

# Divide-and-Conquer Multiple Alignment with Segment-Based Constraints

Michael Sammeth<sup>1</sup>, Burkhard Morgenstern<sup>2</sup> and Jens Stoye<sup>1</sup>

<sup>1</sup>Bielefeld University, Department of Genome Informatics, Faculty of Technology, P.O. 10 01 31, 33594 Bielefeld, Germany and <sup>2</sup>University of Göttingen, Department of Bioinformatics, Institute of Microbiology and Genetics, Goldschmidtstr. 1, 37077 Göttingen, Germany

## ABSTRACT

A large number of methods for multiple sequence alignment are currently available. Recent benchmarking tests demonstrated that strengths and drawbacks of these methods differ substantially. *Global* strategies can be outperformed by approaches based on *local* similarities and vice versa, depending on the characteristics of the input sequences. In recent years, *mixed* approaches that include both global and local features have shown promising results. Herein, we introduce a new algorithm for multiple sequence alignment that integrates the global *divide-and-conquer* approach with the local *segment-based* approach, thereby combining the strengths of those two strategies.

**Contact:** [micha@sammeth.net](mailto:micha@sammeth.net)

## INTRODUCTION

Automatic generation of multiple alignments is a central task of computational biology. Although diverse methods are now available, no final solution applicable in all possible alignment situations has been found (Notredame, 2002). Traditionally, there exist two opposed strategies of alignment construction, one creating *global* alignments and the other one detecting *local* similarities among the input sequences.

For global alignment, *simultaneous* approaches create alignments by synchronising the information of all input sequences in a  $k$ -dimensional lattice. Although highly elaborated algorithms have been developed to narrow regions of interest within this lattice (Gupta et al., 1995; Tönges et al., 1996), these approaches are computationally expensive so that their application is limited. For this reason, alternative approaches have been developed where the multiple alignment problem is reduced to a series of pairwise *profile* alignments (Feng and Doolittle, 1987; Higgins and Sharp, 1988; Taylor, 1988); the most popular of these *progressive* methods is CLUSTAL W (Thompson et al., 1994). However, a serious drawback of this technique is that the resulting multiple alignments crucially depend on the *order* in which the profile alignments are

carried out.

To cope with more locally related sequence sets, a number of alternative approaches have been proposed that focus on locally related segments of the sequences (Depiereux et al., 1997; Morgenstern et al., 1996; Schuler et al., 1991; Vingron and Argos, 1991). These approaches are superior to more traditional strategies in situations where large gaps need to be inserted into the alignment and for data sets that are evolutionarily distantly related. However, they may be outperformed by global methods where sequence sets are related over their entire length (Lassmann and Sonnhammer, 2002; Thompson et al., 1999b).

Obviously, it is highly desirable to have alignment algorithms performing well on both, globally and locally related sequences. Notredame *et al.* proposed an approach where both, local and global alignment information, is pairwise preprocessed and extended to the multiple context in a heuristic solution of the maximum weight trace problem (Kececioğlu, 1993). Biasing those preprocessed similarities improved the results of standard progressive alignment, and the resulting procedure has been implemented in the program T-COFFEE (Notredame et al., 2000). Moreover, Myers *et al.* developed an algorithm for progressive multiple alignment with *constraints* (Myers et al., 1996). Herein, we introduce an algorithm that performs simultaneous multiple alignment under constraints given by pre-calculated local sequence similarities. In our implementation, we combine the global divide-and-conquer algorithm DCA (Stoye, 1998) with the local segment-based program DIALIGN (Morgenstern, 1999). We evaluate this mixed method and compare its results to both of the native protocols and to other successful alignment methods (i.e., T-COFFEE and CLUSTAL W).

## TECHNICAL BACKGROUND

A *global alignment* of a family of  $k$  sequences  $S = (s_1, s_2, \dots, s_k)$  over a finite alphabet  $\Sigma$  can be defined as a  $k \times m$  matrix  $A$  with entries in an extended alphabet  $\Sigma^* = \Sigma \cup \{-\}$ , such that ignoring the blank characters,