Bioinformatics Lunch Seminar (Summer 2014)

UNIVERSITY of WISCONSIN
UWMILWAUKEE
School of Freshwater Sciences

- Every other Friday at 1 pm. 20-30 minutes plus discussion
- Informal, ask questions anytime, start discussions
- Content will be based on feedback
- Targeted at broad audience of various levels of backgrounds and education
- Emphasis on Genomics Center

Contact:
Raymond Hovey
Genomics Center - SFS
rhovey@uwm.edu
414-382-1774
http://www.greatlakesgenomics.uwm.edu/

Pairwise sequence alignment

- Is the most fundamental operation of bioinformatics
- It is used to decide if two proteins (or genes) are related structurally or functionally
- It is used to identify domains or motifs that are shared between proteins
- It is used in the analysis of genomes

Overview of Sequence Alignment

Sequence alignment is the procedure of comparing
two (pairwise alignment)
or
more (multiple sequence alignment)
sequences by searching characters that are in the same order in the sequences.
(order may not change, insertions and deletions are common, rearranging AAs is not)

Sequence alignment is useful for discovering functional, structural, and evolutionary information.

Sequences that are very much alike, or "similar" may have the same function.

Why Do We Align Sequences?

The basic idea of aligning sequences is that similar sequences generally produce similar proteins.

Therefore, if the sequences are similar they are *likely* to be functionally similar proteins.

To predict the characteristics of a protein using only its sequence data, sequence similarity of an unknown sequence with a protein of characterized function can be used to identify the function of the unknown protein

If one of them has known structure, then the alignment gives some insight about the structure of the other sequence.

How do we tell whether two macromolecules are similar?

Sequence Comparisons via Alignments provide an indication of similarity of two sequences.

Furthermore, sequence comparison and alignments permit to map functional information (such as the location of secondary structural elements, domains, active sites, and regulatory regions) present in a well-studied molecule to those that might be present in a new sequence.

The **General Approach** involves the use of a set of algorithms such as **BLAST** programs to compare a query sequence to all the sequences in a specified database.

- Comparisons are made in a pairwise fashion.
- Each comparison is given a score reflecting the degree of similarity between the query and the sequence being compared.
- The higher the score, the greater the degree of similarity.
- The similarity is measured and shown by aligning two sequences.
- In general, alignments can be global or local (this is algorithm specific).
  - A global alignment is an optimal alignment that includes all characters from each sequence, whereas a local alignment is an optimal alignment that includes only the most similar local region or regions.
- Discriminating between real and artifactual matches is done using an estimate of probability that the match might occur by chance.
- Similarity, by itself, cannot be considered a sufficient indicator of function but can be useful as a lead to understanding function.

Comparision that can be done (BLAST programs)

- DNA:DNA or RNA:RNA or DNA:RNA
- polypeptide:polypeptide
- DNA translated in all 6 frames to polypeptide
- DNA translated in all 6 frames to all database DNA sequences translated in all 6 frames

Local vs. global alignments

- Global Alignment

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || | | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

- Local Alignment—better alignment to find conserved segment

```
                  tccCAGTTATGTCAGgggacacgagcatgcagagac
                     ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Local vs. global alignments

Local Alignments

• Two genes in different species may be similar over short conserved regions and dissimilar over remaining regions.

Example:

Homeobox genes have a short region called the *homeodomain* that is highly conserved between species.

• A global alignment would not find the homeodomain because it would try to align the ENTIRE sequence.

• Better for comparing sequences of different lengths.

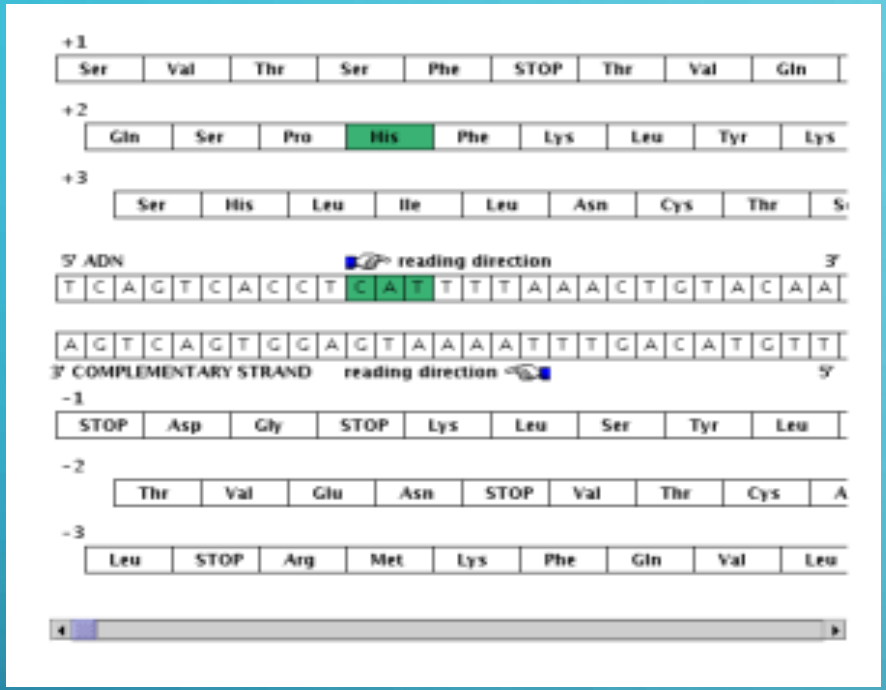
(like checking PCR primers against a genome)

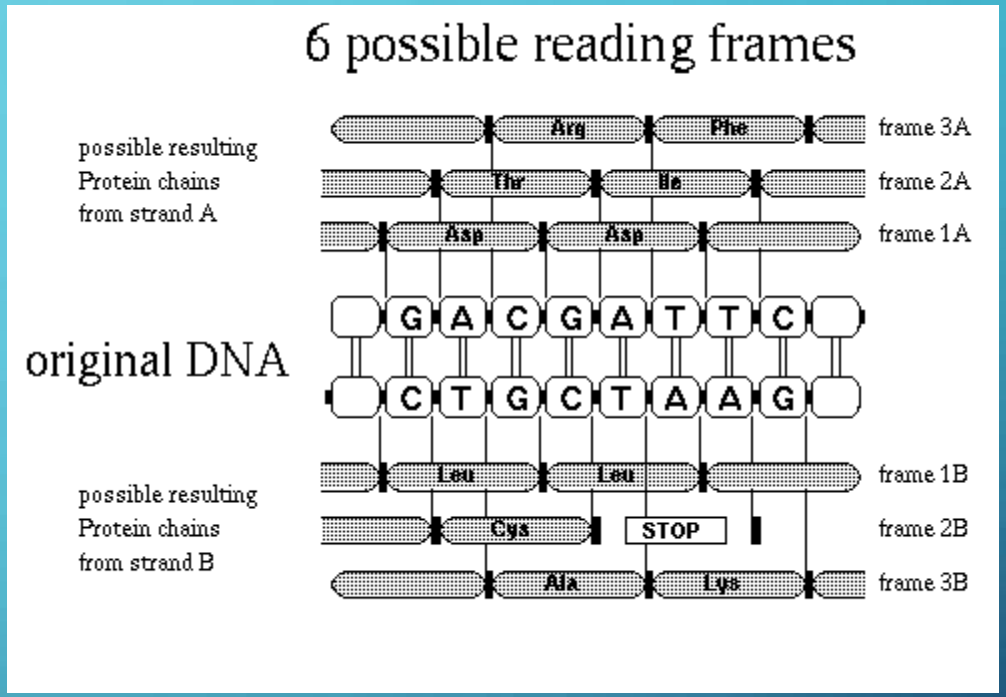
Global alignment

• Finds best possible alignment **across entire length** of 2 sequences

• Aligned sequences assumed to be generally similar over entire length

DNA translated in all 6 frames





Pairwise Alignment

Global

- Best score from among alignments of full-length sequences
- Needleman-Wunsch algorithm (best alignment, dynamic programming)

Local

- Best score from among alignments of partial sequences
- Smith-Waterman algorithm (best alignment, dynamic programming)

http://www.ebi.ac.uk/Tools/psa/

The pros and cons of dynamic programming

These Algorithms will always give you the best possible alignment.

Due to their dynamic programming nature they are really slow
and should not be used with multiple sequences (only pairwise).

Example:
It takes 100^2 seconds to perfectly align 2
protein Sequences of 100 aa each.
Than it takes 100^3 seconds to align 3 sequences.
100^4 for 4 sequences
1.90258×10^{34} Years for 20 sequences

BLAST (Biological Local Alignment Search Tool)

- BLAST programs use a heuristic search algorithm.
- The programs use the statistical methods of Karlin and Altschul.
- BLAST programs were designed for fast database searching, with minimal sacrifice of sensitivity for distantly related sequences.
- The programs search databases in a special compressed format.
- It is possible to use one's private database with BLAST. (required to convert it to the BLAST format)
- Great for comparing sequences of different lengths. (like checking PCR primers against a genome)

Why use BLAST?

BLAST searching is fundamental to understanding the relatedness of any favorite query sequence to other known proteins or DNA sequences.

Applications include
- **identifying orthologs and paralogs**
- **discovering new genes or proteins**
- **discovering variants of genes or proteins**
- **investigating expressed sequence tags (ESTs)**
- **exploring protein structure and function**

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution.

Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions.

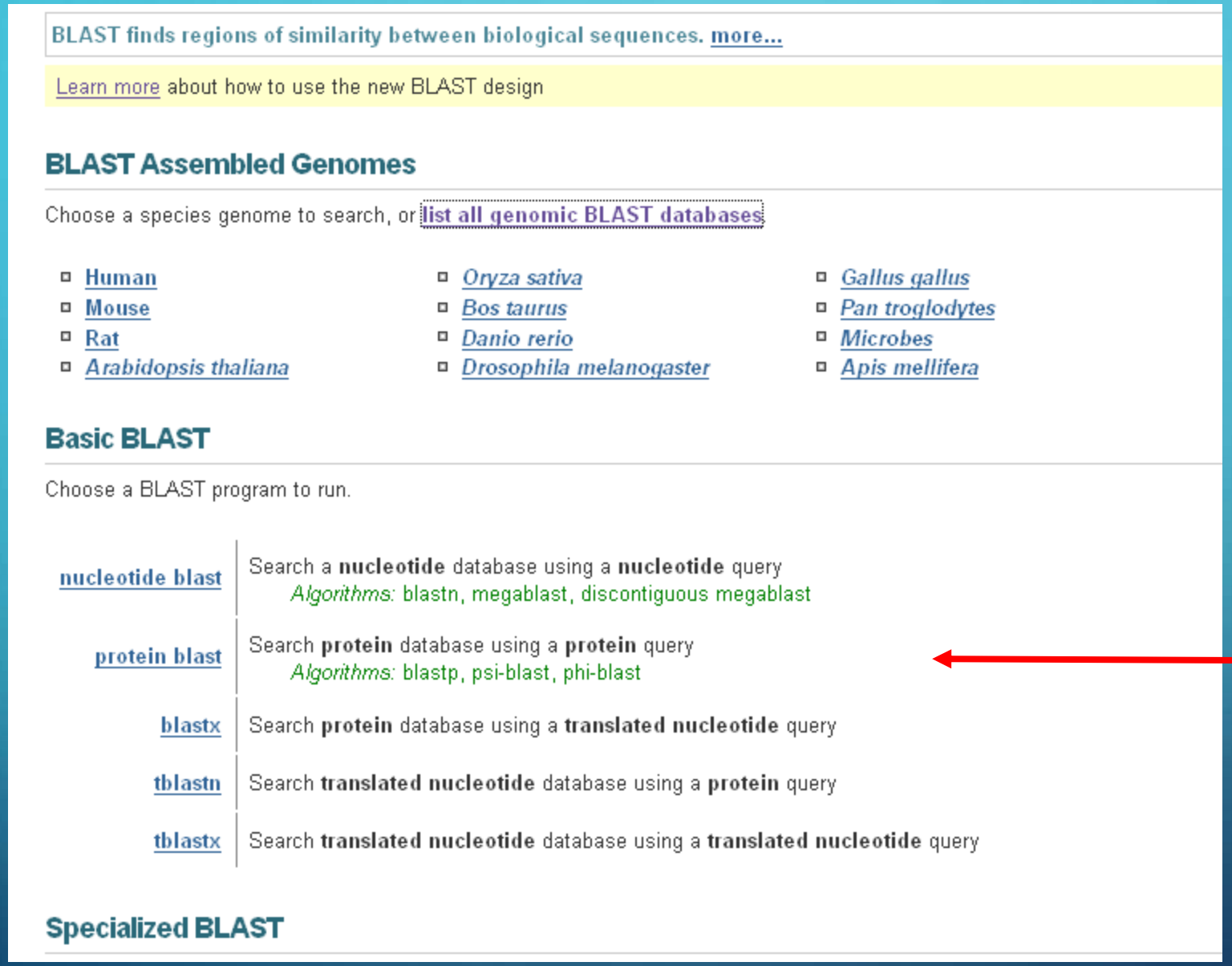
Four components to a BLAST search

(1) Choose the sequence (query)

(2) Select the BLAST program

(3) Choose the database to search

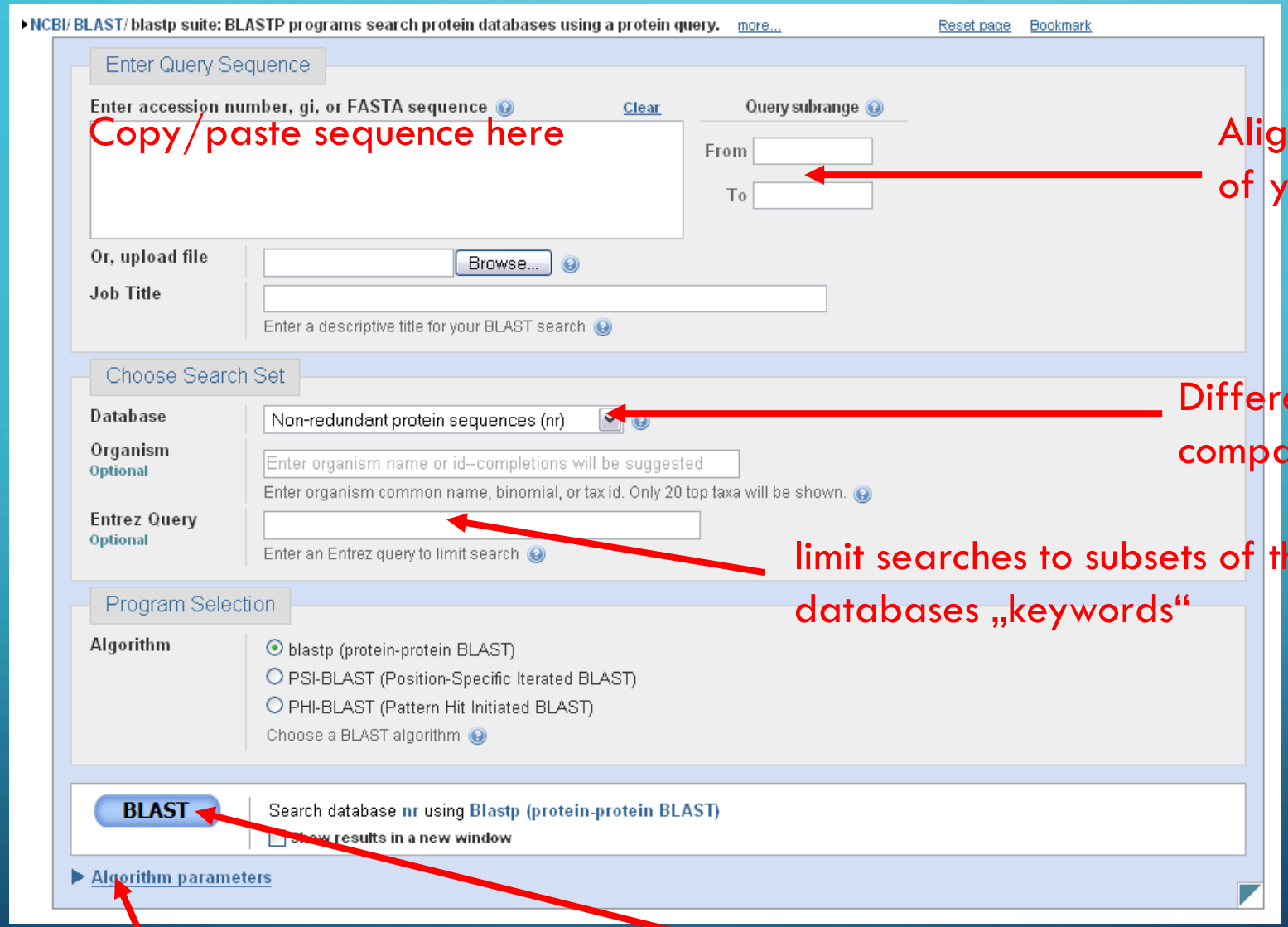
(4) Choose optional parameters



To be discussed first

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

Protein-Protein BLAST (BLASTp): Structure of the NCBI web interface



Align a subsequence of your query only

Different databases for comparison

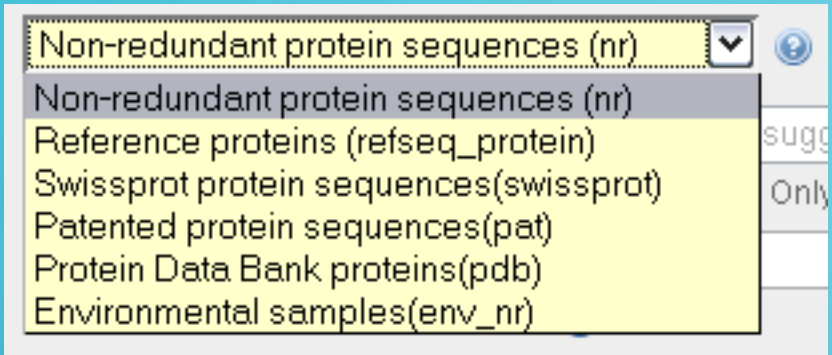
limit searches to subsets of the BLAST databases „keywords“

Start the BLAST

Algorithm Parameters

Open to change matrix, gap/extension penalties,…

Databases for BLAST (Example)



Peptide Sequence Databases

nr

All non-redundant GenBank translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF

refseq

non-redundant set of sequences at NCBI including genomic DNA, transcript (RNA), and protein products, for major research organisms

swissprot

Last major release of the SWISS-PROT protein sequence database

pat

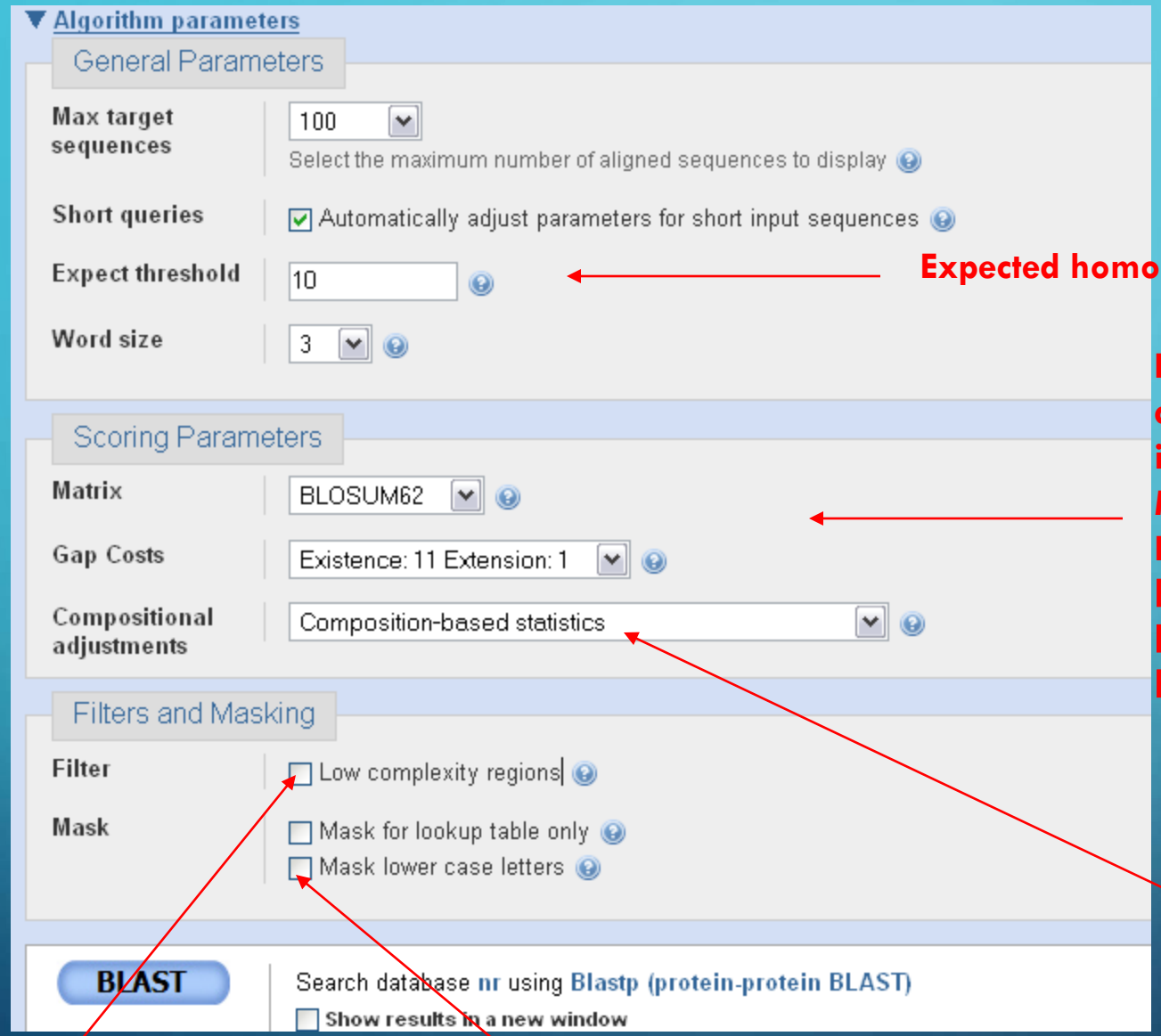
Proteins from the Patent division of GenPept.

pdb

Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank

env_nr

Non-redudant database of environmental sequences



Expected homologs

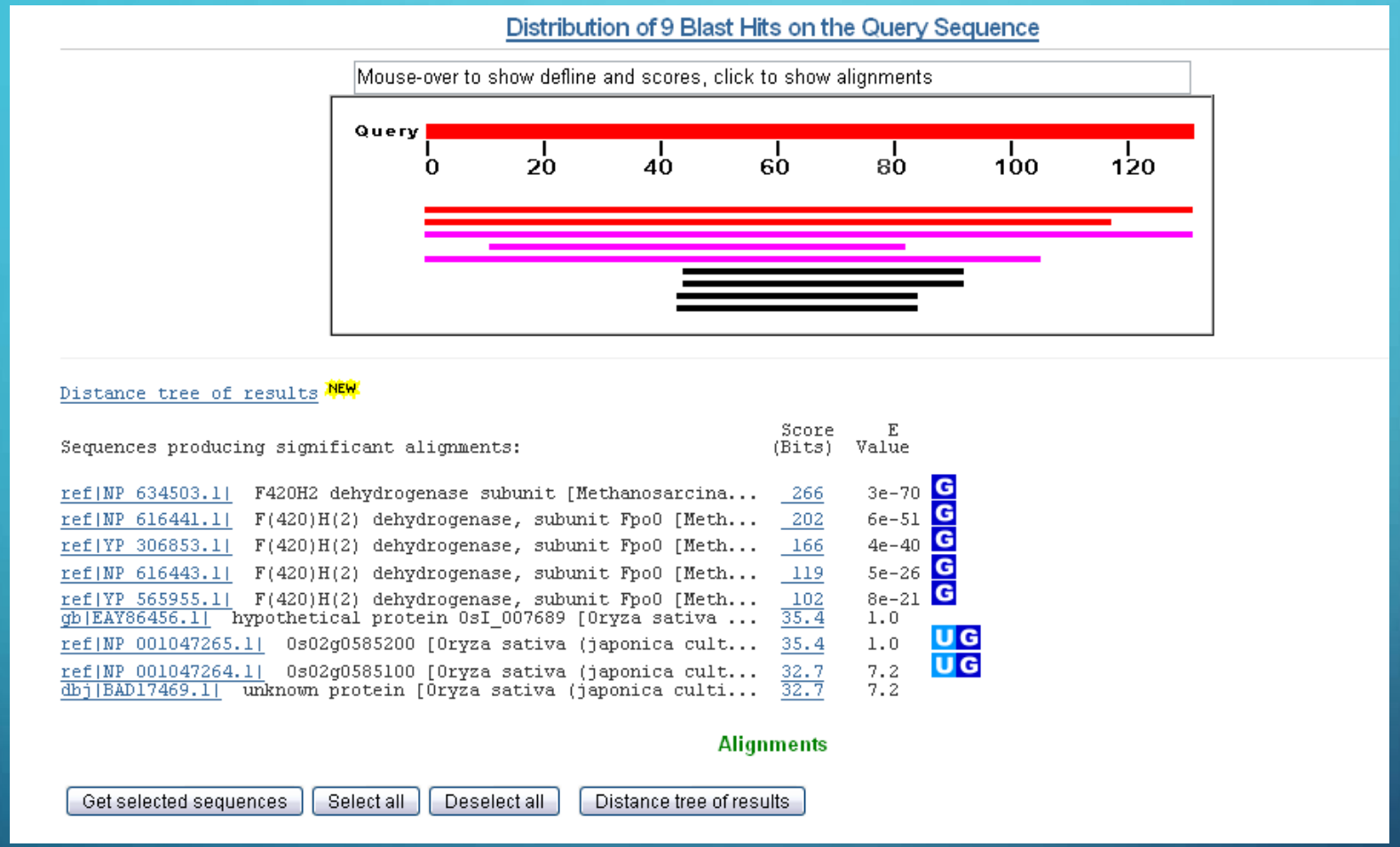
Increasing the Gap Costs will result in alignments which lower number of Gaps introduced. Changes automatically with Matrix.
High Specificity PAM30 Existence: 9, Extension: 1
Low specificity: BLOSUM45 Existence: 15, Extension: 2

Improves E-value accuracy by taking query sequence complexity in consideration and adjusting scoring matrix.

Excludes sequences with low complexity

denote areas you would like filtered with lower case

BLAST Output Example



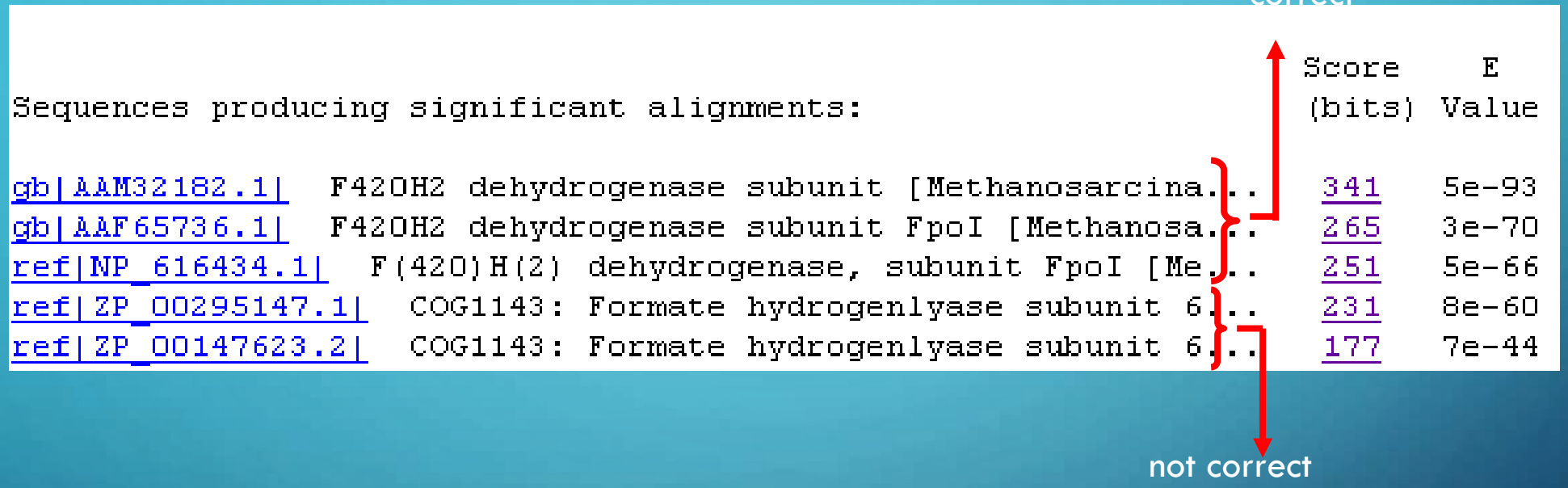
Graphical Overview: An overview of the database sequences aligned to the query sequence is shown. The score of each alignment is indicated by one of five different colors, which divides the range of scores into five groups. At the bottom: Sequences with significant alignments

Link to Entrez

```
>ref|NP_634503.1|  F420H2 dehydrogenase subunit [Methanosarcina mazei Go1]
 gb|AAF65742.1|AF228525_12  F420H2 dehydrogenase subunit Fpo0 [Methanosarcina mazei]
 gb|AAM32175.1|  F420H2 dehydrogenase subunit [Methanosarcina mazei Go1]
Length=131

 Score =  266 bits (679),  Expect = 3e-70, Method: Composition-based stats.
 Identities = 131/131 (100%), Positives = 131/131 (100%), Gaps = 0/131 (0%)

Query  1    MTDCDLCGKGIPTVIPVRTYPPLLRFAYPEGVWKGLCETCLDSAQKTYLEVNRNHTSCRR  60
            MTDCDLCGKGIPTVIPVRTYPPLLRFAYPEGVWKGLCETCLDSAQKTYLEVNRNHTSCRR
Sbjct  1    MTDCDLCGKGIPTVIPVRTYPPLLRFAYPEGVWKGLCETCLDSAQKTYLEVNRNHTSCRR  60

Query  61   GKCSLCGSKTGVFSVELQIPDFSKGIVRKDVDVCYRCLKLVDEAYIRYKREQIEQDHEQG  120
            GKCSLCGSKTGVFSVELQIPDFSKGIVRKDVDVCYRCLKLVDEAYIRYKREQIEQDHEQG
Sbjct  61   GKCSLCGSKTGVFSVELQIPDFSKGIVRKDVDVCYRCLKLVDEAYIRYKREQIEQDHEQG  120

Query  121  RIHGHEHVHPH  131
            RIHGHEHVHPH
Sbjct  121  RIHGHEHVHPH  131
```

Pairwise alignment

Pitfalls in Blast. Wrong annotation of genes/proteins!

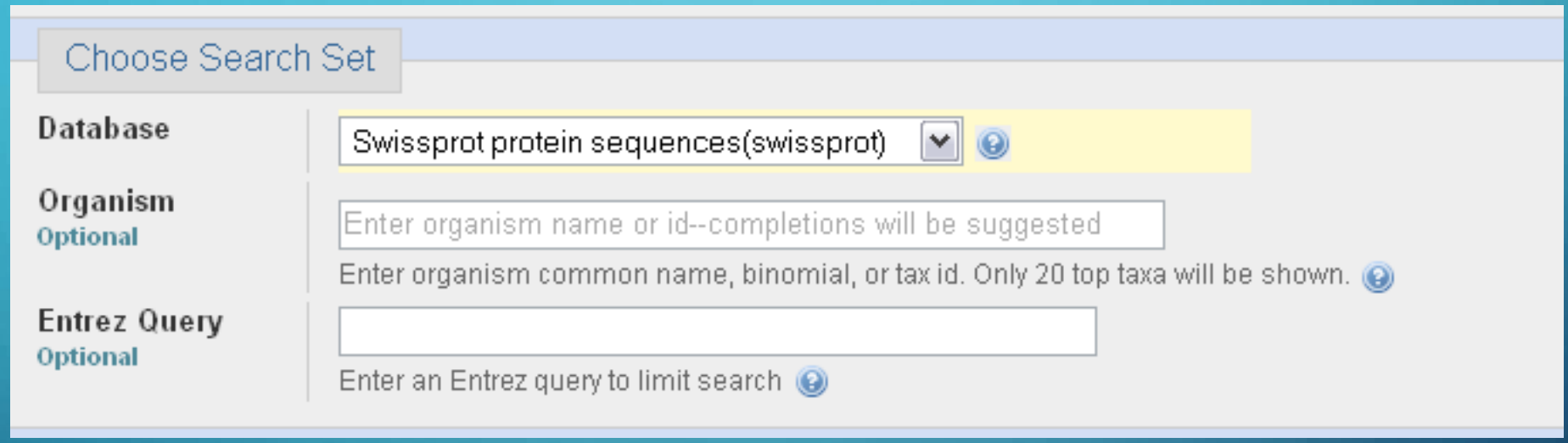


Proteins are 98% identical

First three annotations correct (from Methanosarcina mazei, Ms. acetivorans)– biochemical evidence

No. 4 and 5 mis-annotated – corresponding protein from two closely related methanogenic archaea

Solution: use a curated database like SwissProt



Statistical significance of Blast search: Matrices

The Matrix Option.

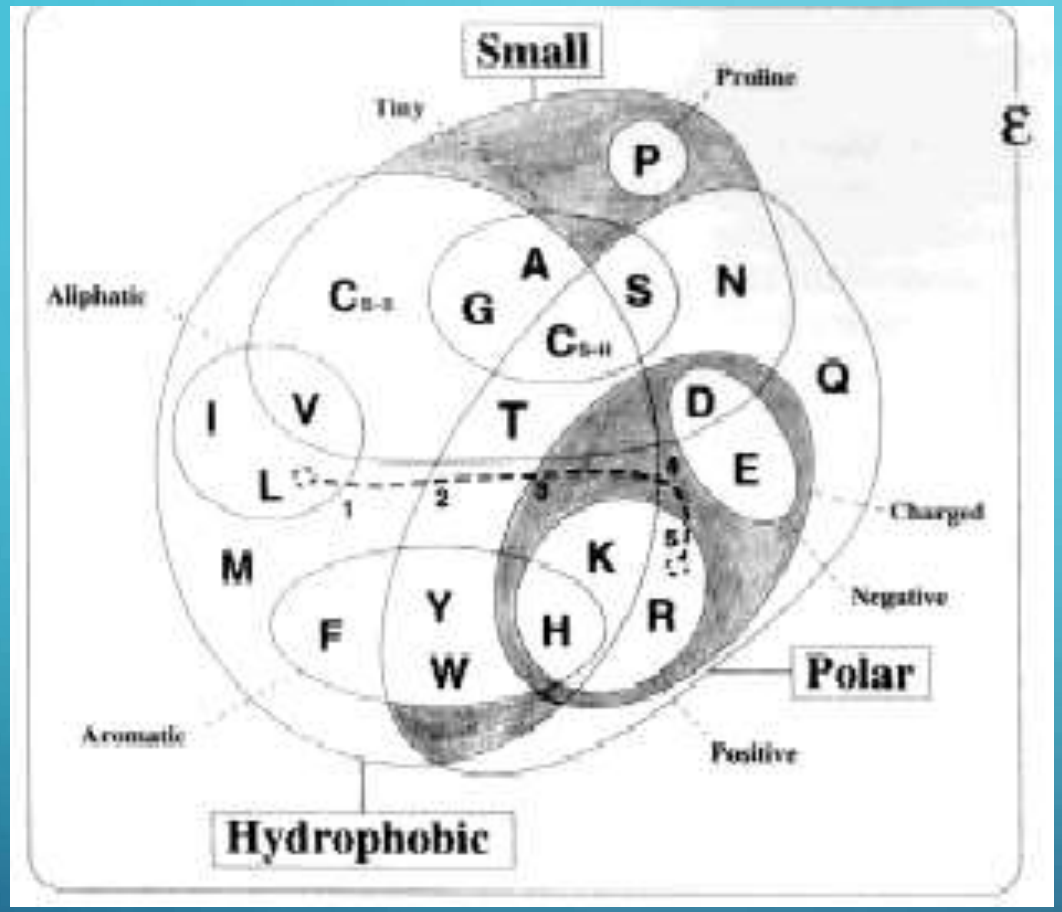
- A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score to aligning any possible pair of residues.

- In general, different substitution matrices are used to detect similarities among sequences that are diverged by differing degrees.

- A single matrix may nevertheless be reasonably efficient over a relatively broad range of evolutionary change.

- The strength of a match is determined by the returned score of the alignment.

How can we evaluate amino acid changes to score them adequately?
-> look at properties of amino acids!



E⟶D 1 border
E⟶Q 2 borders

The 20 natural amino acids are characterized by a set of 10 physico-chemical properties (see above) and are organized according to the overlapping sectors. Lysin (K) for example is positively charged, polar and hydrophobic. Hence, there is a functional similarity to Arginine (R) and also a relationship to the polar and charged amino acid Glutamic Acid(E). To determine the degree of similarity in a position of the alignment the borders are counted that are crossed to connect the amino acid sequences.

Example of an Amino Acid Substitution Scoring Matrix

A	4																		
R	-1	5																	
N	-2	0	6																
D	-2	-2	1	6															
C	0	-3	-3	-3	9														
Q	-1	1	0	0	-3	5													
E	-1	0	0	2	-4	2	5												
G	0	-2	0	-1	-3	-2	-2	6											
H	-2	0	1	-1	-3	0	0	-2	8										
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5							
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Differences in score matrices

Part of the Blosum62 matrix

	A	R	N	D	C	Q	E	G	H	I
A	4	-1	-2	-2	0	-1	-1	0	-2	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3
N	-2	0	6	1	-3	0	0	0	1	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3
E	-1	0	0	2	-4	2	5	-2	0	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4

Part of the Blosum45 matrix

	A	R	N	D	C	Q	E	G	H	I
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1
R	-2	7	0	-1	-3	1	0	-2	0	-3
N	-1	0	6	2	-2	0	0	0	1	-2
D	-2	-1	2	7	-3	0	2	-1	0	-4
C	-1	-3	-2	-3	12	-3	-3	-3	-3	-3
Q	-1	1	0	0	-3	6	2	-2	1	-2
E	-1	0	0	2	-3	2	6	-2	0	-3
G	0	-2	0	-1	-3	-2	-2	7	-2	-4
H	-2	0	1	0	-3	1	0	-2	10	-3
I	-1	-3	-2	-4	-3	-2	-3	-4	-3	5

Different values for substitutions (and gaps)

•Common Substitution Matrix Families

•PAM (Point Accepted Mutation)

Amino acid scoring matrices are traditionally PAM matrices. PAM40 is most sensitive for similar sequences. PAM250 is for more distantly related sequences.

•BLOSSUM (Blocks Substitution Matrix)

•BLOSSUM matrices are most sensitive for local alignment of related sequences.

•BLOSUM62 is optimised for general BLASTP searches, and is suitable for most situations; it will recognise some amino acid substitutions as conservative (e.g. Arg to Lys). It is a good general matrix, set by default for protein BLAST searches

•Caution! You cannot compare the alignment scores (see later) from one matrix directly against the alignment scores from another matrix!

•The following matrices are roughly equivalent...
•PAM100 ==> Blosum90
•PAM120 ==> Blosum80
•PAM160 ==> Blosum60
•PAM200 ==> Blosum52
•PAM250 ==> Blosum45

Statistical significance of Blast search: Scores and E-values

Once BLAST has found a similar sequence to the query in the database, it is helpful to have some idea of whether the alignment is "good" (possible biological relationship), or whether the similarity observed is by chance alone. BLAST uses statistical theory to produce a bit score and expect value (E-value) for each alignment pair.

The raw score (S) gives an indication of how good the alignment is; the higher the score, the better the alignment. The score is calculated from a formula that takes into account the alignment of similar or identical residues, as well as any gaps introduced to align the sequences. Bit scores (S') are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used.

The E-value gives an indication of the statistical significance of a given pairwise alignment. The lower the E-value, the more significant the hit. A sequence alignment that has an E-value of 0.05 means that this similarity has a 5 in 100 (1 in 20) chance of occurring by chance alone. *E* values below 10^{-6} are most probably statistically significant. *E* values above 10^{-6} but below 10^{-3} deserve a second look. *E* values above 10^{-3} should not be considered.

Alignments: Input sequence MM0633

Third hit gives score = 442 and e-value of 1e-122

```
>gi|68212169|ref|ZP_00564014.1|  conserved hypothetical protein [Methanococcoides burtonii DSM
           6242]
 gi|68184324|qb|EAM99061.1|  conserved hypothetical protein [Methanococcoides burtonii DSM
           6242]
           Length=457

 Score =  442 bits (1137),  Expect = 1e-122
 Identities = 218/445 (48%), Positives = 296/445 (66%), Gaps = 11/445 (2%)

Query  5    AGAAEPSGPGDFTSNQFSKSGICSNCHGSSFGEWAGSMHSLADSDFFYNAMLQEYGVAAE  64
            A A +P+G  +  S  F  +  C+NCH     +  GSMH+ A SD  Y    +E  +A+E
Sbjct  17   ASAVDPTGFDELNSEDFVTNSKCANCHAILRSQHEGSMHAFAYSDPLYQ---KEALLASE  73
```

Bit score

Raw score

Determination of scores with Blosum 62

A	G	A	A	E	P	S	G	P	input sequence
A	S	A	V	D	P	T	G	F	homolog
4	0	4	0	2	7	1	6	-4	Sum = Score S

BLAST for Protein query sequences

blastp = Compares a protein sequence with a protein database

If you want to find something about the function of your protein, use **blastp** to compare your protein with other proteins contained in the databases; identify common regions between proteins, or collect related proteins (phylogenetic analysis;

tblastn = Compares a protein sequence with a nucleotide database

If you want to discover new genes encoding proteins (from multiple organisms), use **tblastn** to compare your protein with DNA sequences translated into their six possible reading frames; map a protein to genomic DNA;

BLAST for DNA query sequences

blastn = Compares a DNA sequence with a DNA database;

Mapping oligonucleotides, cDNAs, and PCR products to a genome; annotating genomic DNA; screening repetitive elements; cross-species sequence exploration;

tblastx = Compares a DNA translated into protein with a DNA database translated into protein;

Cross-species gene prediction at the genome or transcript level (ESTs); searching for genes not yet in protein databases;

blastx = Compares a DNA translated into protein with a protein sequence database;

Finding protein-coding genes in genomic cDNA; determining if a cDNA corresponds to a known protein;

Tips

- Use the latest database version.
- As a rule of thumb, run BLAST first, then depending on your results run a finer tool (BLAT, Ssearch, Smith Waterman, Blocks, etc.).
- Whenever possible, use protein or translated nucleotide sequences.
- Splitting large query sequences can help (> 1000 for DNA, > 200 for protein).
- If the query has repeated segments, remove them and repeat the search.

The UCSC and BLAT

- BLAT = Blast Like Alignment Tool
- Optimized for shorter reads (primers and next generation sequences)
- Faster but less homology depth.
- Can link directly into the UCSC browser (and other hosted databases like The Cancer Genome Atlas)
- Supports large amount of queries

https://genome.ucsc.edu/FAQ/FAQblat.html

Bioinformatics Lunch Seminar (Summer 2014)

UNIVERSITY of WISCONSIN
UWMILWAUKEE
School of Freshwater Sciences

- Every other Friday at 1 pm. 20-30 minutes plus discussion
- Informal, ask questions anytime, start discussions
- Content will be based on feedback
- Targeted at broad audience of various levels of backgrounds and education
- Emphasis on Genomics Center

Contact:
Raymond Hovey
Genomics Center - SFS
rhovey@uwm.edu
414-382-1774
http://www.greatlakesgenomics.uwm.edu/

Statistical significance of Blast search: Scores and E-values

Determined by statistical algorithms: score S reflects the similarity of proteins/ genes and is based on the scoring matrix

S (raw score) = Sum of substitutions and gap scores (penalties)

S´ (bit score) = normalized raw score that allows direct comparison of scores from different alignments, scoring matrices in different Blast searches.

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

(λ and K given at the bottom of all blast searches)

E-value: expected value = number of alignments that are expected to occur by change = probability of false-positive results

The E-value is an estimate of how many sequences would score by chance in the database searched. The higher E the higher the probablity that the aligment is made by chance. So we should look to low E values: Rule of thumb: E-values significant below 10^{-3} to 10^{-5}

The E- value is directly connected to S´.

$$E = mn\,2^{-S'}$$

(mn = effective search space; with m = effective length of query; n = effective lenght of database --- numbers given at the bottom of all blast searches).

P-value: other value representing the significance of an alignment

$P = 1 - e^{-E}$ --- $P < 0.05$ traditionally used to define statistical significance

Statistical parameters of blast search (at the end of the results)

```
Gapped
Lambda      K        H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 2905734
Number of Hits to DB: 189399276
Number of extensions: 8306732
Number of successful extensions: 14476
Number of sequences better than 10: 12
Number of HSP's better than 10 without gapping: 1
Number of HSP's gapped: 14481
Number of HSP's successfully gapped: 12
Number of extra gapped extensions for HSPs above 10: 14465
Length of query: 444
Length of database: 1000656685
Length adjustment: 132
Effective length of query: 312
Effective length of database: 617099797
Effective search space: 192535136664
Effective search space used: 192535136664
```

Values for λ and K

Matrix used

Allowed gaps to be inserted into sequence, where they need to be placed, and how long they must be for optimal alignment (see above)

Values for n and m

Used by Blast to calculate E-value (= n x m)