

Protein sequence alignment via anti-translation

Marta Gîrdea, Gregory Kucherov and Laurent Noé

LIFL/CNRS and INRIA-Nord Europe, Parc scientifique de la Haute Borne, 40, Avenue Halley,
59655 Villeneuve d'Ascq Cedex, France

{Marta.Girdea | Gregory.Kucherov | Laurent.Noé}@lifl.fr

Abstract: *We propose a new approach to protein sequence alignment, where the basic idea is to find the best pairwise alignment between two DNA sequences, each of which encodes one of the given proteins. A dynamic programming algorithm was designed for the alignment of directed graphs which combine all these putative DNA sequences. Our goal is to detect protein evolutionary relations that are not captured by other methods.*

Keywords: protein alignment, anti-translation, frameshift, evolution.

1 Motivation

Traditional protein sequence alignment methods, which consist of aligning amino-acid pairs, fail to reveal protein relations when the divergence is caused by frameshift mutations. We aim at inferring such relations by aligning the putative coding DNA, and thus taking into account not only point mutations (substitutions) which affect just one codon, but also frameshift mutations (insertions or deletions of one or more bases) that can alter the reading frame of the ribosomes, affecting all the amino-acids coded after the frameshift. Similar ideas were discussed in [2], where the alignment score of an amino-acid pair is computed from the base pair alignment of the respective codons, and in [3], where several substitution matrices were designed for aligning proteins with a frameshift. We propose a more flexible approach for managing point and frameshift mutations in the putative DNA sequences for an expressive alignment. The goal of this work is to find evolutionary relations between proteins, that are not captured by other methods.

2 The method

Data preparation The basic idea of this approach is to find the best pairwise alignment between two DNA sequences that encode the proteins to analyse. An explicit enumeration and pairwise alignment of all the putative DNA sequences is not an option, since their number increases exponentially with the protein's length. Hence, the protein's "anti-translation" (set of putative DNA sequences) is represented as a directed graph, where a maximal path will retrieve one of these sequences. Basically, these graphs can be seen as generalised sequences, that can have a group of several symbols on each position (which will be further referred to as a "column") and precedence constraints (marked by arcs) between symbols on consecutive positions, as illustrated by the example in Figure 1.

Alignment algorithm We propose a dynamic programming approach, similar to the Smith-Waterman algorithm [5], extended to allow the local alignment of bi-dimensional data structures (anti-translation graphs). Given the input graphs A and B , this algorithm fills a 3-dimensional alignment



Figure 1. Anti-translation graph for sequence MALY; one putative DNA sequence (path) is emphasised.

matrix M , where, for $M[i, j, (\alpha_1, \alpha_2)]$, i and j iterate on the columns of the first and second graph respectively, while (α_1, α_2) enumerates on all possible pairs of symbols (nodes) from $A[i]$, and $B[j]$. The partial scores are computed recursively as shown by equation (1). The conditions $\beta_k \in \text{pred}(\alpha_k)$ ensure that the score for $M[i, j, (\alpha_1, \alpha_2)]$ is computed only from partial scores corresponding to partial alignments of sequences that contain α_1 and α_2 . The third dimension of the matrix being limited by a known, small bound, the time/space complexity remain quadratic in the size of the input.

$$M[i, j, \alpha_1\alpha_2] = \max \begin{cases} 0 \\ M[i-1, j-1, (\beta_1, \beta_2)] + \text{score}(\alpha_1, \alpha_2), & \beta_k \in \text{pred}(\alpha_k), k \in 1, 2 \\ M[i-1, j, (\alpha_1, \beta_2)] + \text{gap-penalty}, & \beta_2 \in \text{pred}(\alpha_2) \\ M[i, j-1, (\beta_1, \alpha_2)] + \text{gap-penalty}, & \beta_1 \in \text{pred}(\alpha_1) \end{cases} \quad (1)$$

The scoring system was designed to discourage the choice of very infrequent codons, which could lead to an apparently good alignment, but unlikely to correspond to the true DNA for the proteins of interest. The knowledge about codon usage appears in the form of a weight (probability, denoted w) computed from a codon usage table and attached to each node. These weights are used for correcting (lowering) the score of two nucleotides $s(a, b)$ as follows: $\text{score}(a, b) = s(a, b) + \log w_a w_b$.

Traditional linear and affine gaps are not expressive enough for this context. Instead, two particular kinds of gaps are considered: *frameshifts* – gaps of size 1 or 2, with high penalty, whose number in a local alignment can be limited, and *codon skips* – gaps of size 3 which correspond to the insertion or deletion of one codon (or amino-acid, if we refer to the protein sequence).

3 First experimental results

Experiments on sequences from *E. coli* and related species from the Pfam [1] database brought out several significant¹ alignments between proteins considered unrelated. Further investigations revealed that consensus sequences for the families of such proteins tend to align better than any pairs of proteins from those families, which suggests the possibility of using this approach to find ancestral sequences.

References

- [1] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, and E.L.L. Sonnhammer, The Pfam Protein Families Database, in *Nucleic Acids Research*, Oxford Univ Press, pp. 276-280, 2002.
- [2] J. Leluk, A non-statistical approach to protein mutational variability, in *BioSystems*, volume 56, number 2-3, Elsevier, pp. 83-93, 2000.
- [3] M. Pellegrini and T.O. Yeates, Searching for Frameshift Evolutionary Relationships Between Protein Sequence Families, in *PROTEINS: Structure, Function, and Genetics*, vol 37, pp.278-283, 1999.
- [4] R. Olsen, R. Bundschuh, and T. Hwa, Rapid assessment of extremal statistics for gapped local alignment, in *Proc. of the 7th Int. Conf. on Intelligent Sys. for Mol. Biol.*, AAAI Press, pp. 211-222, 1999.
- [5] T.F. Smith and M.S. Waterman, Identification of Common Molecular Subsequences, in *J. Mol.Biol.*, volume 147, number 1, Elsevier, pp. 195-197, 1981.

¹ The score significance was estimated according to the Gumbel distribution, where the parameters λ and K were computed with the method described in [4].