Simultaneous Alignment and Structure Prediction of RNAs

Are Three Input Sequences Better Than Two?

Beeta Masoumi and Marcel Turcotte

School of Information Technology and Engineering,
University of Ottawa,
Ottawa, Ontario, Canada
{bmasoumi, turcotte}@site.uottawa.ca

Abstract. Comparative RNA sequence analyses have contributed remarkably accurate predictions. The recent determination of the 30S and 50S ribosomal subunits brought more supporting evidence. Several inference tools are combining free-energy minimisation and comparative analysis to improve the quality of secondary structure predictions. Using many input sequences should improve the accuracy, reduce the likelihood that bad predictions are made, but also lower the sensitivity. To investigate these claims, we have extended the software system Dynalign to use three input sequences, rather than two, and tested our algorithm with 10 tRNAs and 13 5S rRNAs. The following hypotheses were tested: 1) the use of three input sequences improves the average accuracy compared to predictions based on two input sequences. Also, it should be less likely that all three input sequences simultaneously fold into a bad free-energy minimum compared to predictions based on two sequences, consequently, 2) the worse prediction (minimum accuracy) for any sequence should be more accurate when three input sequences are used rather than two. Finally, the consensus structure of three sequences is probably less representative of the individual sequences. 3) Therefore, the average coverage should be less.

1 Introduction

The repertoire of known non-protein coding RNAs (ncRNAs) is growing rapidly[1]. The housekeeping roles of RNAs, such as those of the tRNA, rRNA, RNAseP, snRNA and snoRNA, were established early. In the recent years, it has become clear that RNAs also have important regulatory functions. Examples include microRNAs, which regulate the expression of protein genes by targeting a complementary region of their mRNAs. MicroRNAs constitute one of the most abundant class of regulatory molecules, and are key to many developmental processes[2]. Several discoveries collectively demonstrate that untranslated messenger RNAs can sense the level of metabolites, and modulate the expression of certain genes accordingly. Those RNAs are referred to as RNA sensors and riboswitches[3, 4]. Post-transcriptional regulation of gene expression often involves

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3515, pp. 936–943, 2005.

secondary structure elements located in the untranslated regions of mRNAs[5]. Consequently, detailed knowledge of RNA secondary and tertiary structure is essential to help understand RNA functions.

RNA secondary structure prediction methods have been thoroughly evaluated. In particular, Gardner and Giegerich have performed a comprehensive evaluation of comparative RNA structure prediction methods[6]. Doshi et al. reviewed specifically free-energy minimisation methods that are using the nearest-neighbour model[7]. One of their main conclusions is that free-energy minimisation methods based on the nearest-neighbour model work best for shorter sequences, such as tRNA or 5S rRNA, for which they reported an average accuracy for predictions of 69% and 71% respectively.

Recently, Mathews and Turner developed, and published, a software system combining free-energy minimisation and comparative sequence analysis for finding the minimum free-energy structure common to two input sequences[8]. The computer system, called Dynalign, greatly improves the accuracy of secondary structure predictions compared to free-energy minimisation alone.

Herein, we extend this algorithm to use three input sequences, rather than two, and investigate the performance of the new computer program. We called this software system eXtended-Dynalign, or X-Dynalign for short, to emphasise its origin.

This paper is organised as follows. Section 2 outlines the algorithm. In Section 3, the main hypotheses to be empirically evaluated are laid out, the datasets are described, and the evaluation measures are defined. Section 4 presents the results. Section 5 concludes and discusses the results.

2 Algorithm

Dynalign is a pragmatic implementation of the algorithm proposed by Sankoff for solving simultaneously the RNA folding and alignment problems[9]. Dynalign is restricted to two input sequences, while the original proposal was formulated for an arbitrary set of N input sequences. Also, Dynalign introduces a constraint on the maximum distance between aligned nucleotides so as to reduce the execution time. This is analogous to the banding technique that is used for sequence alignment.

X-Dynalign is a direct extension of Dynalign. It takes as input three sequences and produces a three-way sequence alignment as well as a common secondary structure. The objective function consists of a linear combination of the free-energy of each sequence, given the common secondary structure, and an empirical term for gap penalties.

$$\Delta G^{\circ}_{\text{total}} = \Delta G^{\circ}_{\text{sequence 1}} + \Delta G^{\circ}_{\text{sequence 2}} + \Delta G^{\circ}_{\text{sequence 3}} + \Delta G^{\circ}_{\text{gaps}}$$

where $\Delta G^{\circ}_{\text{sequence } i}$, for $i \in \{1, 2, 3\}$, represents the conformational free-energy of the sequence i when folded onto the common secondary structure, according to the nearest-neighbour model.

Three sets of recurrence equations are defining the objective function: W, V and $W9$. Equations of the form $W(i, j, k, l, m, n)$ represent the minimum free-energy for the optimal alignment and structure prediction of $S_1[i..j]$, $S_2[k..l]$ and $S_3[m..n]$, when i, k and m are aligned, and j, l and n are also aligned, where S_i denotes the sequence i and $S_i[a..b]$ represents the fragment of S_i comprising the nucleotides $a, a + 1 \ldots b$. Equations of the form $V(i, j, k, l, m, n)$ represent the minimum free-energy assuming that i and j, k and l, and m and n are simultaneously aligned but also base-paired. Finally, $W9(i, k, m)$ represents the minimum free-energy for the prefix alignment of $S_1[1..i], S_2[1..k]$ and $S_3[1..m]$. A detailed description of the recurrence equations can be found in [10]. The recurrence equations are solved using dynamic programming. The algorithm requires $\mathcal{O}(|S_1|^2 M^4)$ space and $\mathcal{O}(|S_1|^3 M^6)$ time, where M is a constant that limits the maximum distance between aligned nucleotides.

3 Methodology

3.1 Experiments

The following hypotheses are tested: 1) the use of three input sequences should improve the average accuracy compared to predictions based on two input sequences. When three input sequences are used, the likelihood that they all three fold into a bad free-energy minimum should be less than when two input sequences are used, consequently, 2) the worse prediction (minimum accuracy) should be more accurate when three input sequences are used rather than two. Finally, the secondary structure common to three input sequences should be less representative of the individual sequences, consequently, 3) the average coverage should be less. But first, we determine empirically the optimum gap penalties for these datasets.

3.2 Datasets

Input sequences were selected such that they can be aligned optimally with a small value of M. Obviously, this information would not be known in advance in most cases. Also, the input sequences were filtered so that the maximum pairwise identity was less than 90%. A total of 10 tRNA sequences from the original paper were used. Their pairwise sequence identity varies from 27.3 to 68.8 %. The secondary structure assignments were taken from the compilation by Sprinzl et al.[11, 12]. A set of 13 5S rRNA sequences was built using information obtained from the Comparative RNA Web Site[13, 14, 15]. Their pairwise sequence identity varies from 47.2 to 88.2% .

3.3 Performance Measures

We call **references**, the secondary structures that were obtained from the tRNA compilation by Sprinzl and the Comparative RNA Web Site. We define as **true**

positives (TP) the base pairs that are occurring in both structures, reference and predicted, **false positives** (FP), the base pairs that are occurring in the predicted structure but not in the reference one, and **false negatives** (FN), the base pairs that are occurring in the reference structure but not in the predicted one. Offsets were not allowed.

The **positive predictive value** (sometimes called PPV, specificity or accuracy) is defined as the fraction of the predicted base pairs that are also present in the reference structure, TP/(TP + FP). The **sensitivity** (coverage) is defined as the fraction of the base pairs from the reference structure that are correctly predicted, TP/(TP + FN). Finally, we also measured the **Matthews Correlation Coefficient**, as defined by Gorodkin, Stricklin and Stormo[16]:

$$\sqrt{\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \times \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}}$$

4 Results

4.1 Calibrating Gap Penalties

In [8], the optimal gap penalty was found to depend on the class of RNA; 2.0 and 0.4 Kcal/mol for the tRNA and 5S rRNA, respectively. Accordingly, we performed two sets of experiments to measure the effect of various gap penalty scores on PPV, sensitivity and MCC. Since these experiments are time consuming, only six gap penalty scores were tested, $0.0, 0.25, 0.5, 1.0, 2.0, 4.0$, and only triples that can be aligned with a small value of M, here 5, were selected. In all, 105 and 90 predictions were made for the tRNA and 5S rRNA, respectively. For the experiments presented herein, we have chosen a gap penalty score of 1.0 Kcal/mol, because it corresponds to the maximum sensitivity for both datasets, tRNA and 5S rRNA.

4.2 Comparative Analysis

We present the analysis of the tRNA data first. Nine runs, 27 predictions, were made using X-Dynalign, while 19 runs, 38 predictions, were made using Dynalign. The mean PPV, sensitivity and MCC are 96.8 ± 7.6, 94.4 ± 7.5 and 95.6 ± 7.3 for X-Dynalign, and 92.1 ± 14.6, 89.1 ± 15.7 and 90.5 ± 15.0 for Dynalign. Our data represent a subset of that of Mathews and Turner, the PPV for Dynalign measured on this subset is 5.7 percentage points higher than theirs. We observe that the use of three sequences improves all three indices and reduces their variance, for this particular dataset.

Table 1 presents the performance indices per sequence. Dynalign performed well in the best case scenario. For all the sequences, it was possible to find a pair of input sequences having a high positive predictive value. The maximum PPV for every entry is 100, except for that of RS0380. Further analysis shows that the structure of RS0380 (tRNA$^{\mathrm{Asp}}$ *Haloferax volcanii*) has an extra stem

Table 1. PPV for the tRNA dataset. The subscripts *xd* and *d* are designating X-Dynalign and Dynalign respectively. *N* is the number of predictions

Id	N_{xd}	N_d	Min_{xd}	Min_d	Max_{xd}	Max_d	Ave_{xd}	Ave_d
RD0260	4	5	100	80	100	100	100.0	96.0
RD0500	4	5	76	45	100	100	82.2	80.8
RD4800	5	5	100	80	100	100	100.0	96.0
RE2140	2	4	100	100	100	100	100.0	100.0
RE6781	2	4	100	77	100	100	100.0	94.3
RF6320	4	5	95	45	100	100	96.4	89.1
RL0503	1	2	100	100	100	100	100.0	100.0
RL1141	2	3	100	70	100	100	100.0	90.3
RS0380	1	2	100	83	100	87	100.0	85.2
RS1141	2	3	100	70	100	100	100.0	90.3

in the variable loop, that X-Dynalign predicted more accurately. For 9 out of 10 experiments, the maximum sensitivity for X-Dynalign equals or exceeds that of Dynalign.

Both algorithms are seeking to find a structure that minimises a linear combination of the free-energy of each input sequence given the common structure. Using three input sequences should have a positive impact on the worse case scenario. It should be less likely that all three input sequences jointly fold into the wrong minimum free-energy structure than with two input sequences. Our data support this observation, for all the entries the minimum PPV for X-Dynalign is the same or better than that of Dynalign. For 8 out of 10 sequences, the minimum PPV is 100, in one case, the minimum PPV is 95, and for one case the minimum PPV is 76. The two sequences leading to the worse predictions are RD0500 and RF6320, see Figure 1. Dynalign produces an elongated structure.

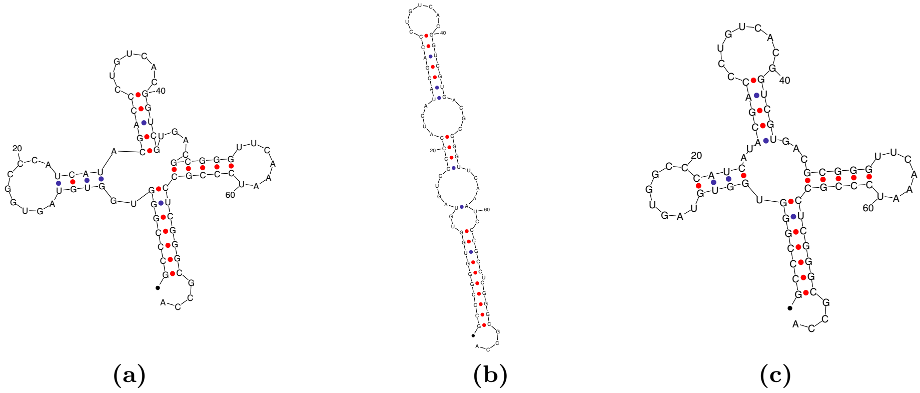

Fig. 1. Reference (a), Dynalign (b) and X-Dynalign (c) secondary structures for the tRNA RD0500

Table 2. PPV for the 5S dataset

Id	N_{xd}	N_d	Min_{xd}	Min_d	Max_{xd}	Max_d	Ave_{xd}	Ave_d
AJ131594	2	3	100	91	100	100	100.0	94.5
AJ251080	6	5	88	82	90	86	90.3	84.8
D11460	6	5	87	66	87	88	87.6	79.4
K02682	8	9	63	88	100	97	89.1	92.0
M10816	3	4	90	85	90	88	90.7	87.8
M16532	1	2	94	77	94	85	94.1	81.8
M25591	6	5	87	82	90	86	89.8	84.8
V00336	3	4	75	65	100	100	91.9	91.4
X02024	9	6	88	82	90	88	90.1	85.8
X02627	1	2	100	92	100	100	100.0	96.0
X04585	2	3	72	68	94	93	83.4	82.7
X08000	5	5	90	88	90	90	90.6	89.4
X08002	5	5	90	88	90	90	90.6	89.4

However, using a third sequence increases the accuracy by more than 30 percentage points. The structure produced by X-Dynalign has the overall cloverleaf shape, however, the nucleotides of the first part and second part of the D-arm are shifted by one and two positions, respectively. The minimum coverage is generally good. For all the sequences the coverage is 75% or better. For all the tests the coverage for X-Dynalign is the same as Dynalign or better.

For our second test set, we have 19 runs, 57 predictions, using X-Dynalign, and 29 runs, 58 predictions, using Dynalign. The mean PPV, sensitivity and MCC are 90.3 ± 5.8, 76.6 ± 5.3 and 83.2 ± 5.5 for X-Dynalign, and 87.7 ± 7.4, 79.2 ± 6.7 and 83.3 ± 6.7 for Dynalign. For this particular dataset, the performance of both systems is comparable on the basis of the Matthews correlation coefficient. What is gained in accuracy is lost in sensitivity.

Table 2 presents the performance indices per sequence. Using three input sequences improves the worse (PPV) prediction for 12 out of 13 sequences. Also, for 10 out of 13 sequences, the minimum PPV obtained is 85% or more. The minimum sensitivity is the same or improved for 11 out of 13 sequences. However, the maximum sensitivity exceeds that of Dynalign for 2 out of 13 sequences.

The prediction of the 5S rRNA of *Micrococcus lutus* (K02682) has an accuracy of 63% only. We believe this is due to the fact that single base pair insertion has not been implemented yet in X-Dynalign. In the triple K02682, V00336 and X04585, the structure of *Rhodobacter capsulatus* (X04585) has a shorter helix IV, 7 base pairs compared to 8 for the other two structures.

5 Conclusion and Discussion

We have extended the software system Dynalign to use three input sequences, rather than two. The resulting system is called eXtended-Dynalign (X-Dynalign for short). Its time/space complexity limits its application to 1) short sequences (say less than

200 nt) and 2) sequences that can be aligned optimally with a small value of M (less than 6), where M is the maximum distance of aligned positions.

The strengths of Dynalign carry over to the new system. Namely, it improves the accuracy of secondary structure predictions compared to predictions based on a single input sequence. It requires no sequence homology.

It also shares some of its limitations. In particular, the gap penalties are treated as a separate term in the objective function. The optimal value has to be determined empirically. In [8], it was found that the optimal value for this term depends on the class of RNA studied. In our limited experiments, the dependency seems less important. It also seems that there is large plateau were several gap penalty scores are leading to a nearly optimal solution; w.r.t. PPV, for example. Our key conclusions are:

- The lowest PPV for any prediction is generally improved when using three input sequences;
- The average accuracy is improved;
- The average sensitivity of the algorithm slightly degraded for the 5S rRNA dataset. However, a per sequence analysis shows that the majority of the lowest sensitivity scores are higher for X-Dyanlign than Dynalign;
- X-Dynalign is able to reproduce subtle details, such as the prediction of a stem in the variable region of certain tRNAs.

There are several obvious directions for extending this class of algorithms, such as handling pseudo-knots and reporting suboptimal structures. However, one of the most urgent improvement is to reduce the time/space complexity. Several runs presented herein take up to week to compute on some of the fastest processors available today.

The detailed knowledge of RNA secondary structure is essential for understanding the sequence-structure-function relationships. X-Dynalign takes advantage of the paramount of data that is accumulating in sequence databases. Because it requires no sequence homology, X-Dynalign should be useful to comparative RNA sequence analyses.

6 Availability

The source code, written in C++, as well as the scripts for calculating the performance indices are made available under the GNU General Public Licence from `http://bio.site.uottawa.ca/software/x-dynalign`. Supplementary material, including additional tables and figures, can be found on our web site.

References

1. Storz, G.: An expanding universe of noncoding RNAs. Science **296** (2002) 1260–1263
2. Bartel, D.P.: MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell **116** (2004) 281–297

3. Nudler, E., Mironov, A.X.: The riboswitch control of bacterial metabolism. Trends Biol. Sci. **29** (2004) 11–17
4. Lai, E.C.: RNA sensors and riboswitches: Self-regulating messages. Current Biology **13** (2003) R285–R291
5. Mignoe, F., Gissi, C., Liuni, S., Pesole, G.: Untranslated regions of mRNAs. Genome Biology **3** (2003) 0004.1–0004.10
6. Gardner, P.P., Robert, G.: A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics **5** (2004) 140
7. Doshi, K.J., Cannone, J.J., Cobaugh, C.W., Gutell, R.R.: Evaluation of the suitability of free-energy minimization using mearest-neighbor energy parameters for rna secondary structure prediction. BMC Bioinformatics **5** (2004) 105
8. Mathews, D., Turner, D.: Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. J. Mol. Biol. **317** (2002) 191–203
9. Sankoff, D.: Simultaneous solution of RNA folding, alignment and protosequence problems. SIAM J. Appl. Math. **45** (1985) 810–825
10. Masoumi, B.: A dynamic programming algorithm for the simultaneous alignment and structure prediction of three RNA sequences. Master's thesis, School of Information Technology and Engineering, Faculty of Engineering, University of Ottawa (2005)
11. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., Steinberg, S.: Compilation of tRNA sequences and sequences of tRNA genes. Nucl. Acids Res. **26** (1998) 148–153
12. Sprinzl, M., Vassilenko, K.S.: Compilation of tRNA sequences and sequences of tRNA genes. http://www.uni-bayreuth.de/departments/biochemie/trna (2003)
13. Gutell, R.R.: Comparative RNA web site. http://www.rna.icmb.utexas.edu (2004)
14. Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., Pande, N., Shang, Z., Yu, N., Gutell, R.R.: The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics **3** (2002)
15. Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., Pande, N., Shang, Z., Yu, N., Gutell, R.R.: The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs: Corrections. BMC Bioinformatics **3** (2002)
16. Gorodkin, J., Stricklin, S.L., Stormo, G.D.: Discovering common stem-loop motifs in unaligned RNA sequences. Nucl. Acids Res. **29** (2001) 2135–2144