Statistics for Fragment Comparison - A Biologically Motivated Approach to Sequence Alignment

Sören W. Perrey, Andreas W.M. Dress, Jens Stoye
Research Center for Studies on Structure Formation (RCSF)
University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany

Summary. Methods for aligning protein sequences need *a priori* defined similarity measures on the set of all possible sequences. Usually, these are based on substitution matrices with scores for all possible exchanges of single amino acids. The underlying model of protein evolution is a Marcovian process where substitutions in a polypeptide chain accumulate independently of both, time and position (the i.i.d.-assumption). The resulting "log-odds" matrices are derived from empirical data [2],[13],[1],[6], [5], from analyses of the chemical properties of the side chains [4],[9],[12], from secondary structural propensities of amino acids [8],[10], or from genetic code [3].
We propose still another approach (related to the methods of deriving the matrices from empirical data) based on calculating frequencies of sites (or, more general, pairs of sites at a given distance) where given sets of amino acids can be observed. Using appropriate statistics, we derive scores not only for pairs of amino acids but also for "amino acid profiles". More generally, we derive scores for pairs of pairs of profiles at a given distance which means a relaxation of the i.i.d.-assumption for the evolution at different sites. Consequently, we believe our approach to be particularly useful when searching for significant local similarity indicating conserved fragments.

Approach. Let $M_{N,k} = (\, s_i(\kappa)\,)_{i=1,..,N\,;\,\kappa=1,..,k}$ $(s_i(\kappa) \in \mathcal{A} \cup \{-\};\ \ k, N \in \mathbf{N})$ be a family of N multiply aligned sequences $s_1, .., s_N$ of length k, whose entries come from the union $\mathcal{A} \cup \{-\}$ of some finite alphabet $\mathcal{A}$ (for example, the set of the 20 amino acids occuring in proteins) enlarged artificially by the "gap letter" $\{-\}$. For any such sequence family and any site $\kappa \in \{1, .., k\}$, we consider the set $\mathcal{A}(\kappa) := \mathcal{A} \cap \{\, s_i(\kappa) \mid i = 1, .., N \,\}$ of amino acids occuring at that site.
For every subset $T \subseteq \mathcal{A}$, we calculate the *number of sites* $\kappa \in \{1, .., k\}$ in $M_{N,k}$ which comprise exactly the letters contained in T. We estimate its probability simply by the relative frequency of sites in the multiple alignment with T as "their exact letter set". Because only some part of the "theoretically complete" multiple alignment, which comprises all members of the family, is available, we are even more interested in the probability $\mathcal{P}(S)$ of observing *at least* a subset $S \subseteq T$ at a site. Hence, we consider the relative frequency of *all* sites κ in the multiple alignment with $S \subseteq \mathcal{A}(\kappa)$ - in particular for subsets S consisting of a few elements, only.

To align some new sequences to an aligned set of sequences, the conditional probability $\mathcal{P}(S \mid S_0)$ of observing some set S under the condition that a subset S_0 of S is observed already, is also important. It is estimated by the fraction of the corresponding (relative) frequencies.

We now derive the correlation between two subsets S_1, S_2 by comparing the product of the conditional probabilities $\mathcal{P}(S_1 \mid S_0)$ and $\mathcal{P}(S_2 \mid S_0)$ with the conditional probability $\mathcal{P}(S_1 \cup S_2 \mid S_0)$ of the union of S_1 and S_2 under the condition S_0. They are positively (negatively) correlated relative to S_0, if the relative frequency of sites comprising $S_1 \cup S_2$ is significantly larger (resp. smaller) than the corresponding product.

At present, we have investigated just one databank [11] containing multiple protein sequence alignments (based of their 3 dimensional coordinates) using this approach. We have obtained a substitution matrix which does fit somehow with some “commonly used” matrices as PAM, blossum etc. Because we count sites which comprise a specific subset of amino acids and do not take any account of multiplicities of the amino acids, the diagonal of the substitution matrix is calculated as the relative frequency of occurence of the corresponding amino acid. We have also derived scores for “amino acid profiles”.

In addition, we consider for pairs $\kappa, \kappa + d$ of sites (for small $d \in \mathbf{N}$) the *set of pairs* $\mathcal{A}(\kappa, d) := \mathcal{A} \times \mathcal{A} \cap \{\, (s_i(\kappa), s_i(\kappa + d)) \mid i = 1, .., N \,\}$ of amino acids occuring at site κ and $\kappa + d$.

Analogous to the approach outlined above, one can estimate the conditional probability $\mathcal{P}_d(R \mid R_0)$ of observing at least a subset $R := R^1 \times R^2 \subseteq \mathcal{A} \times \mathcal{A}$ at a pair of sites at distance d under the condition that $R_0 := R_0^1 \times R_0^2$ $(R_0^i \subseteq R^i \; (i = 1, 2))$ has been observed already, by the fraction of the corresponding (relative) frequencies. To obtain correlations, one has to compare the product of the conditional probabilities computed for some subsets R_1, R_2 (relative to the same subset R_0) with the conditional probability associated with their union $R_1 \cup R_2$.

In a first step, we calculated the most basic correlation coefficients relative to values $d = 1, 2, \ldots, 6$ which are of the form

$$\frac{\mathcal{P}_d(\,\{i, j\} \times \{i, h\}\,)}{\mathcal{P}(\,\{i\}\,) \cdot \mathcal{P}(\,\{j, h\}\,)}.$$

Because the investigated databank is quite small, we also have clustered the amino acids with respect to their chemical properties to infer reasonable correlation coefficients. We intend to develop pairwise and multiple alignment procedures based on ((almost) gapless) fragment alignment using the correlation coefficients derived from the statistics outlined above.

References

[1] S.R. Altschul, *J. Mol. Evol.* 219, pp.555-565, 1991

[2] M.O. Dayhoff, R.V. Eck, C.M. Park, in *Atlas of Protein Sequence and Structure* 1967-68, Vol. 3,pp.33-45, Nat. Biomed. Res. Found., Silver Spring, Maryland, 1968

[3] W.M. Fitch, *J. Mol. Evol.* 16, pp.9-16, 1966

[4] R. Grantham, *Science* 185, pp.862-864, 1974

[5] S. Henikoff, J.G. Henikoff, *Proc. Natl. Acad. Sci. USA* , pp.10915-10919, 1989

[6] D.T. Jones, W.R. Taylor, J.M. Thornton, *CABIOS* 8, pp.275-282, 1992

[7] U. Lessel, D. Schomburg, *Prot. .Eng.* 7, no.10, pp. 1175-1187 199

[8] J.M. Levin, B. Robson, J. Garnier, *FEBS Lett.* 205, pp.303-308, 1986

[9] T. Miyata, S.'Miyazawa, T. Yasunaga, *J. Mol. Evol.* 12, pp.219-236, 1979

[10] K. Niefind, D. Schomburg *J. Mol. Biol.* 219, pp.487-491, 1991

[11] S. Pascarella, P. Argos, *Prot. .Eng.* 5, pp.121-137, 1992

[12] J.K.M. Rao, *Int. J. Peptide Protein Res.* 29, pp.276-281, 1987

[13] J.L. Risler, M.O. Delorme, H. Delacroix, A. Henaut, *J. Mol. Biol.* 204, pp.1019-1029, 1988