Improved heuristic for pairwise RNA secondary structure prediction

Olivier Perriquet, Pedro Barahona

CENTRIA - Centre for Artificial Intelligence
Departamento de Informática, FCT/UNL
Quinta da Torre 2829-516 CAPARICA - Portugal

Abstract

We propose a refined heuristic for pairwise RNA secondary structure prediction based on the recursions of Sankoff [beta version available at http://centria.di.fct.unl.pt/~op/arnica.tar.gz].

Introduction

The discovery of new families of non-coding RNA (ncRNA) demands dedicated tools to predict their structure. The programs that compute secondary structure naturally organize themselves along several noticeable directions that implicitly depend on their context of use. When a large family of not too divergent homologous sequences is available, a multiple alignment of high quality can be obtained and covariation analysis [ED94,KH03] has proven to be very accurate in guessing the structure in that context. When only a small family of poorly conserved RNA sequences are available, a good starting alignment can hardly be constructed, resulting usually in the inapplicability of the methods based on pre-alignment. In that second context, another option is to seek the alignment and the structure at the same time. The outcome is then both a common structure and an alignement of the sequences respecting this structure. For this reason, this formulation of the question is usually termed *“structural alignment”*.

Sankoff [San85] pioneered the field with a set of recursions that optimally compute the best structural alignment of two RNA sequences when the structure is not known a priori. The algorithmic complexity, although polynomial ($0(n^6)$ in time and $0(n^4)$ in space for sequences with sizes in the order of magnitude n), remains prohibitive for real case applications without a good heuristic. Diverse ideas were applied to turn the Sankoff recursions usable. The first adaptation was DYNALIGN [MT02,HSM07] that reduced the complexity to $0(M^2n^2)$ in space and $0(M^3n^3)$ in time, where M is a user-tunable hard constant which bounds the allowed shift between the two sequences. FOLDALIGN [HLS+05,HTG07] combines other different restrictions, namely: “*alignment banding*” (like in DYNALIGN the maximal shift between the sequences is bounded by a constant δ), “*structural banding*” (local alignment, the maximal size for a common motif being bounded by a constant λ), “*multi-loop restriction*” (the structure bifurcativity is limited in multi-loops). The resulting complexity in FOLDALIGN becomes $0(n^2\ \lambda\ \delta)$ in space and $0(n^2\ \lambda^2\ \delta^2)$ in time. The Vienna RNA Package also proposes an alternative attempt PMComp [HBS04] (which is implemented as part of the RNAfold program) based on the McCaskill algorithm that computes the probabilities of base pairings for a single sequence [McC90]. The authors of FOLDALIGN then revisited the same idea [THG07].

Although these heuristics turn the algorithm of Sankoff usable on natural RNA sequences, still the method stays inapplicable once the sequences under consideration show too much divergence. When they have a large difference of length, poor conservation at the primary level or important structural variations, all Sankoff-based method either do not apply or - if they do - the banding heuristics prevent the algorithm from finding the correct structure, while the memory and time consumptions explode. One of the main drawbacks of these heuristics is what we could call their *global* nature: they cannot take advantage of local similarities in the sequences. Moreover, when performing a (global) structural alignment, the shift restriction (δ for FOLDALIGN, M for DYNALIGN) should be at least the difference of length between the sequences and the exploding complexities then turn again the recursions ineffective. FOLDALIGN bypasses the later difficulty by falling back onto local alignment when the case happens. In this paper, we propose a new Sankoff-based algorithm combined with a much more efficient heuristic, resulting in a huge gain, both in memory and time consumption, that allows us to apply it on cases that were unaffordable to other similar methods.

Method

Our program, that we named **ARNICA**, proceeds in three steps:

1. **Setting alignment constraints** – this first phase uses a variant of the standard alignment methods, reminiscent of the algorithm of Waterman [Wat83] to compute alignement constraints (ie. allowed and forbidden matching in the seeked alignment). These constraints are crucial to reduce the space and time consumption, as the size of the Sankoff matrix directly depends on the number of allowed matchings, as discussed in the next paragraph. A simple and unique user-specified threshold THD is used to tune the size of this search space.
2. **Setting structural constraints** – in the second phase, the pairing probabilities are computed independently for each sequence with the McCaskill algorithm [McC90]. Two bases for which the pairing probability is less than 1% are then not allowed to pair (this has no consequence on the computational space and time but simply increases the quality of the solutions found by discarding spurious base pairs).
3. **Seeking common folding and alignment** - the optimal folding and alignment is computed with the recursions of Sankoff. The scoring scheme used in that step takes only indirectly into account the stabilizing and destabilizing effects of stacking.

The size of the 4-dimensional matrix to be allocated in that latest step increases with the alignment threshold THD chosen in the first step. If the threshold is infinite, then any pair of bases is allowed to participate in the final alignment and there is no gain over the complexities of the Sankoff recursions. In practice however, THD is chosen small enough to keep in reasonable time and space consumption. Our framework is quite effective and far more efficient than any alignment banding heuristic, which is no more than a peculiar case of it. The algorithmic complexities for a Sankoff-based pairwise secondary structure alignment with alignment banding δ are 0(n^2 $δ^2$) in space and 0(n^3 $δ^3$) in time. They can be reformulated as O($α^2$) and O($α^3$), where α~δn is the size of the diagonal « band » corresponding to the alignment envelope induced by the banding heuristic. α is in fact the size of the *true* zone in the adjacency matrix of the graph of allowed matching bases in the alignment. As shown on Illustration 1, this zone is an exact diagonal band in the banding heuristic, whereas it can be a tunable zone (the dark area) in our framework.

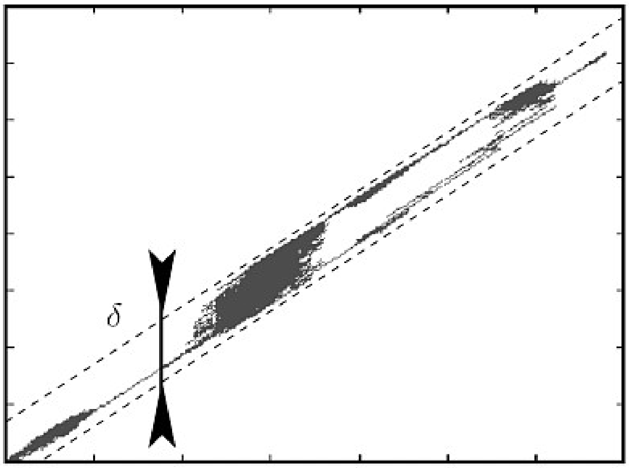

Illustration 1: Suboptimal alignments of two RNase P RNA (D.desulfuricans vs A.eutrophus) showing alternative alignment paths resulting from large zones of deletion. To reach any of these suboptimal alignments with a banding heuristic, the value chosen for the allowed shift δ has to encompass all the possible paths.

In its current beta version, **ARNICA** adopts a rather empirical scoring scheme for sequence alignment and a reduced energy model that partially takes into account the stabilizing effect of base pair stacking in stems. However, even in that simplified model, ARNICA already shows remarkable results on natural sequences and appear to be fully competitive with the other Sankoff-based methods, as discussed in the next section.

Results & Discussion

Focusing on the difficult context of poorly conserved RNA homologs, we have selected 7 RNase P RNA (the sequences are coming from the database of Brown [Bro99]) showing deep variations in structure, which make them difficult candidates for all Sankoff-based methods, as mentioned by [GG04]. The average identity is 64% and the average length around 400. Table 1 gives the average performances of ARNICA with different values for the threshold, compared to FOLDALIGN, which is one of the prominent Sankoff-based structural alignment methods. For FOLDALIGN, we use the default option (local alignment), as the global option can seldom be chosen, the difference of length being too important between the sequences. However, half of the time, the percentage of sequence covered by the structural alignment predicted is close to

100% (in that case, comparing the performances is fully meaningful, as the difference between local and global is weak). The results are averaged over all possible pairwise alignments (we did not compute the standard deviation: ARNICA performs better any time), they are expressed in terms of specificity and sensitivity, which are natural measures for a binary classification test performance assessment (specificity could favourably be replaced by more sophisticated measures dedicated to RNA, such as *positive predictive value* defined by [HSM08] for instance, but at this stage we opted to comply to the standard).

	option	***specificity***	***sensitivity***	***time (s)***	***space (Mb)***
FOLDALIGN	local	56.2 %	40.5 %	1107	142.1
ARNICA	THD 0	73.6 %	43.3 %	6	16.5
	THD 10	74.1 %	45.9 %	7	16.7
	THD 30	75.7 %	51.0 %	10	17.5
	THD 50	76.2 %	55.0 %	20	19.1
	THD 80	75.7 %	58.3 %	77	23.7
	THD 100	74.7 %	58.7 %	148	29.0
	THD 150	73.9 %	60.5 %	536	46.5
	THD 200	73.2 %	61.0 %	1079	64.4

Table 1: Average performance of ARNICA and FOLDALIGN on a set of RNase P (alpha subdivision). All the tests were run on an IBM thinkpad T40 (pentium 1.5GHz - RAM 512Mb). Sequences are from: http://www.mbio.ncsu.edu/RnaseP/alpha-purples.html.

On this data set, ARNICA performs better from any point of view (specificity, sensitivity, time and memory usage) remaining fast and low memory consuming whereas FOLDALIGN seems to be limited by its heuristic restrictions. The average gain in computational time and memory with a threshold 100 is by a factor 7 and 5, while the specificity of ARNICA is neighboring 75%. This globally stable behavior despite the divergence of structure is a promising advantage for the integration of a more complete model. ARNICA is still a work in progress, our ongoing developments aim at incorporating a more complete thermodynamic model while refining even further the method by incorporating dynamic boolean restraints (allowed and forbidden matchings for the alignement but also for the pairings).

References

[Bro99] J.W. Brown. The Ribonuclease P database. NAR, 27(314), 1999.

[ED94] S.R. Eddy, R. Durbin. RNA sequence analysis using covariance models. NAR, 22:2079–2088, 1994.

[GG04] P. P. Gardner and R. Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics, 5(1), September 2004.

[HBS04] I. L. Hofacker, S. H. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. Bioinformatics, 20(14):2222–2227, September 2004.

[HLSG05] Jakob Hull Havgaard, Rune B. Lyngsø, Gary D. Stormo, and Jan Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics, 21(9):1815–1824, 2005.

[HSM07] Arif O. Harmanci, Gaurav Sharma, and David H. Mathews. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in dynalign. BMC Bioinformatics, 8:130, April 2007.

[HSM08] Arif O. Harmanci, Gaurav Sharma, and David H. Mathews. PARTS: Probabilistic Alignment for RNA joinT Secondary structure prediction. NAR - Apr;36(7):2406-17, 2008.

[HTG07] Jakob H. Havgaard, Elfar Torarinsson, and Jan Gorodkin. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Computational Biology, 3(10):e193, Oct. 2007.

[KH03] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res, 31(13):3423–3428, July 2003.

[McC90] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers, 29:1105–1119, 1990.

[MT02] D.H. Mathews and D.H. Turner. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. JMB - Mar 22;317(2):191-203, 2002.

[San85] D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J. Appl. Math., 45(5):810–825, 1985.

[THG07] Elfar Torarinsson, Jakob H H. Havgaard, and Jan Gorodkin. Multiple structural alignment and clustering of RNA sequences. Bioinformatics, February 2007.

[Wat83] M. S. Watermann. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. Applied Mathematical Sciences, 80:3123–3124, May 1983.

Contact: olivier@perriquet.net
Software: http://centria.di.fct.unl.pt/~op/arnica.tar.gz
(available: beta version including sample tests)