Homework assignment 3 for Phys498 Biological Information and Complexity
small submitted by Anoush Aghajani-Talesh on nov. 14, 2001

PSIC: Profile extraction from sequence alignments with position specific counts of independent observations

The PSIC (position specific independent counts) method is a heuristic routine for the extraction of profiles from multiple sequence alignments. It was developed by Sunyaev et al.[1]. In the following I will discuss the theory and the main features of this method. I think it is a good example for what we did in class on sequence alignment methods.
Multiple sequence alignments with up to several hundred protein sequences can be used to identify and to analyse homologies. Applications are for example protein structure prediction or the revealing of evolutionary relationships between species. Similarities between very divergent sequences with remote relationships can usually be better detected by comparing many sequnces at the same time than with a pair-wise alignment.[2]
In an alignment it is often useful to assign different statistical weights to the sequences. Sequences from a database mostly consist of sets of rather similar sequences. Typically one has to compare these with a small number of more divergent sequences. Without sequence weighting techniques the similarities of closely related sequences would be overrepresented and information from remotely related sequences would be surpressed. Within the amino acid sequence of a protein changes in the primary structure appear with different rates depending on the position within the sequence. So called conserved sites change slowly wheras functionally less specific sites might evolve more rapidly. For this reason PSIC does not only use statistical weights that differ from sequence to sequence, but which are also specific with respect to the position within the sequence. This and the comperatively small computational costs are claimed to be the most significant improvement compared to previous profile extracting methods.
As mentioned above PSIC is a tool for the extraction of profiles. Profiles can be regarded as a measure for the probability $p(a,i)$ of finding the amino acid a at the alignment position i. High values of $p(a,i)$ in a certain region of a protein can therefore indicate homologies.
For the purpose of deriving $p(a,i)$ the evolution of a sequence is regarded as a Markov process with position specific transition probabilities. The amino acids are substituting each other with different transition rates. In the limit of infinite evolution time we will observe the frequencies $p(a,i)$. Since in practice sequences are not independent (not sufficent evolutionary time) $p(a,i)$ can not directly be observed and must be estimated. Thus let us consider a given alignment A of N sequences $\{S_1,..,S_N\}$ with L alignment positions i. Wie express $p(a,i)$ in terms of the effective number $n(a,i)_{eff}$ of independent occurences of amino acid type

a at the alignment position i.

$$p(a,i) = \frac{n(a,i)_{eff}}{\sum_b n(b,i)_{eff}} \tag{1}$$

In general $n(a,i)_{eff}$ is not identical with the actually obserbed number $n(a.i)_{obs}$ of counts of a, since sequences are usually not independent. The idea for the estimation of $n(a,i)_{eff}$ from the raw number of occurences $n(a,j)_{obs}$ is to assume that counts from different sequences are less independent the more similar these sequences are. Let A_{aj} be a subset of sequences in which the amino acid a occurs at the alignment position j. We denote by $l(a,j)$ the number of alignment positions within A_{aj} with identical amino acids. Then the probability $P(A_{aj})$ of finding the same amino acid type at any alignment position $i \neq j$ for all sequences in the subset is approximatly given by

$$P(A_{aj}) = \frac{l(a,j)}{m}. \tag{2}$$

Where $m < L$ is the total number of alignment positions in the subset A_{aj} excluding j and positions with gaps.
However thinking of A_{aj} as a set of $n(a,i)_{eff}$ randomly aligned sequences with, according to the default frequencies q_b randomly choosen, aminoacids, we find

$$P(A_{aj}) = \sum_b q_b^{n(a,i)_{eff}}. \tag{3}$$

If we use equation (2) and (3) to identify the frequency of identical positions in a given alignment with the probality of identical alignment positions in random sequences, we obtain

$$\frac{l(a,j)}{m} = \sum_b q_b^{n(a,i)_{eff}} \tag{4}$$

With this equation it is possible to compute $n(a,i)_{eff}$ by dividing the sequences S_k into subsets A_{aj} for which $l(a,j)$ can be determined. Equation (4) can then be solved for $n(a,i)_{eff}$ numerically. A description of the algorithm used by PSIC to create suitable subsets Aaj can by found in [1]. The authors gave no quanitative estimate for the computational costs. However they claimed that they empirically found it to be "negligibly small on standard UNIX workstations".

References:

[1] Sunyaev, R. Shamil et al. , PSIC: profile extraction from sequence alignments with position-specific counts of independent observations, Protein Engeneering vol 12 no.5 pp.387-394, 1999

[2] Molecular Simulations, Inc., manual for the software Homology, Chapter 2: Theory
`http://orson.rz.uni-potsdam.de/orson/software/msi/insight980/homology/980TOC.doc.html`