Aligning proteins

Lukas Käll
lukask@kth.se

Why aligning protein rather than DNA sequences?

- Amino acid sequences are often more conserved than their underlying DNA. That is synonymous mutations are more common than expected by studying the mutation frequencies of non-synonymous ones.

Why aligning protein rather than DNA sequences?

- Amino acid sequences are often more conserved than their underlying DNA. That is synonymous mutations are more common than expected by studying the mutation frequencies of non-synonymous ones.
- Even non-synonymous mutations are more frequently causing shifts to amino acids with similar properties (polarity, size) than expected by studying the frequencies of mutations to amino acids with different properties.

Scoring functions for amino acid sequences.

In principle one could use the same type of score functions as for DNA sequences. However, we can create better scoring systems by using *score matrices*, i.e. score functions that are dependent on which amino acids that are evaluated.

Scoring Matrices

There are two major types of scoring matrices:

- PAM = Percentage Accepted Mutations (Margeret Dayhoff)
- BLOSUM = Blocks Substitution Matrix (Henikoff & Henikoff)

PAM

- Created from global alignments, from tertiary structures.
- Better for global alignments.
- Higher numbers indicates suitability for more diverse sequences.

BLOSUM45 $\sim$ PAM250

BLOSUM

- Created from local alignments, from blocks of similar sequences (the BLOCKS DB)
- Better for local alignments.
- Lower numbers indicates suitability for more diverse sequences.

BLOSUM62

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**

What is the probability that one amino acid is replaced by another?

When scoring a position in an alignment containing the amino acid a and b, we take interest in the ratio between the probability that they appear together if they stem from homologue sequences and if they do not stem from homologues.

$$\frac{\Pr(a, b|\text{homologues})}{\Pr(a, b|\text{not homologues})} = \frac{\Pr(a, b)}{\Pr(a)\Pr(b)}.$$

Substitution scores in score matrices

In scoring matrices this property is used in the following form

$$d(a,b) = \frac{1}{\lambda} \log \frac{\Pr(a,b)}{\Pr(a)\Pr(b)}.$$

Here λ is selected in a manner that the $d(a,b)$'s can be rounded to integer value with as little rounding errors as possible.

The approximate reasoning behind the scores

For the full length sequences we are interested in evaluating

$$\frac{\mathsf{Pr}(\text{ Sequence alignment given the sequences are homologues })}{\mathsf{Pr}(\text{Sequence alignment given the sequences are not homologues })} \approx$$

$$\approx \frac{\prod_i \mathsf{Pr}(\text{align pos } i | \text{homologues})}{\prod_i \mathsf{Pr}(\text{align pos } i | \text{not homologues})} \approx \prod_i \frac{\mathsf{Pr}(a_i, b_i)}{\mathsf{Pr}(a_i)\,\mathsf{Pr}(b_i)} =$$

$$= \exp\left(\log\left(\prod_i \frac{\mathsf{Pr}(a_i, b_i)}{\mathsf{Pr}(a_i)\,\mathsf{Pr}(b_i)}\right)\right) = \exp\left(\sum_i \log\left(\frac{\mathsf{Pr}(a_i, b_i)}{\mathsf{Pr}(a_i)\,\mathsf{Pr}(b_i)}\right)\right)$$

This resembles $\exp\left(\sum_i d(a_i, b_i)\right)$, and hence it makes sense to score alignments based on the sums of $d(a_i, b_i)$.

Thanks!