BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 23 2013, pages 3102–3104
doi:10.1093/bioinformatics/btt536

*Sequence analysis*

Advance Access publication September 16, 2013

# RNA-Pareto: interactive analysis of Pareto-optimal RNA sequence-structure alignments

Thomas Schnattinger[1,2], Uwe Schöning[1], Anita Marchfelder[3] and Hans A. Kestler[2,*]

[1]Institute of Theoretical Computer Science, [2]Medical Systems Biology and [3]Biology II, Ulm University, D-89069 Ulm, Germany

Associate Editor: Ivo Hofacker

**ABSTRACT**

**Summary:** Incorporating secondary structure information into the alignment process improves the quality of RNA sequence alignments. Instead of using fixed weighting parameters, sequence and structure components can be treated as different objectives and optimized simultaneously. The result is not a single, but a Pareto-set of equally optimal solutions, which all represent different possible weighting parameters. We now provide the interactive graphical software tool RNA-Pareto, which allows a direct inspection of all feasible results to the pairwise RNA sequence-structure alignment problem and greatly facilitates the exploration of the optimal solution set.

**Availability and implementation:** The software is written in Java 6 (graphical user interface) and C++ (dynamic programming algorithms). The source code and binaries for Linux, Windows and Mac OS are freely available at http://sysbio.uni-ulm.de and are licensed under the GNU GPLv3.

**Contact:** hans.kestler@uni-ulm.de

Received on July 8, 2013; revised on August 14, 2013; accepted on September 9, 2013

## 1 INTRODUCTION

Noncoding RNAs play important roles in translation or gene regulation (Latchman, 2005). Many families of noncoding RNA show little sequence similarity, but a conserved secondary structure. Furthermore, the function of RNA molecules, besides their sequence, is mainly defined by their secondary structure (cf. Zuker and Sankoff, 1984). If the sequence identity is too low, standard sequence alignment tools fail to produce reliable alignments (Gardner *et al.*, 2005). Consequently, both sequence and secondary structure need to be taken into account.

Sankoff's algorithm (Sankoff, 1985) solves this by a dynamic programming algorithm, which does not only compute a sequence alignment, but solves the consensus folding problem simultaneously. The result is a sequence-structure alignment that is optimal with respect to some fixed objective function, which is the weighted sum of a sequence and a structure component. This algorithm has triggered the development of many tools, which try to improve the original performance by restricting the solution space or by introducing heuristics (Havgaard *et al.*, 2007; Mathews, 2005; Will *et al.*, 2007). The fixed weighting between sequence and structure objectives, which has to be estimated or optimized in advance, is one of the limitations of all of these approaches.

This problem is not new, and approaches using Pareto-optimality are also found in other domains such as economics and classification (Ehrgott, 2005; Müssel *et al.*, 2012). In sequence alignment, e.g. Roytberg *et al.* (1999) constructs a set of Pareto-optimal solutions by using the number of gaps and (mis-)matches as separate objectives. Taneda (2010, 2011) describes an evolutionary algorithm and accompanying web-tool for pairwise RNA sequence alignment and uses a structure score derived from the alignment as a second objective for the approximation of a predefined number of Pareto-optimal solutions.

Here, we now calculate an exact set of Pareto-optimal sequence-structure alignments using distinct objectives for sequence and structure (Schnattinger *et al.*, 2012, 2013). Based on this method, we present an interactive tool, which allows the user to investigate and explore these Pareto-optimal solutions.

## 2 DETAILS AND IMPLEMENTATION

### 2.1 Multi-objective dynamic programming

An alignment is scored by two independent objective functions. The first one (1) is the score for the sequence alignment $R$. The second objective (2) function scores the alignment with respect to its consensus secondary structure $S$:

$$f_{seq}(R,S) = \gamma(R) + \sum_{\substack{(i,k)\in R \\ unpaired}} \sigma(A_i, B_k) \quad (1)$$

$$f_{str}(R,S) = \sum_{(ij;kl)\in S} \Psi^A_{ij} + \Psi^B_{kl} \quad (2)$$

$f_{seq}$ sums up the gap penalties of the sequence alignment ($\gamma$), where $\gamma$ can be a linear or an affine gap cost function, and the sequence matches of unpaired columns ($\sigma$). $f_{str}$ is the sum of all log transformed base pair probabilities ($\Psi$) of the constructed consensus structure. For more details, see Schnattinger *et al.* (2013).

In general, there is no single optimal solution that maximizes both objectives. Therefore, we extended the concept of optimality to vector valued solutions. A scoring vector $a = (a_{seq}, a_{str})$ dominates $b = (b_{seq}, b_{str})$ if either $a_{seq} \geq b_{seq}$ and $a_{str} > b_{str}$ or $a_{seq} > b_{seq}$ and $a_{str} \geq b_{str}$. An alignment is Pareto-optimal if its scoring vector is not dominated by the scoring vector of any other valid alignment (cf. Ehrgott, 2005). The computation of the Pareto-optimal alignments is done by a multi-objective dynamic programming algorithm, which is based on two

*To whom correspondence should be addressed.

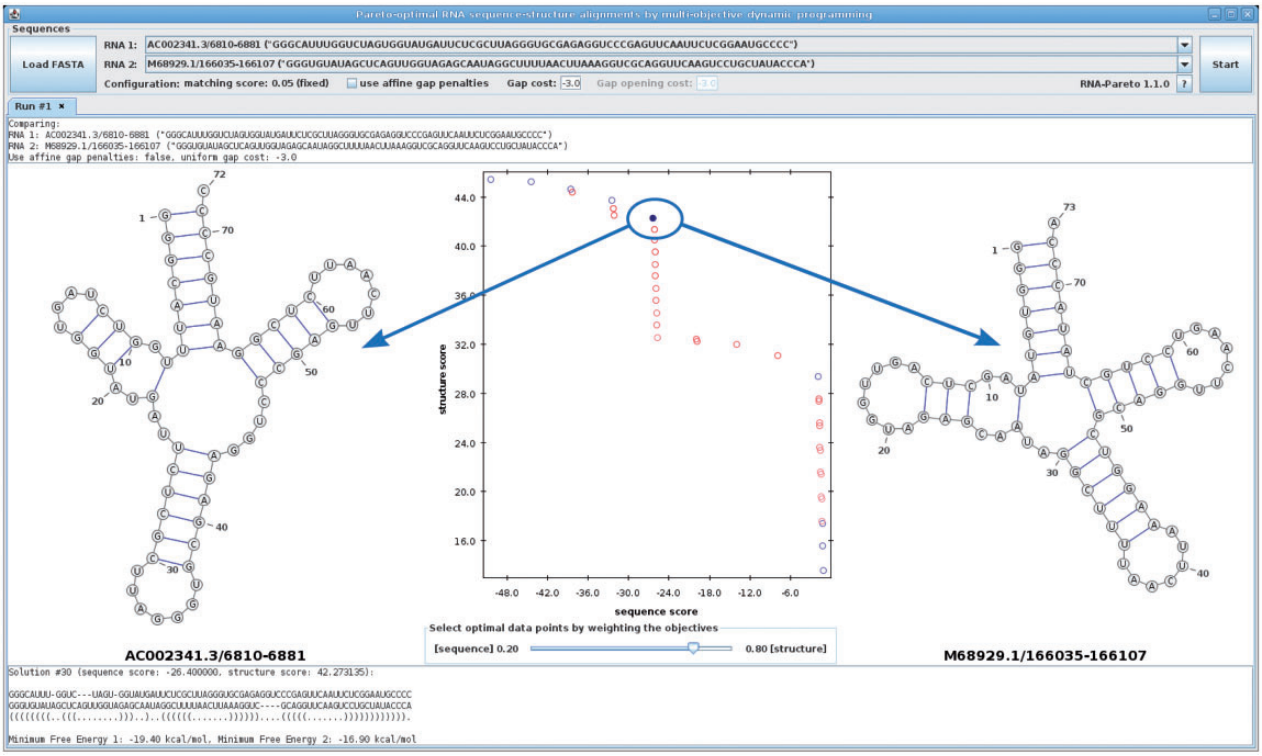


**Fig. 1.** Interactive RNA-Pareto graphical user interface. Upper panel: Import of two sequences from a FASTA file (RNA 1 and RNA 2). Center panel: Selectable scatterplot of the Pareto-optimal solutions (Pareto front) choosing an alignment (horizontal axis: sequence score; vertical axis: structure score). The selected solution is marked by a filled circle. Solutions on the convex hull of the Pareto front are shown as blue circles. The slider below the Pareto-front is hidden if a solution not on the convex hull is selected or one clicks into the background of this panel. On both sides (arrows), the two corresponding RNAs are drawn with the consensus structure of this solution. In the bottom panel, scores of the selected solution are given and the sequence alignment together with the consensus secondary structure is shown

well-known algorithms for mono-objective sequence-structure alignment (Hofacker *et al*., 2004; Will *et al*., 2007), and generalizes the dynamic programming approach to a vector valued scoring function. The structure information is incorporated using precomputed base pair probability matrices (McCaskill, 1990), which are computed using the Vienna RNA library (Lorenz *et al*., 2011).

## 2.2 Exploring optimal solutions

Our method results not in one solution, but a set of solutions. These solutions are all equally good with respect to the concept of Pareto-optimality, but differ in biological implication. To give the researcher a useful tool for the analysis of these solutions, we developed a platform independent graphical user interface. It allows the user to select two RNA sequences from a *FASTA* database file, for which the Pareto-set of solutions is then computed. The main view offers four different sections (Fig. 1). In the center, there is a 2D scatterplot of the Pareto-optimal scoring vectors. By using the mouse wheel, mouse clicks or keyboard shortcuts, it can be used to select a data point that corresponds to an optimal alignment. Under this subpanel there is a slider on which the user can set a specific weighting between the two objectives. As a result, the solution that maximizes the weighted sum of the two objectives is then automatically selected. Note that only those points that lie on the convex hull of the Pareto-set can maximize a weighted sum. This slider can be turned off by clicking into background or by selecting a solution that is not on the convex hull of the Pareto front. The sequence-structure alignment for the currently selected solution is displayed at the bottom, together with the minimum free energies of the two RNAs. On the left and on the right side, the two RNA sequences featuring the consensus secondary structure are drawn (*VARNA* drawing library, Darty *et al*., 2009). By right clicking, images can be rotated, printed or exported to various popular graphics formats.

## 3 CONCLUSION

Lifting the arbitrary restrictions of fixed weighting parameters, the sequence-structure alignment results in a set of mathematically equivalent solutions. If this set becomes large, a manual examination becomes infeasible. Since we do not want to restrict ourselves to a subset of these Pareto-optimal solutions, one needs an exploration aid. To this end, we developed an interactive tool to guide the researcher through the exploration process (cf. Shneiderman, 1996). It assists in generating reliable consensus secondary structures, and helps to better understand the interplay between sequence and structure in the alignment process.

*Funding*: This work is supported by the German Research Foundation (DFG, Scho302/8-2 to U.S. and MA1538/14-2 to A.M.); and the Federal Ministry of Education and Research (BMBF, Forschungskern SySTaR, project ID 0315894A to H.A.K.).

*Conflict of Interest*: none declared.

## REFERENCES

Darty,K. *et al.* (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Ehrgott,M. (2005) *Multicriteria Optimization*. 2nd edn. Springer Verlag, Berlin.

Gardner,P.P. *et al.* (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.

Havgaard,J.H. *et al.* (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, **3**, e193.

Hofacker,I.L. *et al.* (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.

Latchman,D. (2005) *Gene Regulation: A Eukaryotic Perspective*. Taylor & Francis Group, New York.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Mathews,D.H. (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics*, **21**, 2246–2253.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Müssel,C. *et al.* (2012) Multi-objective parameter selection for classifiers. *J. Stat. Softw.*, **46**, 1–27.

Roytberg,M. *et al.* (1999) Pareto-optimal alignment of biological sequences. *Biophysics*, **44**, 565–577.

Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.

Schnattinger,T. *et al.* (2012) Pareto-optimal RNA sequence-structure alignments. In: *9th International Workshop on Computational Systems Biology 2012 (WCSB 2012)*. Ulm, Germany, pp. 83–86.

Schnattinger,T. *et al.* (2013) Structural RNA alignment by multi-objective optimization. *Bioinformatics*, **29**, 1607–1013.

Shneiderman,B. (1996) The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of IEEE Symposium on Visual Languages*. IEEE Computer Society Press, Los Alamitos, CA, pp. 336–343.

Taneda,A. (2010) Multi-objective pairwise RNA sequence alignment. *Bioinformatics*, **26**, 2383–2390.

Taneda,A. (2011) A web server for multi-objective pairwise RNA sequence alignment with an index for selecting accurate alignments. *IPSJ Trans. Bioinform.*, **4**, 2–8.

Will,S. *et al.* (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol.* **3**, e65.

Zuker,M. and Sankoff,D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621.