Week 2: Sequence alignment & search

Substitution, BLAST

- Substitution matrix
  - Construction & properties
- Fast sequence searches
  - BLAST; Statistics of similarity search

Substitution matrix to measure similarity in sequence alignments

Dr. Margaret Dayhoff

Applying math & computational techniques to the sequencing of proteins and nucleic acids.

- 1965: First collection of protein seqs.
- Single-letter code for amino acids.
- 1966: ‘Evolutionary trees’.
- **1978: First AA similarity-scoring matrix.**
- 1980: Launched the Protein Information Resource, the first online database system that could be accessed by telephone line.

Substitution matrix: A collection of scores for aligning nucleotides or amino acids with one another.

- The scores represent the relative ease with which one nucleotide or amino acid may mutate into or substitute for another.
- Purely statistical, nothing directly to do with structure/biochemistry.

Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

Wikipedia; Eddy (2004)

Substitution matrix to measure similarity in sequence alignments

Substitution matrix: Each score is a log-odds score equal to the logarithm of the ratio of the likelihoods of two hypotheses: i) the residues can substitute for one another, or ii) not.

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

- p_{ab}: likelihood of these two residues being correlated because they're homologous.
  - p_{ab} are the target frequencies: the probability that we expect to observe residues *a* and *b* aligned in homologous sequence alignments.
- $f_a f_b$: likelihood of these two residues being uncorrelated and unrelated, occurring independently.
  - f_a and f_b are background frequencies: the probabilities that we expect to observe amino acids *a* and *b* on average in any protein sequence.
- λ: a scaling factor, usually set to something that lets helps round off all the terms in the score matrix to sensible integers.

Eddy (2004)

Substitution matrix to measure similarity in sequence alignments

Substitution matrix: Each score is a log-odds score equal to the logarithm of the ratio of the likelihoods of two hypotheses: i) the residues can substitute for one another, or ii) not.

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

- p_{ab}: likelihood of these two residues being correlated because they're homologous.
- $f_a f_b$: likelihood of these two residues being uncorrelated and unrelated, occurring independently.
- λ: a scaling factor

Assuming that each aligned residue pair is statistically independent of the others (biologically dubious, but mathematically convenient):

- The score of an alignment ("**alignment score**") = sum of individual log-odds scores for each aligned residue pair.

Eddy (2004)

Substitution matrix to measure similarity in sequence alignments

BLOSUM (BLOcks SUbstitution Matrix) for protein sequence alignment.

- Scan the BLOCKS database for very conserved regions of protein families (w/o gaps in the alignment).
- Count the relative frequencies of AA and their substitution probabilities.
- Calculate a log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids.

Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

Henikoff & Henikoff (1992)

Substitution matrix to measure similarity in sequence alignments

Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	$\boxed{4}$									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	$\boxed{11}$		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

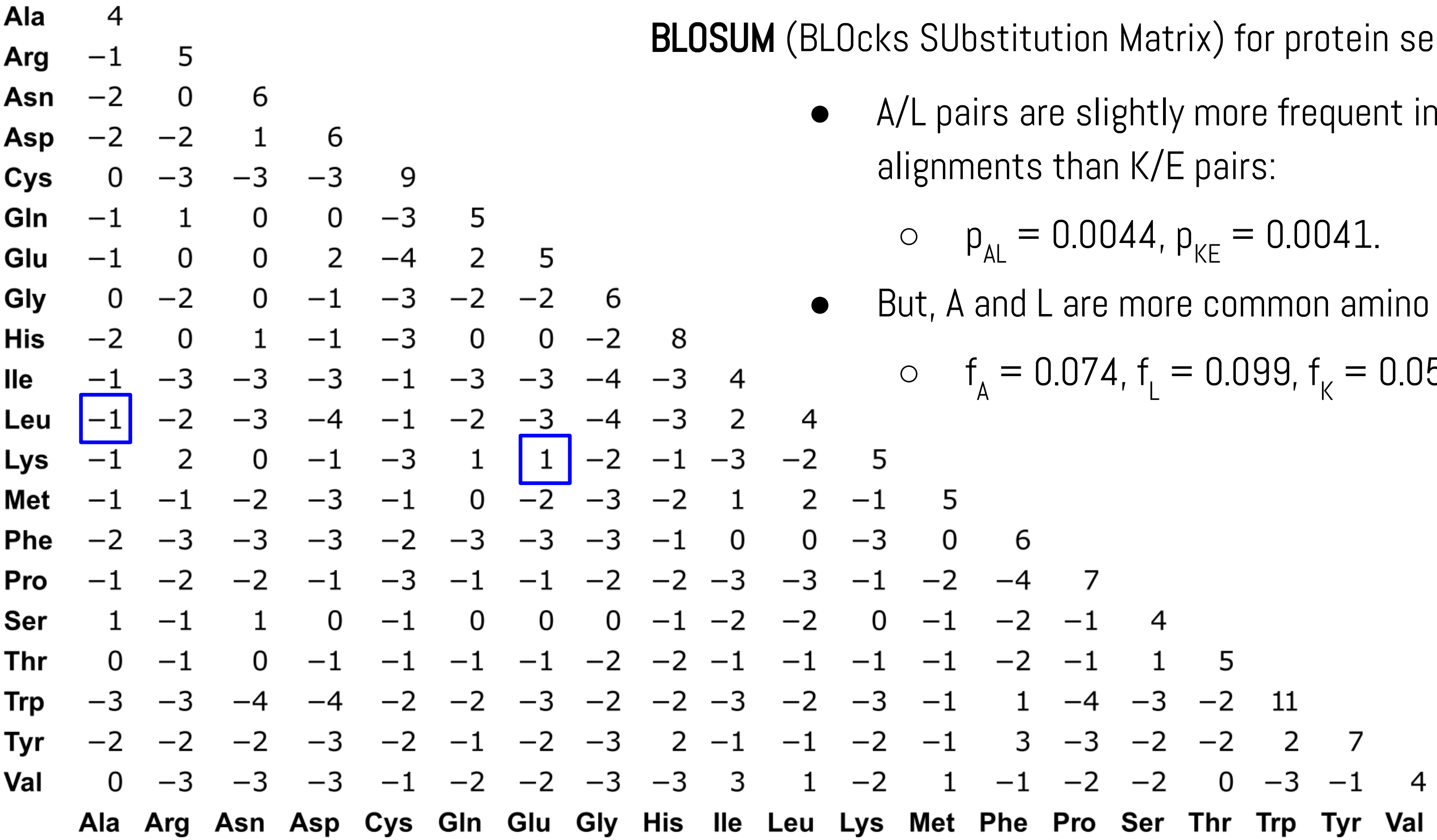
BLOSUM (BLOcks SUbstitution Matrix) for protein sequence alignment.

- The rarer the amino acid is, the more surprising it would be to see two of them align together by chance.
- L/L more common than WW:
  - ○ $p_{LL} = 0.0371$, $p_{WW} = 0.0065$
- W is a much rarer amino acid:
  - ○ $f_L = 0.099$, $f_W = 0.013$.

Check with $\lambda = 0.347$.

Eddy (2004)

Substitution matrix to measure similarity in sequence alignments



Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

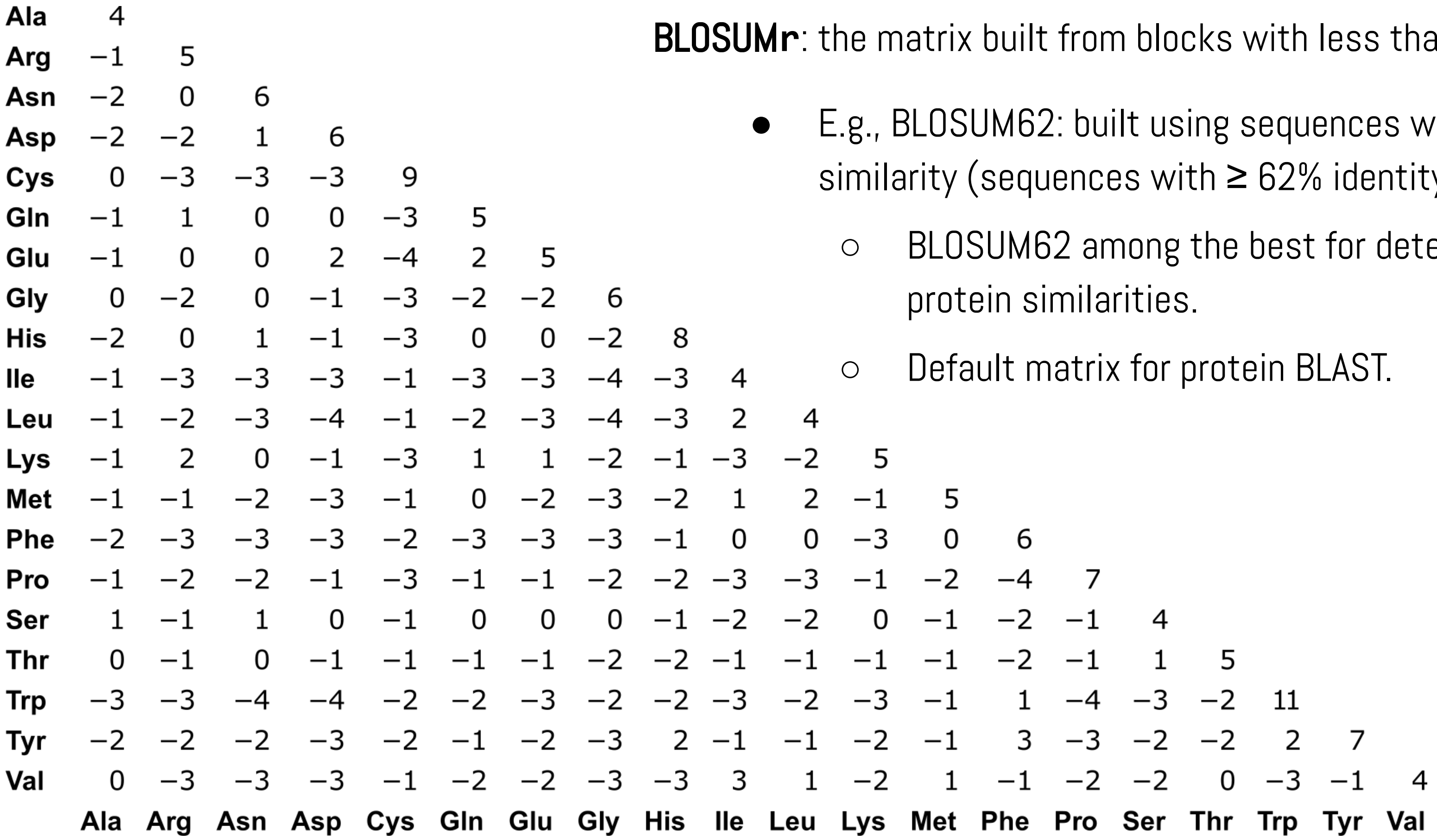
BLOSUM (BLOcks SUbstitution Matrix) for protein sequence alignment.

- A/L pairs are slightly more frequent in homologous alignments than K/E pairs:
  - $p_{AL} = 0.0044$, $p_{KE} = 0.0041$.
- But, A and L are more common amino acids:
  - $f_A = 0.074$, $f_L = 0.099$, $f_K = 0.058$, $f_E = 0.054$.

Check with $\lambda = 0.347$.

Eddy (2004)

Substitution matrix to measure similarity in sequence alignments



Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

BLOSUMr: the matrix built from blocks with less than r% of similarity.

- E.g., BLOSUM62: built using sequences with less than 62% similarity (sequences with ≥ 62% identity were clustered)
  - BLOSUM62 among the best for detecting most weak protein similarities.
  - Default matrix for protein BLAST.

Eddy (2004)

Substitution matrix to measure similarity in sequence alignments

Substitution matrix for DNA

A				
T				
G				
C				
	A	T	G	C

Making-up an arbitrary matrix by fixing the p_{ab} values ⟶ directly describes what homologous alignments are expected to look like.

- The resulting score matrix is optimal for detecting alignments that match these target frequencies.

Say, the matrix should be optimized for finding 88% identity alignments.

- Assume that all mismatches are equiprobable, and composition of both alignments and background sequences is uniform at 25% for each nucleotide (f_a, f_b = 0.25 for all a,b). Then,
  - Four identities: $p_{aa} = 0.22$
  - 12 types of mismatch: $p_{ab} = 0.01$.
- If we set $\lambda = 1$, this gives +1.26 for a match and −1.83 for a mismatch.
- Setting $\lambda = 0.25$ and round off: we have a new scoring system of +4/−7.

Eddy (2004)

Substitution matrix to measure similarity in sequence alignments

Substitution matrix for DNA

A				
T				
G				
C				
	A	T	G	C

Given a scoring matrix, we can back calculate target frequencies if two conditions are met:

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

1. It must have at least one positive score, and
2. The expected score for random sequence alignments must be negative.

True for most score matrices:

- These properties are necessary to make local sequence alignment algorithms like BLAST and Smith-Waterman work.
- Both conditions are met by definition for matrices derived as log-odds scores, except for the useless case of $p_{ab} = f_a f_b$ for all a,b.

Examples:

- FASTA & WU-BLASTN: arbitrary +5/−4 scoring system;
  Optimal for detecting alignments that are 65% identical.
- NCBI BLASTN: +1/−2 scoring system; Optimal for detecting alignments that are 95% identical.

Eddy (2004)

How do we scale this up to search an entire sequence database?

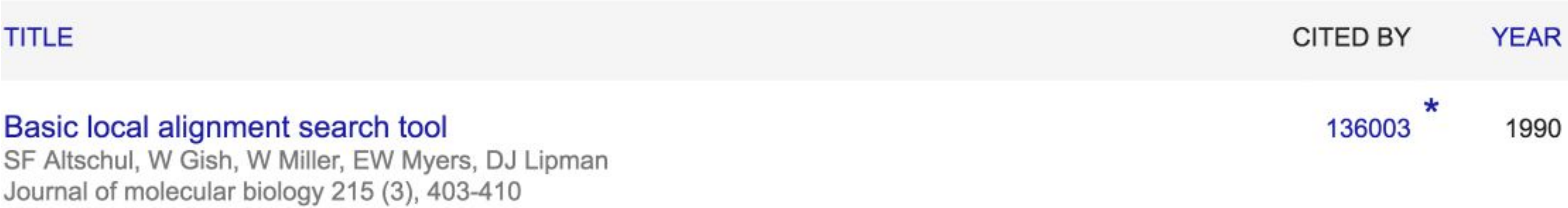
Given a query sequence, and a large set of target sequences (millions), which target sequences (if any) are related to the query?

- Individual alignments need not be perfect: Once initial matches are found, they can fine-tune them later.
- Must be very fast.

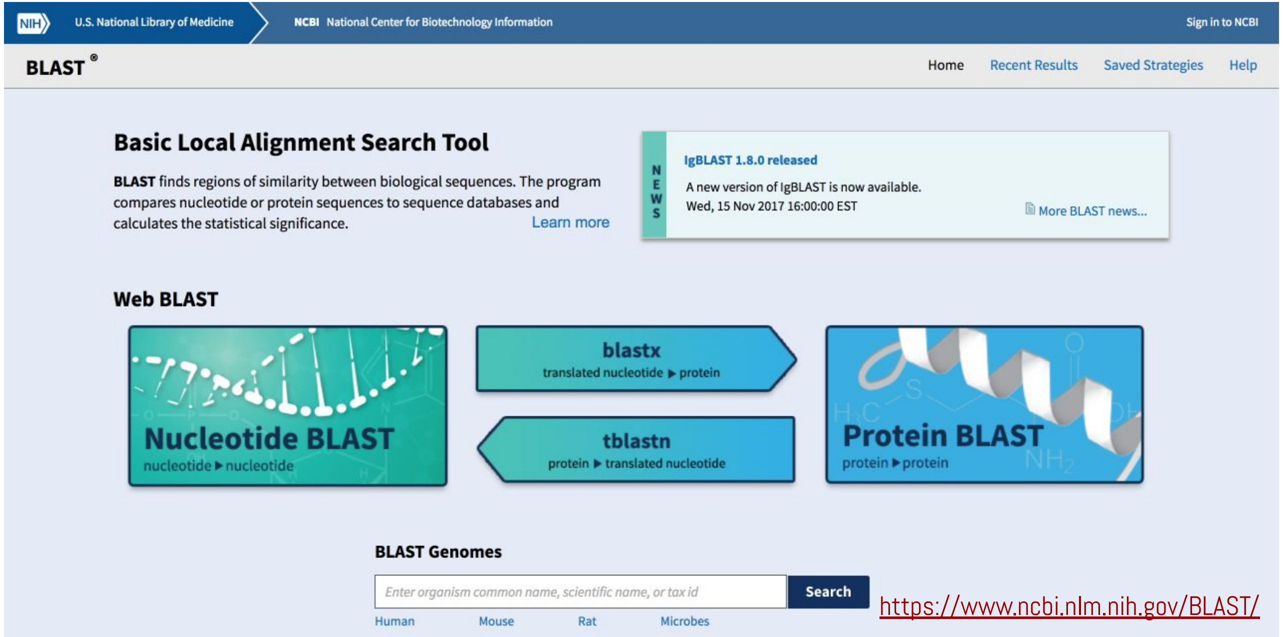
Exploit the nature of the problem (most sequences will be unrelated to the query):

- If any match with % identity ≤ 90 is going to be rejected, can ignore sequences which don't have a stretch of 10 nucleotides in a row.
- Pre-screen sequences for common long stretches.
- Pre-process the database offline and index k-mers.

BLAST



TITLE	CITED BY	YEAR
Basic local alignment search tool SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman Journal of molecular biology 215 (3), 403-410	136003 *	1990



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST ®
Home
Recent Results
Saved Strategies
Help

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. Learn more

NEWS

IgBLAST 1.8.0 released

A new version of IgBLAST is now available.
Wed, 15 Nov 2017 16:00:00 EST

More BLAST news...

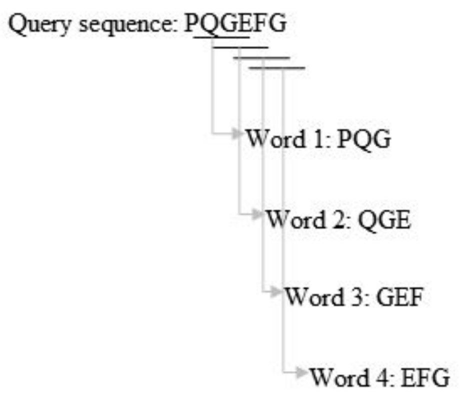
Web BLAST

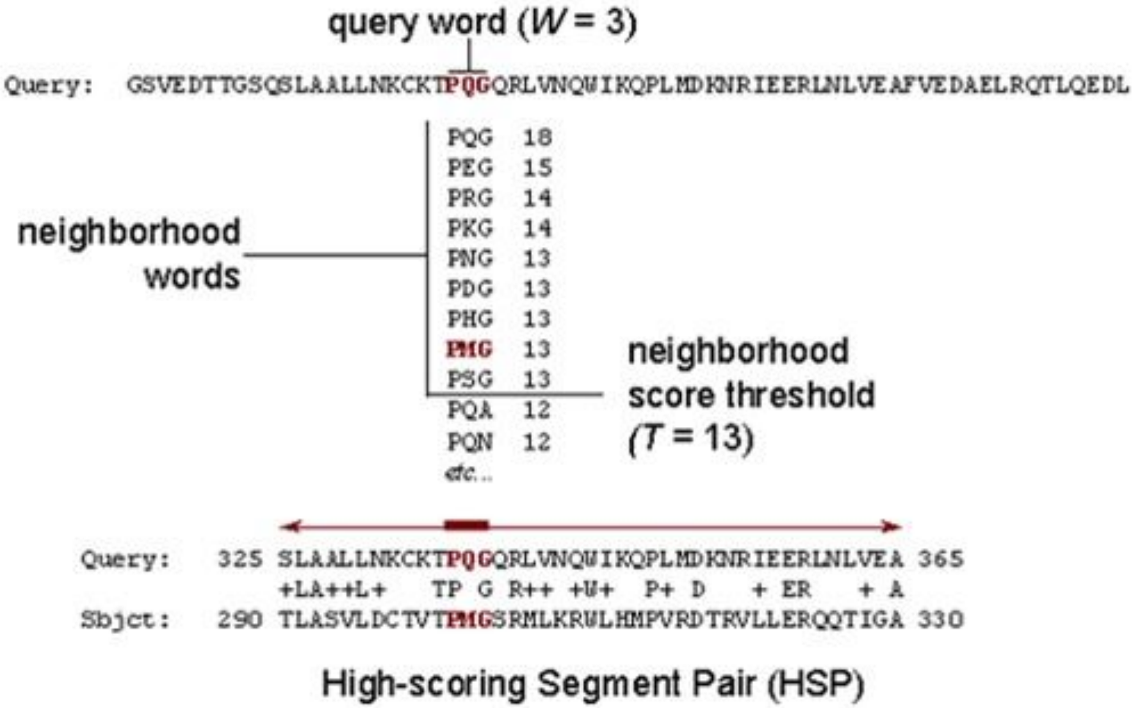
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein

BLAST Genomes

https://www.ncbi.nlm.nih.gov/BLAST/

BLAST





query word (W = 3)
neighborhood words
neighborhood score threshold (T = 13)
High-scoring Segment Pair (HSP)

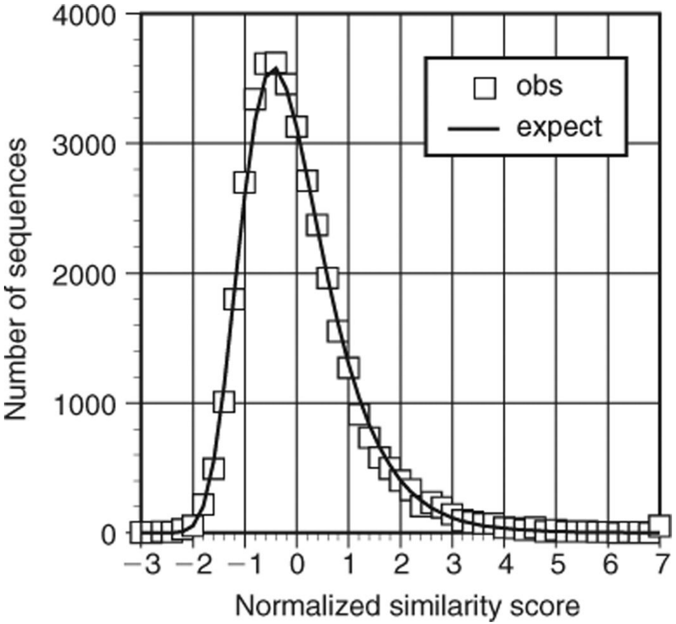
```
Query:   GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

         PQG  18
         PEG  15
         PRG  14
         PKG  14
         PNG  13
         PDG  13
         PHG  13
         PMG  13
         PSG  13
         PQA  12
         PQN  12
         etc...

Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

NCBI Handbook: https://www.ncbi.nlm.nih.gov/books/NBK153387/

Some uses of BLAST

- Finding the right/relevant species:
  - If you have a DNA sequence from unknown species, BLAST can help identify the correct/related species.
- Finding protein domains:
  - If you a protein sequence (or a translated nucleotide sequence), BLAST can be used to look for known protein domains in the query sequence.
- Mapping the phylogeny of a gene/protein:
  - BLAST can be used to find potential homologs of your gene/protein of interest across many species, which you can then use to generate a phylogenetic tree.
- Mapping DNA to a known chromosome:
  - If you are sequencing a gene from a known species but have no idea of the chromosome location, BLAST can help you. BLAST will show you the position of the query sequence in relation to the hit sequences.
- Annotations:
  - BLAST can also be used to map gene/protein annotations from one organism to another.

Statistics of similarity search



Distribution of real (squares) & expected similarity scores (Gumbel extreme value distribution).

P-value:

- The probability of observing a score equal to or greater than the observed score S.

E-value:

- The expected number of HSPs with score at least S.
- $E = Kmne^{-\lambda S}$

Database E-value:

- E-value after thousands/millions of searches ≈ E*D.

Bit score:

- Normalized raw score.