44. Contextual Multiple Sequence Alignment (Context helps in aligning orphan genes)

Rafał Otto,[1] **Anna Gambin**[2]

Keywords: context dependency, multiple alignment, orphan genes, amino acids substitution tables, BAliBASE

1 Introduction

The multiple alignment of biological sequences has become an essential tool in molecular biology. It is used to find conserved regions and motifs in protein families, to detect the homology between new sequences and groups of sequences having already known function and in a preliminary phase of protein structure prediction. Multiple alignment is also extensively used in molecular evolutionary analysis.

The various genome projects have provided the biologist with a great number of new protein sequences, and the rate of appearance of these data is steadily increasing. The development of accurate and reliable multiple alignment program which is capable to handling many (often very divergent) sequences simultaneously is still of major importance.

The complexity of the problem do not allow to find the exact solution in the reasonable computational time [1]. Traditionally, the most popular heuristic approach has been the progressive alignment method.

We propose to explore new model for sequence alignment, in which the score for the substitution depends on its neighborhood in the sequence, too. Such a *contextual alignment model* has been proposed recently in [2] for the pairwise alignment problem. To use the notion of context while considering multiple alignment we have decided to relax slightly this model. However, we still need the family of contextual amino acid substitution matrices. The novel approach to construct such a tables have been developed. We present preliminary experimental results that illustrate the advantage of using contextual approach in progressive alignment algorithm. It turned to be particularly useful in aligning the family of sequences containing several *orphans* (i.e. distantly related sequences, sometimes sharing the common fold).

It should be clear that the existence of orphan genes is unavoidable. Despite of the accumulation of genetic information, newly sequenced genomes continue to reveal a high proportion (even to 50%) of uncharacterized genes. Among them there is a significant number of strictly orphan genes without any resemblance to previously determined protein sequences. Moreover, most genes found in databases have only be predicted by computer methods and have never been experimentally validated. Hence, for the alignment method it is important to tolerate orphans (some existing programs exclude the divergent orphans as unrelated or unalignable sequences) and to keep the stability of the family alignment when orphans are introduced into the sequence set.

[1] Warsaw School of Economics, ul. Rakowiecka 24, 02-554 Warsaw, Poland.
E-mail: rafal@rafalotto.com
[2] Institute of Informatics, Warsaw University, ul. Banacha 2, 02 097, Warsaw, Poland.
E-mail: aniag@mimuw.edu.pl

2 Results

Contextual alignment model considered in [2] cannot be directly applied for the problem of multiple alignment. In this model the score of an alignment depends on the order of operations (substitutions and indels) performed, as a substitution at one position can change the context for neighboring sites. The optimal alignment for the pair of sequences was defined as the alignment having the maximal score, when we maximize over all possible chronologies of evolutionary changes.

To deal with several sequences simultaneously and to keep the context-dependency, we propose relaxed contextual model. In this model we penalize substitution also considering two surrounding letters but we do not care about the relative order of operations. We consider all possible contexts (maximum 9 if the substitution is surrounded by two indels) for the substitution and take an average of the contextual scores. Notice, that standard non-contextual e.g. Blosum matrix entry can be viewed as an average over all 400 possible pairs of context.

Our contextual multiple alignment algorithm can be viewed as a contextual extension of popular ClustalW algorithm [3], which belongs to the family of progressive alignment algorithms.

The BAliBASE (Benchmark Alignments dataBASE) is a database of manually-refined multiple sequence alignments. It is specifically designed for the evaluation and comparison of multiple sequence alignment programs.

In our experiment the multiple alignments are calculated for all reference sets in two settings: contextual and non-contextual. Then, the results obtained were compared with the reference alignments from the database. For this comparison the measure *sum-of-pairs score* from [4] has been used. It is the frequency of properly aligned pairs of residues w.r.t. the reference alignment. In the Table below we present several families, for which the contextual approach yields much better results.

protein family	# of seq.	context	non-context	% of improvement
1ycc: cytochrome e	4	0,765	0,665	15,04
2trx: thioredoxin	4	0,671	0,468	43,38
1aboA: sh3	15	0,683	0,580	17,76
1uky: uridyl kin	24	0,541	0,464	20,91
sh3-2-ref6: sh3	6	0,553	0,454	21,81
sh3-3-ref6: sh3	5	0,430	0,214	100,93
AVG		**0,606**	**0,474**	**29,11**

Table 1: The influence of context-dependency to the quality of the alignment.

References

[1] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.

[2] Gambin, A., Lasota, S., Szklarczyk, R., Tiuryn, J. and Tyszkiewicz, J. Contextual Alignment of Biological Sequences, In: *Proc. of European Conference on Computational Biology ECCB, Saarbrücken, Germany* Bioinformatics, vol. **18 (2)**, 2002, pp. 116-127.

[3] Thompson, J.D., Higgins, D.G. and Gibson, T.J. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, vol. **22**, 1994, pp. 4673-4680.

[4] Thompson, J.D., Plewniak, F. and Poch, O. A comprehensive comparison of multiple sequence alignment programs, *Nucleic Acids Research*,1999, vol. **27**, pp. 2682-2690.