BLAST

(and MMSeqs2 & Foldseek)

Slides adapted & edited from a set by Cheryl A. Kerfeld (UC Berkeley/JGI) & Kathleen M. Scott (U South Florida)

Kerfeld CA, Scott KM (2011) Using BLAST to Teach ‘‘E-value-tionary’’ Concepts. *PLoS Biology* 9(2):e1001014

Starts with a Query Sequence in FASTA Format

Amino acid sequence:

```
>ribosomal protein L7/L12 [Thiomicrospira crunogena XCL-2]
MAITKDDILEAVANMSVMEVVELVEAMEEKFGVSAAAVAVAGPAGDAGAA
GEEQTEFDVVLTGAGDNKVAAIKAVRGATGLGLKEAKSAVESAPFTLKEG
VSKEEAETLANELKEAGIEVEVK
```

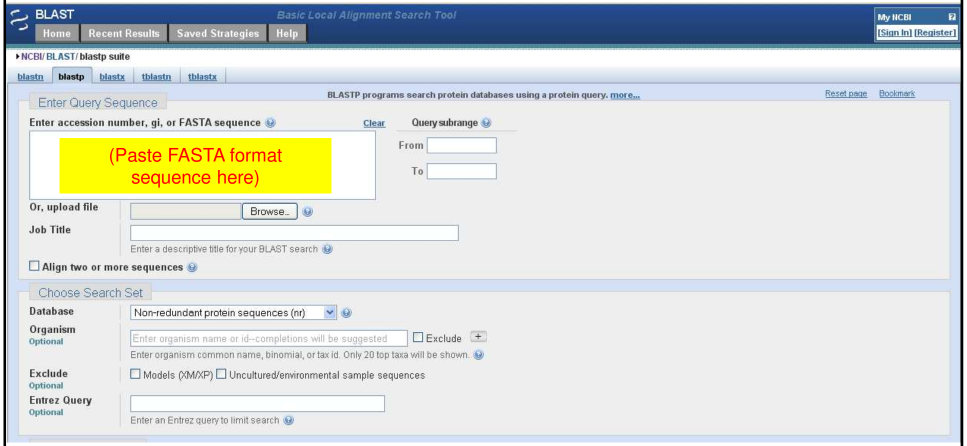
Note the description line
Starts with “>”, ends with carriage return
Not read as sequence data

Nucleotide sequence:

```
>gi|118139508:333094-333465 Thiomicrospira crunogena XCL-2
ATGGCAATTACAAAAGACGATATTTTAGAAGCAGTTGCTAACATGTCAGTAATGGAAG
TTGTTGAACTTGTTGAAGCAATGGAAGAGAAGTTTGGTGTTTCTGCAGCAGCAGTTGC
GGTTGCAGGTCCTGCAGGTGATGCTGGCGCTGCTGGTGAAGAACAAACAGAGTTTGAC
GTTGTCTTGACTGGTGCTGGTGACAACAAAGTTGCAGCAATCAAAGCCGTTCGTGGCG
CAACTGGTCTTGGGCTTAAAGAAGCGAAAAGTGCAGTTGAAAGTGCACCATTTACGCT
TAAAGAGGGTGTTTCTAAAGAAGAAGCAGAAACTCTTGCAAATGAGCTTAAAGAAGCA
GGTATTGAAGTCGAAGTTAAATAA
```

Kerfeld and Scott, PLoS Biology 2011

NCBI BLAST Interface
(blastp: for protein-protein alignments)



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

(Paste FASTA format sequence here)

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ Exclude

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

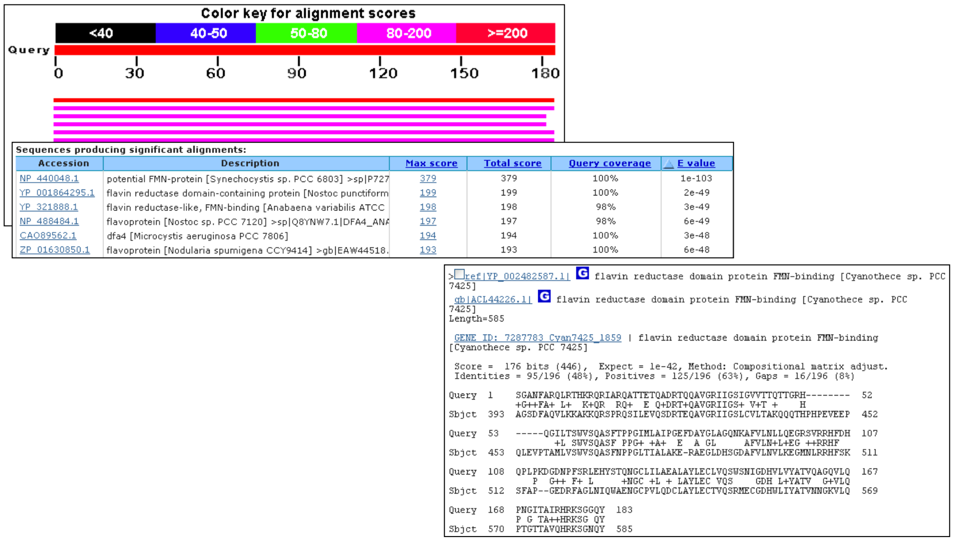
Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional)

Enter an Entrez query to limit search

Kerfeld and Scott, PLoS Biology 2011

NCBI BLAST Results Page:
Potential homologs retrieved from database



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
NP_440048.1	potential FMN-protein [Synechocystis sp. PCC 6803] >sp|P727	379	379	100%	1e-103
YP_001864295.1	flavin reductase domain-containing protein [Nostoc punctiform	199	199	100%	2e-49
YP_321888.1	flavin reductase-like, FMN-binding [Anabaena variabilis ATCC	198	198	98%	3e-49
NP_488484.1	flavoprotein [Nostoc sp. PCC 7120] >sp|Q8YNW7.1|DFA4_ANA	197	197	98%	6e-49
CAO89562.1	dfa4 [Microcystis aeruginosa PCC 7806]	194	194	100%	3e-48
ZP_01630850.1	flavoprotein [Nodularia spumigena CCY9414] >gb|EAW44518.	193	193	100%	6e-48

```
>□ref|YP_002482587.1| G flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
 gb|ACL44226.1| G flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
Length=585

 GENE ID: 7287783 Cyan7425_1859 | flavin reductase domain protein FMN-binding
[Cyanothece sp. PCC 7425]

 Score =  176 bits (446),  Expect = 1e-42, Method: Compositional matrix adjust.
 Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)

Query  1    SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTGRH--------  52
            +G++FA+ L+  K+QR  RQ+  E Q+DRT+QAVGRIIGS+ V+T +    H
Sbjct  393  AGSDFAQVLKKAKKQRSPRQSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP  452

Query  53   -----QGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRSVRRHFDH  107
                  +L SWVSQASF PPG+ +A+  E  A GL     AFVLN+L+EG ++RRHF
Sbjct  453  QLEVPTAMLVSWVSQASFNPPGLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK  511

Query  108  QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVLQ  167
               P  G++ F+ L      +NGC +L + LAYLEC VQS     GDH L+YATV  G+VLQ
Sbjct  512  SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVNNGKVLQ  569

Query  168  PNGITAIRHRKSGGQY  183
            P G TA++HRKSG QY
Sbjct  570  PTGTTAVQHRKSGNQY  585
```

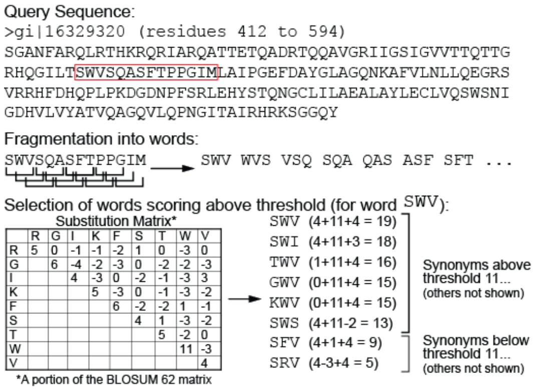
Kerfeld and Scott, PLoS Biology 2011

Overview of BLAST

1. Segment the query sequence into short “words”
2. Use the query sequence segments to scan the database for matching sequences
3. Extend the matched segments in either direction to find local alignments.
4. Create a list of hits & alignments, with best matches first

Kerfeld and Scott, PLoS Biology 2011

BLAST Phase 1: Segment the query sequence and identify words that <u>could form potential alignments</u>



Query Sequence:

```
>gi|16329320 (residues 412 to 594)
SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTG
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRS
VRRHFDHQPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNI
GDHVLVYATVQAGQVLQPNGITAIRHRKSGGQY
```

Fragmentation into words:

```
SWVSQASFTPPGIM ——> SWV WVS VSQ SQA QAS ASF SFT ...
```

Selection of words scoring above threshold (for word SWV):

Substitution Matrix*

	R	G	I	K	F	S	T	W	V
R	5	0	-1	-1	-2	1	0	-3	0
G		6	-4	-2	-3	0	-2	-2	-3
I			4	-3	0	-2	-1	-3	3
K				5	-3	0	-1	-3	-2
F					6	-2	-2	1	-1
S						4	1	-3	-2
T							5	-2	0
W								11	-3
V									4

*A portion of the BLOSUM 62 matrix

SWV (4+11+4 = 19)
SWI (4+11+3 = 18)
TWV (1+11+4 = 16)
GWV (0+11+4 = 15)
KWV (0+11+4 = 15)
SWS (4+11-2 = 13)

Synonyms above threshold 11... (others not shown)

SFV (4+1+4 = 9)
SRV (4-3+4 = 5)

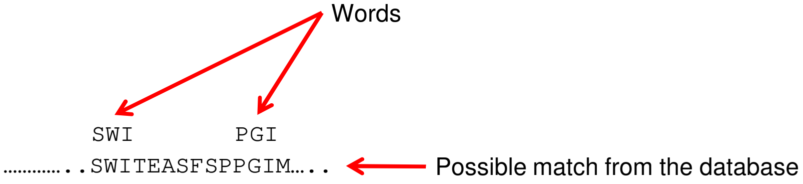
Synonyms below threshold 11... (others not shown)

- Segment the query sequence into pieces (“words”)
  - Default word length: 3 amino acids or 11 nucleic acids
- Create a list of synonyms and their scores for comparing query words to target words
  - Uses scoring matrix to calculate scores for synonyms that might be found in the database
- Save the scores (and synonyms) exceeding a given threshold T

Kerfeld and Scott, PLoS Biology 2011

BLAST Phase 2: Using the query sequence word list, scan the database for synonyms (hits)

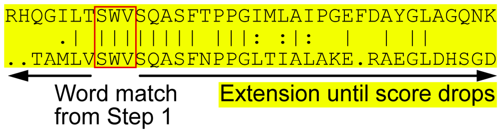
- Scan the database for matches to the word list with acceptable T values
- Require two matches (“hits”) within the target sequence
- Set aside sequences with matches above T for further analysis



Kerfeld and Scott, PLoS Biology 2011

BLAST Phase 3: Extending the hits

- Search 5’ and 3’ of the word hit on both the query and target sequence
- Add up the score for sequence identity or similarity until value exceeds S
- Alignment is dropped from subsequent analyses if value never exceeds S



```
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNK
    .| |||||||| |||: :|:  |  | ||
..TAMLVSWVSQASFNPPGLTIALAKE.RAEGLDHSGD
```

Kerfeld and Scott, PLoS Biology 2011

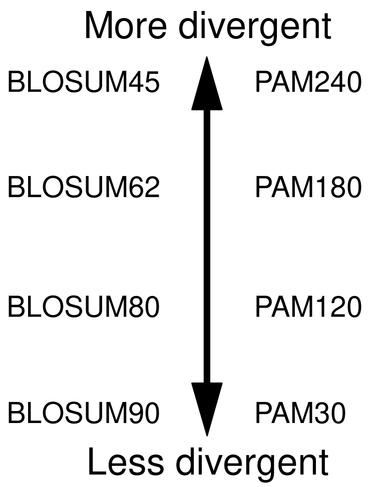
So, to summarize:

- BLAST segments query sequence into “words” and scores potential word matches
- Scans this list for alignments that meet a threshold score T
  - uses a scoring matrix to calculate this (e.g., **BLOSUM62**)
- Uses this list of ‘synonyms’ to scan the database
- Extends the alignments to see if they meet a cutoff score S
  - uses a scoring matrix to calculate this
- Reports the alignments that exceed S

Kerfeld and Scott, PLoS Biology 2011

PAM and BLOSUM Matrices

- Scoring matrices are calibrated to capture different degrees of sequence similarity
- In practice, this means choosing a matrix appropriate to the suspected degree of sequence identity between the query and its hits
- PAM: empirically derived for close relatives
- BLOSUM: empirically derived for distant relatives



More divergent

BLOSUM45 PAM240

BLOSUM62 PAM180

BLOSUM80 PAM120

BLOSUM90 PAM30

Less divergent

Kerfeld and Scott, PLoS Biology 2011

Raw Scores (*S* values) from an Alignment

$$S = (\Sigma M_{ij}) - cO - dG,$$

where

M = score from a similarity matrix
for a particular pair of amino acids (ij)

c = number of gaps

O = penalty for the existence of a gap

d = total length of gaps

G = per-residue penalty for extending the gap

Kerfeld and Scott, PLoS Biology 2011

Limitations of Raw Scores

- S values depend on the substitution matrix, gap penalties
- Impossible to compare S values from hits retrieved from BLAST searches when different matrices and gap penalties are used

Kerfeld and Scott, PLoS Biology 2011

Going from Raw Scores to Bit Scores

(as in 0 vs 1)

$$S' = [\lambda S - \ln(K)]/\ln(2)$$

where

S' = bit score

λ and K = normalizing parameters of the specific matrices and search spaces

- Larger raw scores result in larger bit scores
- Allows user to compare scores obtained by using different matrices and search spaces

Kerfeld and Scott, PLoS Biology 2011

Limitations of Bit Scores

- How high does a bit score have to be to suggest common ancestry?
  - Hard to evaluate hits as homologs or not, based solely on bit scores

Kerfeld and Scott, PLoS Biology 2011

E-value

- Number of distinct alignments with scores greater than or equal to a given value expected to occur in a search against a database of known size, based solely on chance, not homology.
  - Large E-values suggest that the query sequence and retrieved sequence similarities are due to chance
  - Small E-values suggest that the sequence similarities are due to shared ancestry (or potentially convergent evolution)

Kerfeld and Scott, PLoS Biology 2011

Calculating E-values

$$E = (n \times m) / 2^{S'}$$

where

m = effective length of the query sequence

= length of query sequence – average length of alignments

(Controls for fewer alignments occurring at the ends of the query sequence)

n = effective length of the database sequence

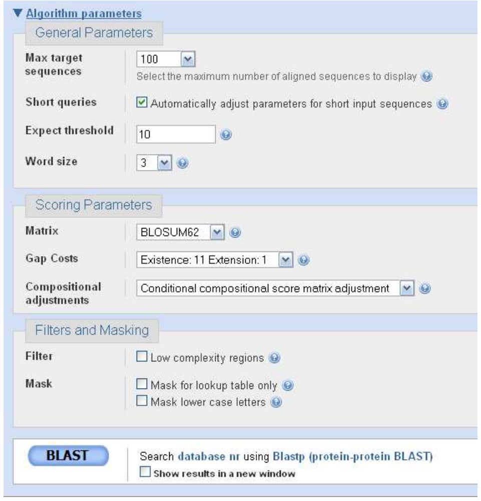
(total number of bases)

The value of E decreases exponentially with increasing S

Kerfeld and Scott, PLoS Biology 2011

BLAST Parameters

- Expect
- Word size
- Matrix
- Gap costs
- Filter
- Mask



▼ Algorithm parameters

General Parameters

Max target sequences: 100 — Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

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

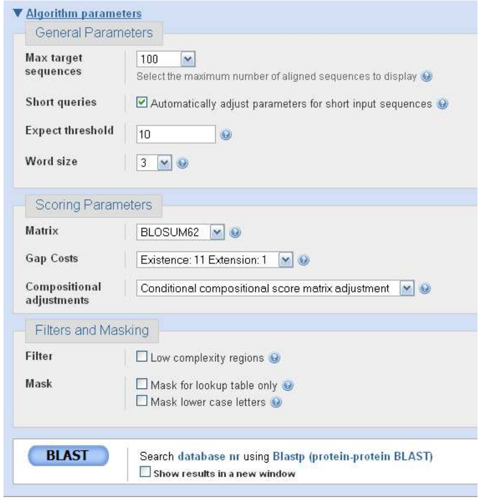
Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

Kerfeld and Scott, PLoS Biology 2011

E value Threshold

- Alignments will be reported with E-values less than or equal to the expect values threshold
  - Setting a larger E threshold will result in more reported hits
  - Setting a smaller E threshold will result in fewer reported hits



▼ Algorithm parameters

General Parameters

Max target sequences: 100 — Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

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

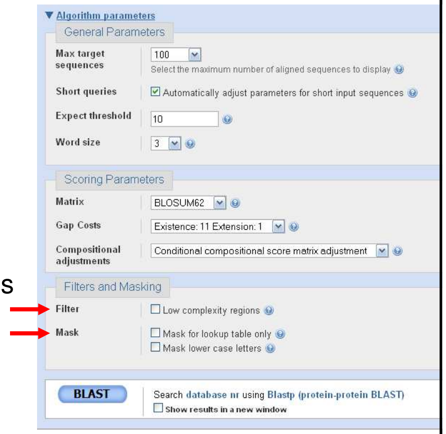
Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

Kerfeld and Scott, PLoS Biology 2011

Filter and Mask

- Filter: Low complexity
  - Replaces the following with N (nucleotides) or X (amino acids)
    - Dinucleotide repeats
    - Amino acid repeats
    - Leader sequences
    - Stretches of hydrophobic residues
- Mask: Lower case
  - Replaces lowercase letters in sequence with N or X
    - Lowercase letters typically indicate base or amino acid not known with certainty



▼ Algorithm parameters

General Parameters

Max target sequences: 100 — Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

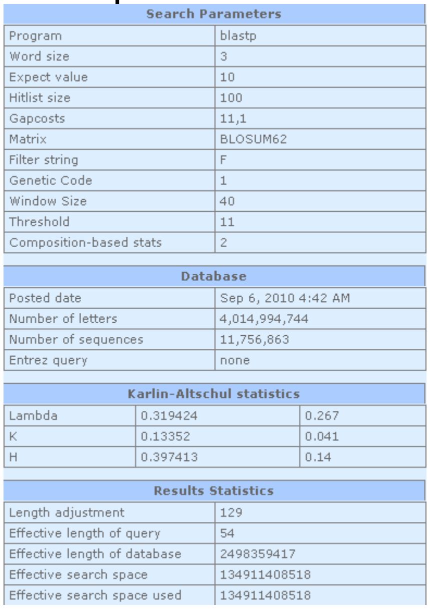
Filter: ☐ Low complexity regions

Mask: ☐ Mask for lookup table only
☐ Mask lower case letters

Search database nr using Blastp (protein-protein BLAST)
☐ Show results in a new window

Kerfeld and Scott, PLoS Biology 2011

Parameter Summary is Found at the Bottom of the Output…..



Search Parameters	
Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	2

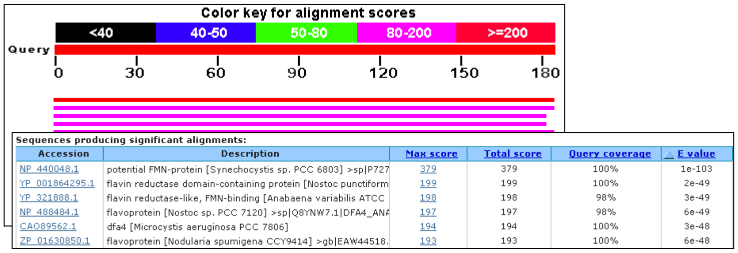
Database	
Posted date	Sep 6, 2010 4:42 AM
Number of letters	4,014,994,744
Number of sequences	11,756,863
Entrez query	none

Karlin-Altschul statistics		
Lambda	0.319424	0.267
K	0.13352	0.041
H	0.397413	0.14

Results Statistics	
Length adjustment	129
Effective length of query	54
Effective length of database	2498359417
Effective search space	134911408518
Effective search space used	134911408518

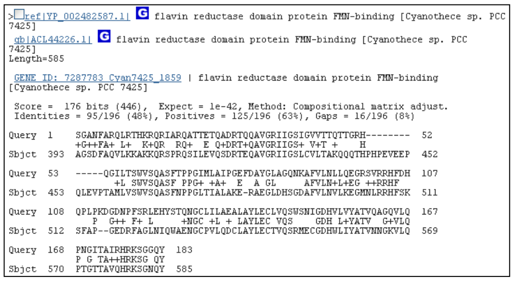
Kerfeld and Scott, PLoS Biology 2011

Evaluating BLAST Results



Sequences producing significant alignments:

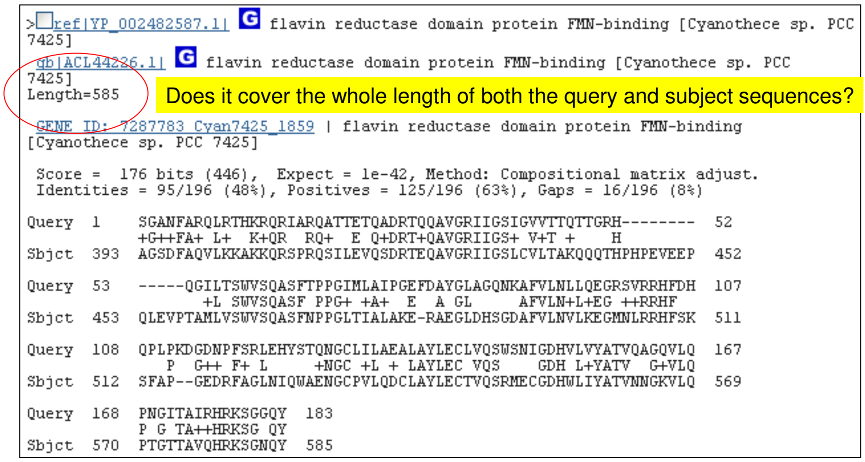
Accession	Description	Max score	Total score	Query coverage	E value
NP_440048.1	potential FMN-protein [Synechocystis sp. PCC 6803] >sp|P727	379	379	100%	1e-103
YP_001864295.1	flavin reductase domain-containing protein [Nostoc punctiform	199	199	100%	2e-49
YP_321888.1	flavin reductase-like, FMN-binding [Anabaena variabilis ATCC	198	198	98%	3e-49
NP_488484.1	flavoprotein [Nostoc sp. PCC 7120] >sp|Q8YNW7.1|DFA4_ANA	197	197	98%	6e-49
CAO89562.1	dfa4 [Microcystis aeruginosa PCC 7806]	194	194	100%	3e-48
ZP_01630850.1	flavoprotein [Nodularia spumigena CCY9414] >gb|EAW44518.	193	193	100%	6e-48



```
>☐ref|YP_002482587.1| G flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
 gb|ACL44226.1| G flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
Length=585

 GENE ID: 7287783 Cyan7425_1859 | flavin reductase domain protein FMN-binding
[Cyanothece sp. PCC 7425]

 Score =  176 bits (446),  Expect = 1e-42, Method: Compositional matrix adjust.
 Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)

Query  1    SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTGRH--------  52
            +G++FA+ L+  K+QR  RQ+  E Q+DRT+QAVGRIIGS+ V+T +    H
Sbjct  393  AGSDFAQVLKKAKKQRSPRQSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP  452

Query  53   -----QGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRSVRRHFDH  107
                  +L SWVSQASF PPG+ +A+  E  A GL      AFVLN+L+EG ++RRHF
Sbjct  453  QLEVPTAMLVSWVSQASFNPPGLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK  511

Query  108  QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVLQ  167
               P  G++ F+ L      +NGC +L + LAYLEC VQS     GDH L+YATV  G+VLQ
Sbjct  512  SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVNNGKVLQ  569

Query  168  PNGITAIRHRKSGGQY  183
            P G TA++HRKSG QY
Sbjct  570  PTGTTAVQHRKSGNQY  585
```

Kerfeld and Scott, PLoS Biology 2011

Examine the BLAST Alignment



Does it cover the whole length of both the query and subject sequences?

```
>☐ref|YP_002482587.1| G flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
 gb|ACL44226.1| G flavin reductase domain protein FMN-binding [Cyanothece sp. PCC
7425]
Length=585

 GENE ID: 7287783 Cyan7425_1859 | flavin reductase domain protein FMN-binding
[Cyanothece sp. PCC 7425]

 Score =  176 bits (446),  Expect = 1e-42, Method: Compositional matrix adjust.
 Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)

Query  1    SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTTQTTGRH--------  52
            +G++FA+ L+  K+QR  RQ+  E Q+DRT+QAVGRIIGS+ V+T +    H
Sbjct  393  AGSDFAQVLKKAKKQRSPRQSILEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP  452

Query  53   -----QGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRSVRRHFDH  107
                  +L SWVSQASF PPG+ +A+  E  A GL      AFVLN+L+EG ++RRHF
Sbjct  453  QLEVPTAMLVSWVSQASFNPPGLTIALAKE-RAEGLDHSGDAFVLNVLKEGMNLRRHFSK  511

Query  108  QPLPKDGDNPFSRLEHYSTQNGCLILAEALAYLECLVQSWSNIGDHVLVYATVQAGQVLQ  167
               P  G++ F+ L      +NGC +L + LAYLEC VQS     GDH L+YATV  G+VLQ
Sbjct  512  SFAP--GEDRFAGLNIQWAENGCPVLQDCLAYLECTVQSRMECGDHWLIYATVNNGKVLQ  569

Query  168  PNGITAIRHRKSGGQY  183
            P G TA++HRKSG QY
Sbjct  570  PTGTTAVQHRKSGNQY  585
```

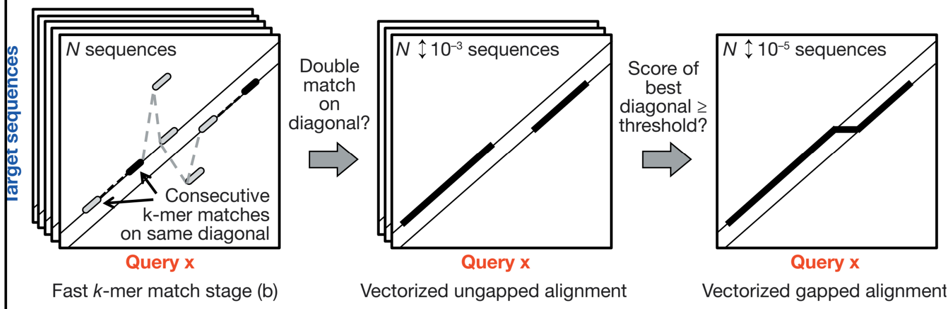
Kerfeld and Scott, PLoS Biology 2011

High E-value: Discovery of a Distant Homolog or Garbage?

- Take another look at the target (subject) sequence(s) that have high E-values
  - Similar length?
  - Recurring motifs?
  - Similar biological functions?
- Use target sequences as query sequences for another BLAST search
  - Does the original query sequence come up in report?

Kerfeld and Scott, PLoS Biology 2011

MMSeqs2 = speeding up BLAST-style database searches by >200X

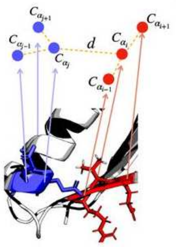

Target sequences
N sequences
Consecutive k-mer matches on same diagonal
Query x
Fast *k*-mer match stage (b)
Double match on diagonal?
N ↕ 10⁻³ sequences
Query x
Vectorized ungapped alignment
Score of best diagonal ≥ threshold?
N ↕ 10⁻⁵ sequences
Query x
Vectorized gapped alignment

Uses a combination of parallelization and clever pre-filtering:
“MMseqs2 searching is composed of three stages: a short word (‘*k*-mer’) match stage, vectorized ungapped alignment, and gapped (Smith–Waterman) alignment. The first stage is crucial for the improved performance. For a given query sequence, it finds all target sequences that have two consecutive similar-*k*-mer matches on the same diagonal.”

Steinegger & Söding, *Nature Biotech* 35:1026–1028 (2017)

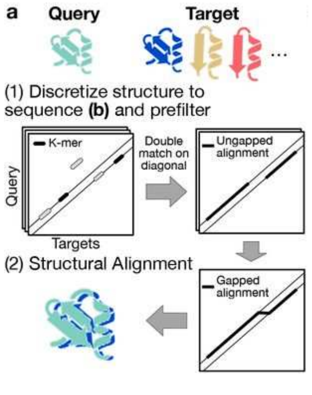
How might you perform fast 3D structure- structure matching instead of sequence-sequence matching?

The current best algorithm to compare a protein’s 3D structure to a database of 3D structures operates like BLAST/MMSeqs2



FoldSeek 1st converts 3D structures to sequences of characters representing 3D neighborhoods

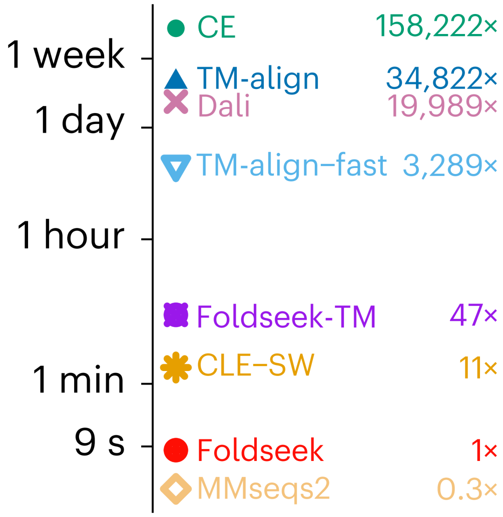
→ 20 “3Di” states instead of 20 amino acids



then finds matching k-mers in database targets & builds alignments around those

https://www.nature.com/articles/s41587-023-01773-0

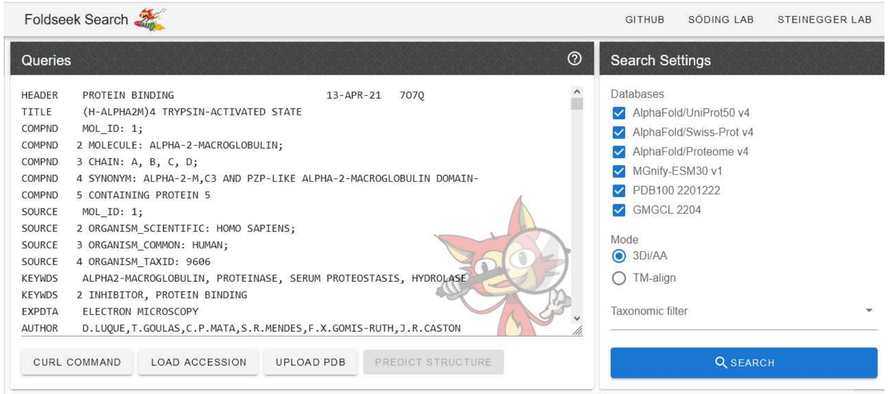
FoldSeek is orders of magnitude faster at finding similar full-length 3D protein structures in large databases



Time to search 100 proteins against the full 200M structure AlphaFold database

https://www.nature.com/articles/s41587-023-01773-0

You can try it out at https://search.foldseek.com & search >800M protein structures



```
HEADER    PROTEIN BINDING                         13-APR-21   7O7Q
TITLE     (H-ALPHA2M)4 TRYPSIN-ACTIVATED STATE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: ALPHA-2-MACROGLOBULIN;
COMPND   3 CHAIN: A, B, C, D;
COMPND   4 SYNONYM: ALPHA-2-M,C3 AND PZP-LIKE ALPHA-2-MACROGLOBULIN DOMAIN-
COMPND   5 CONTAINING PROTEIN 5
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606
KEYWDS    ALPHA2-MACROGLOBULIN, PROTEINASE, SERUM PROTEOSTASIS, HYDROLASE
KEYWDS   2 INHIBITOR, PROTEIN BINDING
EXPDTA    ELECTRON MICROSCOPY
AUTHOR    D.LUQUE,T.GOULAS,C.P.MATA,S.R.MENDES,F.X.GOMIS-RUTH,J.R.CASTON
```

Search Settings

Databases
☑ AlphaFold/UniProt50 v4
☑ AlphaFold/Swiss-Prot v4
☑ AlphaFold/Proteome v4
☑ MGnify-ESM30 v1
☑ PDB100 2201222
☑ GMGCL 2204

Mode
● 3Di/AA
○ TM-align

Taxonomic filter

(courtesy of the Steinegger lab @ steineggerlab.com,
who also make great illustrations for all their programs)