MSARC: Multiple Sequence Alignment by Residue Clustering

Michał Modzelewski and Norbert Dojer

Insitute of Informatics, University of Warsaw, Poland
dojer@mimuw.edu.pl

Abstract. Progressive methods offer efficient and reasonably good solutions to the multiple sequence alignment problem. However, resulting alignments are biased by guide-trees, especially for relatively distant sequences.

We propose MSARC, a new graph-clustering based algorithm that aligns sequence sets without guide-trees. Experiments on the BAliBASE dataset show that MSARC achieves alignment quality similar to best progressive methods and substantially higher than the quality of other non-progressive algorithms. Furthermore, MSARC outperforms all other methods on sequence sets whose evolutionary distances are hardly representable by a phylogenetic tree. These datasets are most exposed to the guide-tree bias of alignments.

MSARC is available at http://bioputer.mimuw.edu.pl/msarc

Keywords: multiple sequence alignment, stochastic alignment, graph partitioning.

1 Introduction

Determining the alignment of a group of biological sequences is among the most common problems in computational biology. The dynamic programming method of pairwise sequence alignment can be readily extended to multiple sequences but requires the computation of an n-dimensional matrix to align n sequences. Consequently, this method has an exponential time and space complexity.

Progressive alignment [21] offers a substantial complexity reduction at the cost of possible loss of the optimal solution. Within this approach, subset alignments are sequentially pairwise aligned to build the final multiple alignment. The order of pairwise alignments is determined by a guide-tree representing the phylogenetic relationships between sequences.

There are two drawbacks of the progressive alignment approach. First, the accuracy of the guide-tree affects the quality of the final alignment. This problem is particularly important in the field of phylogeny reconstruction, because multiple alignment acts as a preprocessing step in most prominent methods of inferring a phylogenetic tree of sequences. It has been shown that, within this approach, the inferred phylogeny is biased towards the initial guide-tree [23,11].

A. Darling and J. Stoye (Eds.): WABI 2013, LNBI 8126, pp. 259–272, 2013.

Second, only sequences belonging to currently aligned subsets contribute to their pairwise alignment. Even if a guide-tree reflects correct phylogenetic relationships, these alignments may be inconsistent with remaining sequences and the inconsistencies are propagated to further steps. To address this problem, in recent programs [15,2,8,1,17] progressive alignment is usually preceded by *consistency transformation* (incorporating information from all pairwise alignments into the objective function) and/or followed by *iterative refinement* of the multiple alignment of all sequences.

In the present paper we propose MSARC, a new multiple sequence alignment algorithm that avoids guide-trees altogether. MSARC constructs a graph with all residues from all sequences as nodes and edges weighted with alignment affinities of its adjacent nodes. Columns of best multiple alignments tend to form clusters in this graph, so in the next step residues are clustered (see Figure 1a). Finally, MSARC refines the multiple alignment corresponding to the clustering.

Experiments on the BAliBASE dataset [22] show that our approach is competitive with the best progressive methods and significantly outperforms current non-progressive algorithms [20,19]. Moreover, MSARC is the best aligner for sequence sets with very low levels of conservation. This feature makes MSARC a promising preprocessing tool for phylogeny reconstruction pipelines.

2 Methods

MSARC aligns sequence sets in several steps. In a preprocessing step, following Probalign [17], *stochastic alignments* are calculated for all pairs of sequences and consistency transformation is applied to resulting posterior probabilities of residue correspondences. Transformed probabilities, called residue alignment affinities, represent weights of an *alignment graph*[1]. MSARC clusters this graph with a top-down hierarchical method (Figure 1c). Division steps are based on the Fiduccia-Mattheyses graph partitioning algorithm [3], adapted to satisfy constraints imposed by the sequence order of residues. Finally, multiple alignment corresponding to resulting clustering is refined with the iterative improvement strategy proposed in Probcons [1], adapted to remove clustering artefacts.

2.1 Pairwise Stochastic Alignment

The concept of stochastic (or probability) alignment was proposed in [13]. Given a pair of sequences, this framework defines statistical weights of their possible alignments. Based on these weights, for each pair of residues from both sequences, the posterior probability of being aligned may be computed. A consensus of highly weighted suboptimal alignments was shown to contain pairs with significant probabilities that agree with structural alignments despite the optimal alignment deviating significantly. Mückstein et al. [14] suggest the use

[1] Our notion of alignment graph slightly differs from the one of Kececioglu [9]: removing edges between clusters transforms the former into the latter.