

BLAST

**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  
GEEQTEFDVVLTGAGDNKVAATKAVRGATGLGLKEAKSAVESAPFTLKEG  
VSKEEAETLANELKEAGIEVEVK
```

Nucleotide sequence:

```
>gi|118139508:333094-333465 Thiomicrospira crunogena XCL-2  
ATGGCAATTACAAAAGACGATATTTAGAACGAGTTGCTAACATGTCAGTAATGGAAG  
TTGTTGAACCTGTTGAAGCAATGGAAGAGAAGTTGGTGTCTGCAGCAGCAGTTGC  
GGTTGCAGGTCCTGCAGGTGATGCTGGCGCTGCTGGTGAAGAACAAACAGAGTTTGAC  
GTTGTCTTGACTGGTGTGTTGACAACAAAGTTGCAGCAATCAAAGCCGTTTCGTGGCG  
CAACTGGTCTTGGGCTTAAAGAAGCGAAAAGTGCAGTTGAAAAGTGCACCATTTACGCT  
TAAAGAGGGTGTCTTAAAGAAGAAGCAGAAAAGTCTTGCAAATGAGCTTAAAGAAGCA  
GGTATTGAAGTCGAAGTTAAATAA
```

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

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

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

Query Sequence:

```
>gi|16329320 (residues 412 to 594)
SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTQTG
RHQGILTSWVSQASFTPPGIMLAIPGEFDAYGLAGQNKAFVLNLLQEGRS
VRRHFDHQPLPKDGNPFPSRLEHYSTQNGCLILABALAYLECLVQSWNSI
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
R	5	0	-1	-1	-2	1
G	6	-4	-2	-3	0	-2
I		4	-3	0	-2	-1
K			5	-3	0	-1
F				6	-2	-2
S					4	1
T						5
W						
V						

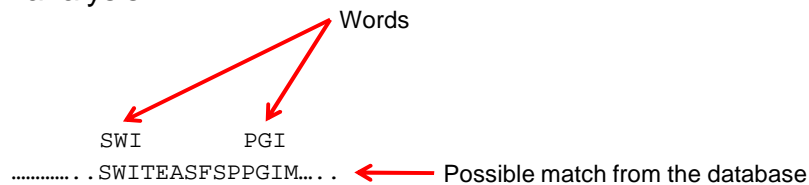
SWV (4+1+1+4 = 19)	} Synonyms above threshold 11... (others not shown)
SWI (4+1+1+3 = 18)	
TWV (1+1+1+4 = 16)	
GWV (0+1+1+4 = 15)	
KWV (0+1+1+4 = 15)	
SWS (4+1+1-2 = 13)	} Synonyms below threshold 11... (others not shown)
SEV (4+1+4 = 9)	
SRV (4-3+4 = 5)	

*A portion of the BLOSUM 62 matrix

- 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

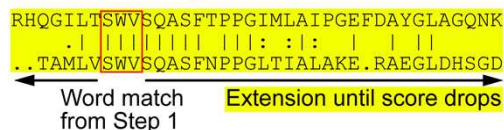
BLAST Phase 2: Using the query sequence word list, scan the database for synonyms (hits)

- 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



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

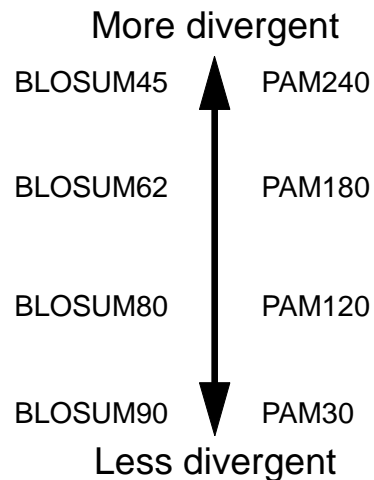


So, to summarize:

- 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

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



Raw Scores (S values) from an Alignment

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

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

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

Going from Raw Scores to Bit Scores

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

where

S' = bit score

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

(as in 0 vs 1)

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

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

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)

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

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

BLAST 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

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

Kerfeld and Scott, PLoS Biology 2011

Filter and Mask

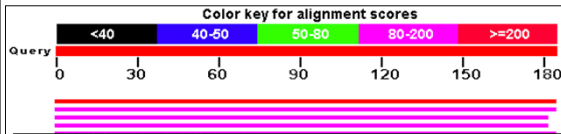
- Filter: Low complexity
 - Replaces the following with N (nucleotides) or X (amino acids)
 - Dinucleotide repeats
 - Amino acid repeats
 - Leader sequences
 - Stretches of hydrophobic residues
- Mask: Lower case
 - Replaces lowercase letters in sequence with N or X
 - Lowercase letters typically indicate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

The screenshot shows the 'Algorithm parameters' window for BLAST. Under the 'Filters and Masking' section, the 'Filter' checkbox is checked, and the 'Mask' checkbox is unchecked. Red arrows point to these two checkboxes from the text on the left. Other parameters include 'Max target sequences' set to 100, 'Expect threshold' set to 10, and 'Word size' set to 3.

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	

Evaluating BLAST Results



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
NP_440048.1	potential FMN-protein [Synecocystis sp. PCC 6803] >sp P727	373	379	100%	1e-103
YP_001864295.1	flavin reductase domain-containing protein [Nostoc punctiforme]	192	199	100%	2e-49
YP_321893.1	flavin reductase-like, FMN-binding [Anabaena variabilis ATCC	190	198	99%	3e-49
NP_438454.1	flavoprotein [Nostoc sp. PCC 7120] >sp O81YW7.1 DFA4_AN	157	197	98%	6e-49
CAO95562.1	fla4 [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 [Cyanothecae sp. PCC 7425]
gb|ACL44226.1| G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
Length=585
GENE ID: 7287783 Cyan7425_1859 | flavin reductase domain protein FMN-binding [Cyanothecae 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  SGANFARQLRTHKRQRIARQATTETQADRTQAVGRIGSIGVVTQTGRH----- 52
          +G++FA+ L+ K+QR RQ+ E Q+DR+QAVGRIGS+ V+T + H
Sbjct 393  AGSDFAVLKKAKKQKRSRQSIQVSDRTEQAVGRIGSLCVLTAQKQQTTHPPEVEEP 452
Query 53  -----QGILTSVVSQASFTPPGIMLAIPEGFDAYLAGQNKAFVNLLEQGRSVRRHFDH 107
          +L SVVSQASF PPG+ +A+ E A GL AFVNL+L+EG ++RRHF
Sbjct 453  QLEVPTAMLVSVVSQASFPVGLTIALAKE-RAEGLDHSGDAFVNLVLEKGMMLRRHFSK 511
Query 108 QPLPKDCGNPFSRLEHYSTQNGCLLAEALAYLECLVQSWSNIGDHVLYATVQAGQVLQ 167
          P G++ F+ L +NGC +L+ LAYLEC VQS GDH L+YATV G+VQ
Sbjct 512  SFAP--GEDRFAGLNIQWENGCPVLQDCLAYLECTVQSRMECGDHVLYATVQAGQVLQ 569
Query 168 PNGITAIRHRKSGGQY 183
          P G TA++HRKSG QY
Sbjct 570  FTGTTAVQHRKSGNQY 585
```

Examine the BLAST Alignment

```
>|ref|YP_002482587.1| G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
gb|ACL44226.1| G flavin reductase domain protein FMN-binding [Cyanothecae sp. PCC 7425]
Length=585
GENE ID: 7287783 Cyan7425_1859 | flavin reductase domain protein FMN-binding [Cyanothecae 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  SGANFARQLRTHKRQRIARQATTETQADRTQAVGRIGSIGVVTQTGRH----- 52
          +G++FA+ L+ K+QR RQ+ E Q+DR+QAVGRIGS+ V+T + H
Sbjct 393  AGSDFAVLKKAKKQKRSRQSIQVSDRTEQAVGRIGSLCVLTAQKQQTTHPPEVEEP 452
Query 53  -----QGILTSVVSQASFTPPGIMLAIPEGFDAYLAGQNKAFVNLLEQGRSVRRHFDH 107
          +L SVVSQASF PPG+ +A+ E A GL AFVNL+L+EG ++RRHF
Sbjct 453  QLEVPTAMLVSVVSQASFPVGLTIALAKE-RAEGLDHSGDAFVNLVLEKGMMLRRHFSK 511
Query 108 QPLPKDCGNPFSRLEHYSTQNGCLLAEALAYLECLVQSWSNIGDHVLYATVQAGQVLQ 167
          P G++ F+ L +NGC +L+ LAYLEC VQS GDH L+YATV G+VQ
Sbjct 512  SFAP--GEDRFAGLNIQWENGCPVLQDCLAYLECTVQSRMECGDHVLYATVQAGQVLQ 569
Query 168 PNGITAIRHRKSGGQY 183
          P G TA++HRKSG QY
Sbjct 570  FTGTTAVQHRKSGNQY 585
```

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

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?