Improved pairwise alignment of proteins in the Twilight Zone using local structure predictions

Yao-ming Huang and Christopher Bystroff

CENTER FOR BIOINFORMATICS, DEPT OF BIOLOGY, RENSSELAER POLYTECHNIC INSTITUTE, TROY, NEW YORK 12180 USA
{bystrc,huangy2}@rpi.edu

Abstract

Recent advances in the ability to discriminate between homologous and non-homologous proteins in the "Twilight Zone" of sequence similarity, must be accompanied by accurate alignments if they are to be of value to molecular modelers. Pairwise alignments require a measure of evolutionary distance, traditionally modeled using global amino acid substitution matrices. But real differences in the likelihood of substitutions may exist for different structural contexts within proteins, since structure contributes to the selective pressure. HMMSUM (HMMSTR-based SUbstitution Matrices) is a new model for structure-dependent amino acid substitution probabilities consisting of a set of 281 matrices, one for each of the sequence-structure contexts defined in HMMSTR (a Hidden Markov Model for protein STRucture). HMMSUM does not require the structure of the protein to be known, using HMMSTR predictions instead. Alignments using the HMMSUM compare favorably BLOSUM50 alignments when validated against curated remote homolog alignments from BAliBASE.

1. Introduction

Amino acid substitution matrices are evolutionary models that seek to explain the cost of mutating any one of the twenty amino acids to any other, relative to the cost of not mutating. The most widely used substitution matrices, such as PAM [5] and BLOSUM [6], are derived from high confidence multiple sequence alignments. Counting statistics are used to estimate the frequency of each possible mutation, and a ratio of the observed mutation probability to the probability one would expect by chance is calculated. The logarithm of this likelihood ratio is the number we use when scoring pairwise alignments such as those generated by Dynamic Programming algorithms (for example, LALIGN [7]) and by database search algorithms.

2. Methods

2.1. Local structure predictions

The observed amino acid substitution frequencies were summed from a non-redundant training set (PDBselect25) of multiple sequence alignments (MSAs) produced by searching the 'nr' protein database using PSI-BLAST [1], with e-value cutoff 0.001. Sequence weighting [9] was used to correct for unbalanced sampling within an MSA.

HMMSTR [4] assigns structural descriptors to each position in each MSA. HMMSTR takes as input the sequence profile, and optionally the protein structure expressed as backbone angles, and produces as output a set of conditional probabilities $\gamma_{qt}=P(q|t)$, for each sequence position t. Each state q = 1..281 represents a position in one of the I-sites local structure motifs [3].

2.2. Structure-dependent substitution matrices

Substitution and background frequencies were summed in a manner similar to the one described earlier for BLOSUM, except that sequence weights were used instead of binning similar sequences. Also, in our case γ-values were used as positional weights in order to separate the substitution counts into HMMSTR states.

$$F(i,j\,|\,q)=\sum_{\substack{m\in\\ all\ MSAs}}\ \sum_{a\in m}\ \sum_{\substack{b\in m,\\ b<a}}\ \sum_{\substack{t\in\\ a_t=i,\\ b_t=j}} f(a_t,b_t)\gamma_{qt}^{m} \qquad (1)$$

where f depends on the sequence weights. The observed substitution frequencies F are normalized to give probabilities $P(i,j|q)$. The ratio of the *observed* and *expected* substitution frequencies is used as the alignment score.

For the expected frequencies of substitution, we considered two models, called "Dayhoff" (D) and "Lipman" (L) in the spirit of two classic bioinformatics experiments for estimating expectation values for

sequence alignments [5,8]. Our D-model assumes that the expected frequency of substitution is not dependent on the structure. Our L-model uses the structural context to estimate the expected frequency, assigning a different background amino acid frequency distribution to each HMMSTR state.

$$F(i) = \sum_{m \in \text{all MSAs}} \sum_{a \in m} \sum_{t \in a_t = i} f(a_t) \quad \text{(D-model)} \qquad (2)$$

$$F(i \mid q) = \sum_{m \in \text{all MSAs}} \sum_{a \in m} \sum_{t \in a_t = i} f(a_t)\gamma_{qt}^{m} \quad \text{(L-model)} \qquad (3)$$

A structure-dependent, weighted combination of the substitution scores gives the number used (as a log-likelihood ratio, LLR) in the dynamic programming alignment matrix A . For example,

$$P(a_i, b_j \mid Q) = \frac{\sum_{q=1}^{281} P(i, j \mid q)\gamma_{iq}^{a}\gamma_{jq}^{b}}{\sum_{q=1}^{281} \gamma_{iq}^{a}\gamma_{jq}^{b}} \qquad (4)$$

where a_i is the i[th] position of sequence a. To calculate the LLR match score A_{ij} in the alignment matrix, the value in Eq 4 is divided by the expected value (Eqs 2 and 3), and we take the logarithm. Alignments were carried out using Smith-Waterman local Dynamic Programming. Gap penalties and matrix bias were optimized over the sequences being used to validate the method. Therefore we report the optimal alignment accuracy for all methods.

3. Results

To assess the performance of HMMSUM models compared to BLOSUM50, we used a well-documented benchmark database of alignments, BAliBASE [2]. We used 167 MSAs, having a total of 33,977 "True" matches. The alignments have similarity in the "twilight-zone" range (from 7-25% percent identity). All alignments are based on three-dimensional structural superpositions.

Table 1 shows the accuracy and coverage for correctly aligned positions. The statistical significance (P-value) of the differences between methods was evaluated using the Wilcoxon Signed-Rank Test [10].

Better pairwise alignments will lead to better multiple sequence alignments and more sensitive database searches.

Table 1. Comparison of HMMSUM substitution matrices with BLOSUM50.

Matrix	Gap penalty		Correct matches		Accuracy		Coverage (%)
	Opening	Extension	Counts	P value	%	P value	
BLOSUM50	8	2.3	12,211	-	41.8	-	35.9
HMMSUM-M	15	0.9	13,850	<0.001	42.4	0.316	40.8
HMMSUM-L	12	1.2	13,551	<0.001	43.8	0.110	39.9
HMMSUM-D	21	0.5	15,927	<0.001	46.0	<0.001	46.9

4. References

[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **25,** 3389-3402.

[2] Bahr, A., Thompson, J.D., Thierry, J.C., and Poch, O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.,* **29,** 323-326.

[3] Bystroff, C., and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.,* **281,** 565-577.

[4] Bystroff, C., Thorsson, V., and Baker, D. (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.,* **301,** 173-190.

[5] Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In Dayhoff M. (ed), *Atlas of Protein Sequence and structure.* **5 Suppl 3,** National Biomedical Research Foundation, Silver Spring, Maryland, USA, pp. 345-352.

[6] Henikoff, S., and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA,* **89,** 10915-10919.

[7] Huang, X.Q., and Miller, W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.,* **12,** 337-357.

[8] Lipman, D.J., Wilbur, W.J., Smith, T.F., and Waterman, M.S. (1984) On the statistical significance of nucleic acid similarities. *Nucleic Acids Res.,* **12,** 215-226.

[9] Vingron, M., and Argos, P. (1989) A fast and sensitive multiple sequence alignment algorithm. *Comput. Appl. Biosci.,* **5,** 115-121.

[10] Wilcoxon, F. (1945) "Individual Comparisons by Ranking Methods." *Biometrics* **1,** 80-83.