BIOINFORMATICS 1
Sequence Alignment, BLAST and Substitution Matrices

Annotating Sequences

Pairwise Alignments

- **GLOBAL**
  Most simple relationship - spans the whole length of both sequences - aligns the whole *e.g. same protein in 2 different species*
  (ClustalO)

- **LOCAL**
  Alignment of a pair of sequences such that homologous sub-sequences are aligned, surrounded by areas of non-related (and unaligned) sequence
  e.g. align two protein sequences that share a single common domain, align cDNA to genomic where no matches to introns!
  (BLAST)

BLAST (Basic Local Alignment Search Tool) is the most common tool used to search sequence databases. Based on the Smith-Waterman algorithm.
Quick to run and the first stage in identifying potential similarity targets.
The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.

Program name	Query type	Database type
blastn	nucleotide	nucleotide
blastp	amino acid	amino acid
tblastn	amino acid	translated nucleotide
blastx	translated nucleotide	amino acid
tblastx	translated nucleotide	translated nucleotide

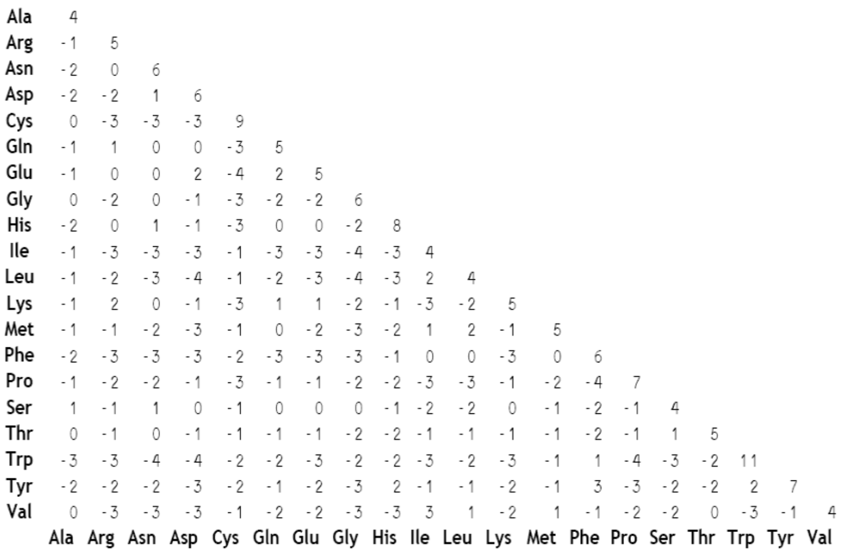
Substitution Matrix

During the course of evolution amino acids at particular locations can change due to mutations.
When aligning protein sequences a non-exact amino acid match can be as informative as a perfect match. It is the nature of the amino acid that is important.
If the chemical/physical properties of that amino acid is the important factor then mutation to another amino acid with the same properties more likely to be maintained.
This homology is the basis of a substitution matrix. Matrices built by selecting a group of similar proteins and scoring based on the observed frequency of the amino acids within the protein.

PAM *(Point Accepted Mutation)*
One of the first amino acid **substitution matrices**, the PAM matrix was developed by Margaret Dayhoff in the 1970s.
This matrix is calculated by observing the differences in closely related proteins.
Set of matrices based on 1572 observed mutations in 71 families of closely related proteins
Higher number in naming scheme donates lower sequence similarity and larger evolutionary distance e.g. PAM30, PAM70

BLOSUM *(BLOck SUbstitution Matrix)*
The PAM matrix, created by comparing closely related species, does not work very well when aligning evolutionarily divergent sequences.
This problem was rectified by the BLOSUM matrix, which uses multiple alignments of evolutionarily divergent proteins.
The probabilities used in the matrix calculation are computed by looking at "blocks" of conserved sequences found in multiple protein alignments.
These conserved sequences are assumed to be of functional importance within related proteins.

Negative score – unlikely to happen (eg Gly/Leu -4)
Positive score – conservative substitution (eg Lys/Arg 2)
High score for identical matches – rare amino acids (eg Cys, Trp)



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**

Log Odds
The logarithm (usually the natural logarithm) of the odds of success

The probability of event E occurring is p
The probability of E not occurring is 1-p

The log odds on E is $\log_e(p/(1-p))$

BLAST scoring matrices values are derived from log odds ratios:
log2 (observed frequency of amino acid pairing/
expected random frequency of pairing)

Note: BLAST scoring matrices use base 2 and not natural logarithms

Deriving Matrix Scores

Suppose the frequencies of methionine (M) and leucine (L) in the data set are 0.01 and 0.1, respectively.
By random pairing 1/1000 (0.001) amino acid pairs would be M-L.
If the observed frequency of pairing is 1/500 (0.002) then the odds ratio is 2/1:
observed/expected = 0.002/0.001 = 2
Converting this to a base 2 logarithm gives a log odds (lod) score of +1, or 1 bit.
log2 (2) = 1

Another example:
Frequency of arginine (R) is 0.1 and leucine 0.1
Actual frequency of pairing (observed) with R-L is 1/500 = 0.002
Random frequency (expected) would be 0.1 * 0.1 = 0.01
Odds ratio is therefore:
observed/expected = 0.002/0.01 = 0.2
Converting this to a base 2 logarithm gives:
log2 (0.2) = -2.322

Simplifying Scores

Lod scores are real numbers but for simplicity are represented as integers in scoring matrices
To retain precision the scores are multiplied by a scaling factor before being converted to integers
For example, a lod score of -1.609 may be scaled by a factor of 2 (-3.218) and then rounded off to an integer value of -3
Scores that have been scaled and converted to integers have no units quantity and are called *raw scores*

The BLOSUM MATRICES raw scores are $\text{int}(\log_2 *2)$

Lambda and Bit Scores

Raw score can be a misleading quantity because scaling factors are arbitrary
A normalized score (bit score), corresponding to the original lod score, is therefore a more useful measure
Converting a raw score to a normalized score requires a matrix-specific
constant called lambda (or λ)
Lambda is approximately the inverse of the original scaling factor, but its value may be slightly different due to integer rounding errors

Gap Scores

In order to build the best possible alignment BLAST introduces gaps where necessary

From an evolutionary perspective insertions/deletions are costly so a penalty is applied when a gap is introduced

The overall score is calculated using the chosen matrix and the following rules:

-11 for opening a gap

-1 to extend a gap

(For nucleotide alignments: +1 for a nucleotide match , -3 for a mismatch,
-5 for opening a gap, -2 to extend a gap)

Interpreting BLAST Output

Score: The score is calculated by incrementing for matches/similarities and decrementing for mismatches gaps.

Identities: The number of residues that are identical in the alignment.

Positives: The number of similar residue matches in the alignment.

Gaps: The number of gaps in the alignment. Gaps are introduced where required to give the best overall alignment.

Does the alignment cover all/most of the query and subject or only part?

Former may identify the sequence if the alignment is perfect, latter shows a homologous region.

BLAST E-value

The **E**, or **Expect**, value is the most important reported statistic

It is a measure of how reliable the alignment is, or how likely it is to be correct

It is the expected number of alignments of sequences of this length with at least the given score that would be found by chance in a search of the database used

The E-value takes in to account the size of the query sequence, the length of the alignment and overall score and the size of the queried database

A good E-value is one that is less than 1e-2 but 1e-4 is better

Alignment tables expose patterns of amino acid conservation, from which distant relationship may be more reliable detected. To be informative a MSA should contain a distribution of closely and distantly related sequences. If all sequences are very closely related, information is redundant and few interferences can be drawn. If all sequences are distant related, difficult to contruct an alignment.

PROFILES

Profiles express the patterns inherent in a multiple sequence alignment of a set of homologous sequences.

- ⇒ They permit greater accuracy in alignments of distantly related sequences
- ⇒ Sets of residues that are highly conserved are likely to be part of the active site, and give clues to function
- ⇒ The conservation patterns facilitate identification of other homologous sequences.

To use profile patterns to identify homologues, the basic idea is to match the query sequences from the database against the sequences in the alignment table, giving higher weight to positions that are conserved than to those that are variable.

PSI-BLAST

It is a program that searches a data bank for sequences similar to query sequence. It begins with a one-at-a-time search. It then derives pattern information from a MSA of the initial hits and reprobes tha database using the pattern. Then it repeats the process, fine tuning the pattern at each cycle.
Workflow:

1. Probe each sequence in the chosen database independently for local regions of similarity to the query sequence, using a BLAST-type search but allowing gaps
2. Collect significant hits. Construct a MSA table between the query sequence and the significant local matches
3. Form a profile from the MSA
4. Reprobe the database with the profile, still looking only for local matches
5. Decide which hits are statistically significant and retain these only
6. Go back to step 2 until a cycle produces little or no change. This accounts for the “iterated” in the title of the PSI-BLAST program

PSI-BLAST, using iterated pattern searching, is much more powerful than simple pairwise BLAST in picking up distant relationships. PSI-BLAST correctly identifies three times as many homologues as BLAST in the region below 30% identity.
It is a good method for analysing whole genomes.

HIDDEN MARKOV MODELS (HMM)

A HMM is a computational structure for describing the subtle patterns that define families of homologous sequences. HMMs are powerful tools for detecting distant relatives and prediction of protein folding pattern. They solely rely on sequences (not structures at all)

The dynamics of the system is such that only the current state influences the choice of its successor: the system has no “memory” of its history. Only the succession of characters emitted is visible; the state sequence that generated the characters remains internal to the system (hidden).