
QAlign: quality-based multiple alignments with dynamic phylogenetic analysis

M. Sammeth^{1, 2, 3, *, †}, J. Rothgänger⁴, W. Esser¹, J. Albert¹,
J. Stoye³ and D. Harmsen⁵

¹Department of Computer Science II, University of Würzburg, 97074 Würzburg, Germany, ²International Graduate School of Bioinformatics, Bielefeld University, 33594 Bielefeld, Germany, ³Genome Informatics, Department of Technology, Bielefeld University, 33594 Bielefeld, Germany, ⁴RIDOM bioinformatics, 97082 Würzburg, Germany and ⁵Institute for Hygiene, University of Münster, 48149 Münster, Germany

Received on December 11, 2002; revised on February 12, 2003; accepted on March 4, 2003

ABSTRACT

Summary: Integrating different alignment strategies, a layout editor and tools deriving phylogenetic trees in a 'multiple alignment environment' helps to investigate and enhance results of multiple sequence alignment by hand. QAlign combines algorithms for fast progressive and accurate simultaneous multiple alignment with a versatile editor and a dynamic phylogenetic analysis in a convenient graphical user interface.

Availability: QAlign is freely available over the internet at <http://www.ridom.de/qalign/>. The platform-independent JAVA technology used provides distributions for various operating systems and hardware architectures.

Contact: qalign@ridom.de

INTRODUCTION

The correct alignment of multiple DNA and protein sequences is a fundamental problem in computational biology. Results produced by the commonly used progressive multiple alignment methods can be obtained rapidly but they are highly dependent on the degree of similarity of the input. Simultaneous alignment algorithms synchronize the information in all sequences to construct the multiple alignment and are therefore more sensitive. However, even these optimal alignment layouts may need some manual editing. Furthermore, downstream analyses (e.g. methods to derive phylogenetic trees) are linked dynamically to the multiple alignment. Thus, a stronger interaction between the creation of the alignment and the phylogenetic analysis enables evolutionary trees of high quality to be found.

IMPLEMENTATION

Due to its modular and layered structure, our program QAlign may easily be extended to support additional algorithms for both, multiple alignment and phylogenetic reconstruction. Herein we outline the features included in the current version.

Multiple alignment algorithms

QAlign is a new graphical environment integrating multiple features in the construction of the best multiple alignment for a specific set of sequences (FASTA and MSF sequence format supported). The algorithm monitor controls the construction of the multiple alignment where a fast progressive or a more accurate simultaneous approach may be chosen to align the sequences or parts of them (see Fig. 1 right). The heuristics used in the progressive approach (QAlign uses the variant of the MSA protocol, Gupta *et al.*, 1995) of global multiple sequence alignment allow the alignment of even very large data sets. However, the drawback is that the resulting alignment is a fast approximation of the solution (McClure *et al.*, 1994; Hickson *et al.*, 2000).

In addition, QAlign contains an efficient and stable re-implementation of the NCBI's MSA (multiple sequence alignment) program (Gupta *et al.*, 1995). This is based on the simultaneous alignment strategy, an exact algorithm capable of finding the optimal mathematical solution. In addition to the optimizations used in MSA, the divide-and-conquer algorithm DCA (Tönges *et al.*, 1996; Stoye *et al.*, 1997) was used to achieve the simultaneous alignment of larger data sets. The desired quality-time tradeoff ratio for simultaneous alignment construction can be balanced by a slider. Both the progressive and the simultaneous alignment strategy may be used in a complementary manner on the same alignment layout.

*To whom correspondence should be addressed.

† Present address: Genome Informatics, Department of Technology, Bielefeld University, 33594 Bielefeld, Germany.

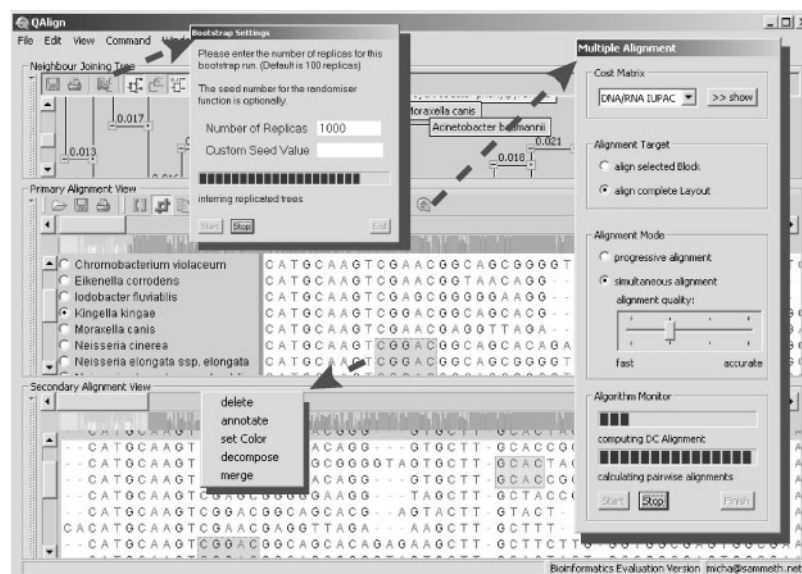


Fig. 1. The graphical user interface of QAlign: the neighbour joining tree is updated dynamically (top) different algorithms and their parameters are accessible by the algorithm monitor (centre-right) and context menus support the editing functions for each block (bottom).

Alignment editor features

After aligning the sequences, the graphical editor of QAlign provides features to analyse the result and modify the multiple alignment layout (as in Fig. 1, bottom). Gaps may be inserted or deleted and marked blocks may be moved within the alignment providing that the aligned sequences have the same length. An immediately updated consensus sequence with coloured bars shows the matching ratio of each column. These bars represent the conservation of different clusters across the alignment layout. They are also displayed as a bird's eye view underneath a scrollbar thus allowing easy navigation to areas of low similarity. A secondary view is provided which may be used to extend the editor capabilities on one alignment or to compare two different multiple alignment layouts.

Dynamic phylogenetic analysis

A dynamic phylogenetic tree view makes visible the consequences of a change in the alignment with regard to the phylogenetic relationship (see Fig. 1, top) where branch lengths may change and nodes may swap according to the neighbour-joining method (Saitou and Nei, 1987). The tree may also be bootstrapped at any time to reveal its current stability. Thus, a phylogenetic reflection of the dynamics of the multiple alignment layout is obtained.

A variety of visual rearrangements is provided for the tree (e.g. subtrees may be collapsed or rearranged). Finally, the phylogenetic tree may be exported, either to a vectorial data format for drawing tools (SVG) or to a common format used by tree plotters (Newick).

CONCLUSION

QAlign provides a practical solution for the creation of refined multiple alignments: layouts produced by various algorithms may be used as a starting-point for changes done by hand, while the phylogenetic consequences are visualised on the fly. Furthermore, the comparison of multiple alignments is made easier because of the two alignment views integrated in the user interface of QAlign.

REFERENCES

- Gupta,S.K., Kececioğlu,J.D. and Schäffer,A.A. (1995) Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *J. Comp. Biol.*, **2**, 459–472.
- Hickson,R.E., Simon,C. and Perrey,S.W. (2000) The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. *Mol. Biol. Evol.*, **17**, 530–539.
- McClure,M.A., Vasi,T.K. and Fitch,W.M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.*, **11**, 571–592.
- Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
- Stoye,J., Moulton,V. and Dress,A.W. (1997) DCA: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. *Comput. Appl. Biosci.*, **13**, 625–626.
- Tönges,U., Perrey,S.W., Stoye,J. and Dress,A.W.M. (1996) A general method for fast multiple sequence alignment. *Gene*, **172**, GC33–GC41.