MOSAL: software tools for multiobjective sequence alignment

CISUC Technical Report TR2013/XX

Luís Paquete[1], Pedro Matias[1], Maryam Abbasi[1], Miguel Pinheiro[2]

[1]CISUC, Department of Informatics Engineering,
University of Coimbra, Portugal.
paquete/maryam@dei.uc.pt,
pamatias@student.dei.uc.pt
[2]School of Medicine, University of St. Andrews, UK.
mmp2@st-andrews.ac.uk

Abstract

Multiobjective sequence alignment brings the advantage of providing a set of alignments that represent the trade-off between performing insertion/deletions and matching symbols from both sequences. Each of these alignments provide a potential explanation of the relationship between the sequences. We introduce MOSAL, a software tool that provides an open-source implementation and an on-line application for multiobjective pairwise sequence alignment.

1 Background

Sequence alignment is in the core of many bioinformatics applications. It aims to identify regions of similarity in sequences of biological data, such as nucleotide and amino acid residues. The procedure consists of inserting gaps between the residues so that similar symbols from several sequences become aligned. For two sequences, dynamic programming algorithms can compute the optimal alignment in an efficient manner [8]. However, for very large DNA or protein databases, heuristic approaches like FASTA and BLAST have been used [2, 7]. See [5] for an extensive review from a computational point of view.

Any of these approaches rely on the a priori definition of coefficients that are assigned to the components of the score function. These weights

are usually defined by default in most of the software packages for sequence alignment and are usually not modified by the practitioner. However, there is a considerable disagreement about how to weight each coefficient. A small change in the weights can lead to a completely different alignment.

One way of overcoming the problem of setting weights is to consider a multiobjective formulation, where the practitioner is provided a set of optimal alignments representing the trade-off between components of the score function, for instance, substitution score given by a substitution matrix and the number of gaps; in this case, an alignment is optimal if there is no other alignment with better substitution score value and lesser number of gaps. Usually, there is not only one optimal alignment but several for which this notion of optimality holds; such set of all optimal alignments is called *the Pareto optimal alignment set*.

Under a multiobjective formulation, no weights are needed to be set up. Moreover, according to a classical result in the multiobjective optimization field [4], this optimal set contains not only all of the optima of a weighted sum formulation, but also many other alignments that are not possible to find at all by the weighted sum approach. Each of these alignments can be seen as a potential explanation of the relationship between the sequences and may be of interest for the practitioner for a more in-depth analysis. In fact, several other problems in bioinformatics have been already reformulated from a multiobjective point of view [6].

A multiobjective approach to pairwise sequence alignment has been explored by several researchers, both from a problem formulation and algorithmic point of view [1, 3, 9–11]. Recently, it has been applied to the construction of phylogenetic trees, which has shown to provide complementary information to that obtained by common methods [1].

2 Implementations

MOSAL is a software tool that results from the problem formulation given in [1] with the aim of providing an open-source implementation and an on-line application where this implementation can be tested. The web-server is available at `http://mosal.dei.uc.pt` and physically located at the Department of Informatics Engineering, University of Coimbra, and is one of the outcomes of a national funded research project on multiobjective sequence alignment.

Table 1: Command line options

option	explanation
`F1`	path to the 1st sequence file (FASTA)
`F2`	path to the 2nd sequence file (FASTA)
`i|g`	use indels or gaps
`dp | -dpp -b=N`	do not use or use pruning technique. If yes, specify the size of lower bound (`N`)
`-ss=F`	use substitution score instead of matches (`F` is the path to the matrix file)
`--no-traceback`	output only the scores without the alignments

2.1 Code

The code is written in `C` and provided under a GNU General Public License. A makefile is available for compilation under GNU/Linux. The implementation can be setup for several multiobjective score functions as described in [1]: maximization of the number of matches or substitution score and minimization of gaps or indels.

Speed-up techniques described in [1] are also implemented and can be parameterized, in particular, the maximum size of the lower bound set for the pruning technique. This parameter should be defined with some care; if too small, the pruning has a reduced effect, and if too large, a excessive number of comparisons may reduce the advantage of pruning in terms of CPU-time. For most of the benchmarks tested, a value of 10 seems to be the most appropriate [1].

The command line options available are described in Table 1. The implementation outputs the Pareto optimal set of alignments and the corresponding score function values by default.

2.2 On-line application

The web-server provides also an on-line application, written in PHP, that is available for sequences up to 2000 symbols. Four steps are needed to produce the set of Pareto optimal alignments:

Step 1: Insertion of each sequence in FASTA format in a text box. The user can choose either Protein or DNA type of sequence in a switch button.

Step 2: Choice of the score function with switch buttons. The user can choose either matches or substitution score for the first score function component and either indels or gaps for the second score function component. If substitution score is chosen, the user can choose a substitution score matrix (PAM 100, 250 and BLOSUM 62, 75, 80, 85 if Protein option is chosen in the previous step) or can even provide one in a predefined text format.

Step 3: Choice of the sequence alignment options: with or without the alignments and with or without pruning technique. If pruning is chosen, the number of bounds must be provided (10 is given by default). The option without alignment provides only the score function values of the alignments.

Step 4: Submit to the server, with the option of sending an e-mail to the user with the output files.

Once the Pareto optimal alignment set is computed, the score function values are shown in an iterative plot; the user can zoom and choose a given point to see the corresponding alignment, see Figure 1. No information about the submissions is stored in the web-server. During the benchmark testing, the application was able to retrieve the output in less than 10 seconds for the largest sizes.

A visualization tool in the on-line application allows to visualize all the alignments and the corresponding score function values produced by the implementation or by the on-line application.

3 Conclusions

MOSAL provides a set of tools for the practitioner to perform a more in-depth analysis on the relation between a pair of biological sequences. The multiobjective formulation that is explored by the framework provides further insight into the confidence of the alignments obtained by common methods; for instance, a large number of optimal scores suggests that a single alignment may be insufficient to understand the relation between the sequences and that further investigation is required. Moreover, the output can be used to construct phylogenetic trees as suggested in [1].

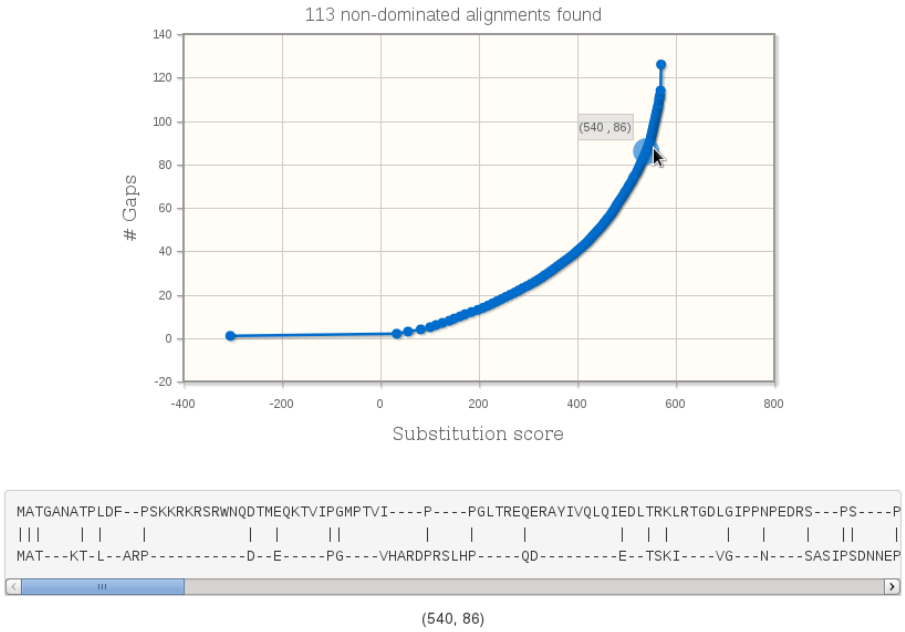

Figure 1: The figure illustrates the output of the on-line application for two sample sequences. The plot in the top shows the score function values as points for substitution score (bottom axis) and number of gaps (left axis). The alignment for a given score function value is given below.

Acknowledgements

This work was support by the Fundação para a Ciência e Tecnologia, project MOSAL - Multiobjective sequence alignment (PTDC/EIA-CCO/098674/2008) and by FEDER, Programa Operacional Factores de Competitividade do QREN, ref. COMPETE: FCOMP-01-0124-FEDER-010024.

References

[1] M. Abbasi, L. Paquete, A. Liefooghe, M. Pinheiro, and P. Matias. Improvements on bicriteria pairwise sequence alignment: algorithms and applications. *Bioinformatics*, 29(8):996–1003, 2013.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[3] K.W. DeRonne and G. Karypis. Pareto optimal pairwise sequence alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(2):481–493, 2013.

[4] M. Ehrgott. *Multicriteria optimization.* Springer, Berlin, 2005.

[5] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences.* Computer Science and Computational Biology. Cambridge University Press, New York, 1997.

[6] Julia Handl, Douglas B. Kell, and Joshua D. Knowles. Multiobjective optimization in bioinformatics and computational biology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(2):279–292, 2007.

[7] D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

[8] Saul Ben Needleman and Christian Dennis Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[9] M.A. Roytberg, M.N. Semionenkov, and O.I. Tabolina. Pareto-optimal alignment of biological sequences. *Biophysics*, 44(4):565–577, 1999.

[10] T. Schnattinger, U. Schöning, and H. Kestler. Structural RNA alignment by multi-objective optimization. *Bioinformatics*, 29(13):1607–1613, 2013.

[11] Akito Taneda. Multi-objective pairwise RNA sequence alignment. *Bioinformatics*, 26(19):2383–2390, 2010.