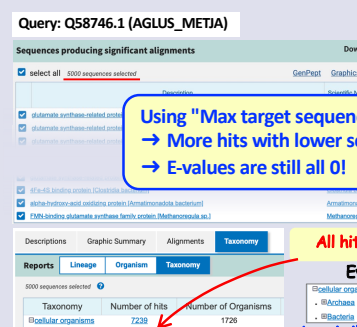
blastp search parameters:
Max target sequences



BIOS477/877 L12 - 29

29

blastp search parameters:
Max target sequences



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

All hits are from bacteria/archaea
Even using ClusteredNR!
Any similar proteins in eukaryotes?

BIOS477/877 L12 - 30

30