Bioinformatics – Lecture Notes

Announcements

Campus Bookstore has more books.

Seminar – "Quality Control in Manufacturing Oligo Arrays"
Professor Charlie Colbourn (over 230 journal papers)
Friday February 1, 11: am
EC 2.112

Correction – PAM stands for Point Accepted Mutation

Class 4

1. Point Accepted Mutation (PAM) and Amino Acid Pair Probabilities

We mentioned that we must choose an appropriate evolutionary model $E((p_{AB})_{AB})$ for the homologous hypothesis, ie we have to find p_{AB} for each pair of amino acids A and B. Since we are using a statistical approach, this has to be estimated from data. If we know that two sequences s and s' are homologous, we could estimate p_{AB} by finding the value of p_{AB} that would maximize

$$P(E((p_{AB})_{AB})|s,s')$$

This can be done by using the maximum likelihood approach (section 2.1.6 pp 52-53) – Review Method
Lagrange Multipliers (Section 2.2) – Review Example and Proof of Method
Appendix (Chapter 3) – Find p_{AB} using the maximum likelihood approach and Lagrange multipliers

We now have $p_{AB} = \frac{n_{AB}(s,s')}{n}$ which is the relative frequency of a pair (A,B) in the alignment of s and s' where $n_{AB}(s,s')$ is the number of times the amino acids A and B are aligned in one column in the alignment of s and s' and n is the length of s and s'.

To find a value for n_{AB}, some homologous sequences are needed. To do this Dayhoff and co-workers used local sequence alignment.

Problem – They used sequence alignment to find a substitution matrix (substitution score matrix) for sequence alignment – which comes first, the chicken or the egg?

Answer – Use only very closely related sequence (sequences differ in at most 15% of the amino acid.

Caveat – The substitution matrix is only valid for closely related protein sequences