ISMB-97, 120-123

Finding Common Sequence and Structure Motifs in a set of RNA Sequences

Jan Gorodkin[1], Laurie J. Heyer[2], Gary D. Stormo[3]
[1]Center for Biological Sequence Analysis, The Technical University of Denmark, Building 206, 2800 Lyngby, Denmark, (gorodkin@cbs.dtu.dk)
[2]Department of Applied Mathematics, (heyer@colorado.edu)
[3]Department of Molecular, Cellular and Developmental Biology, (stormo@colorado.edu)
University of Colorado, Boulder, CO 80309, USA *

Abstract

We present a computational scheme to search for the most common motif, composed of a combination of sequence and structure constraints, among a collection of RNA sequences. The method uses a simplified version of the Sankoff algorithm for simultaneous folding and alignment of RNA sequences, but maintains tractability by constructing multi–sequence alignments from pairwise comparisons. The overall method has similarities to both CLUSTAL and CONSENSUS, but the core algorithm assures that the pairwise alignments are optimized for both sequence and structure conservation. Example solutions, and comparisons with other approaches, are provided. The solutions include finding consensus structures identical to published ones.

Introduction

Locating sequence as well as structure motifs in a set of RNA sequences is of general interest. For example all of the methods that do structure prediction based on phylogenetic data require that the alignment of the sequences be known in advance. That alignment process is usually done by hand and is one of the biggest problems in using that approach. The method presented here promises to automate the alignment and structure determination process, and can be used on normal phylogenetic data, on SELEX (Tuerk & Gold 1990) type data where the RNAs have been selected *in vitro*, and when one has a collection of genes that are coordinately regulated at the translational level. With the rapid increase of the genomic databases, and the expanded use of selected RNAs, the need for such a structural alignment method which is fast and accurate is still growing.

The problem of finding the best structural alignment among N sequences of size L, has been solved by (Sankoff 1985), but unfortunately the time complexity is $O(L^{3N})$, which is impractical. However just finding the best alignment between N sequences is not always what we are interested in; some of the N sequences might not be really related to the rest, but have been included in the set erroneously; or some of the sequences could be functionally related but fall into two (or more) structural classes so that there is not a single motif that is common to all of them. Here we search for *the best core structure shared by $M \leq N$ of the sequences.* The idea is that one should be able to proceed with other existing RNA folding methods using this core structure as a solid starting point. However, here we only present structural alignments directly obtained using our approach FOLDALIGN.

To reduce the time usage, we present an algorithm consisting of a core alignment algorithm and a greedy algorithm which successively adds new sequences to be aligned using the core algorithm. The core algorithm, which essentially is the approach of (Sankoff 1985) (for two sequences), structurally aligns two sequences (or two collections of aligned sequences), essentially using a combination of plain sequence alignment (Smith & Waterman 1981) and a basic algorithm which maximizes the number of basepairs (Nussinov & Jacobson 1980). In this approach branching configurations are neglected thus reducing the time complexity from $O(L^6)$ to $O(L^4)$, allowing for many more comparisons using the greedy part of the overall algorithm.

Method

Here we present the basic ideas of the core algorithm and the greedy algorithm, but refer to (Gorodkin, Heyer, & Stormo 1997) for further details.

The best structural alignment of the subsequences $(a_i, \ldots, a_j)$ and $(b_k, \ldots, b_l)$ of the sequences $\vec{a}$ and $\vec{b}$ can be found by introducing a dynamic programming algorithm, presented below. Similar to the sequence alignment matrix, which is a 5×5 matrix (when including gaps) defining the similarity for substituting

bases with each other, we can here define a 25×25 scoring matrix, $S_{ij,kl}$, over all quadruples, including gaps. Such a matrix defines the similarity of substituting one pair of bases with another. This score matrix can, in a natural way, be directly related to the usual sequence alignment matrix along with another 5×5 matrix listing the degree of complementarity of between the various bases. Hence for any scoring matrix and the constraint of non-branching structures, the maximum scoring subsequence alignment can be obtained from the four-dimensional matrix $D_{ij,kl}$ from the following recursion:

$$D_{ij,kl} = \max \left\{ \begin{array}{l} D_{(i+1)(j-1),(k+1)(l-1)} + S_{ij,kl}, \\ D_{i(j-1),(k+1)(l-1)} + S_{-j,kl}, \\ D_{(i+1)j,(k+1)(l-1)} + S_{i-,kl}, \\ D_{(i+1)(j-1),k(l-1)} + S_{ij,-l}, \\ D_{(i+1)(j-1),(k+1)l} + S_{ij,k-}, \\ D_{(i+1)(j-1),kl} + S_{ij,--}, \\ D_{ij,(k+1)(l-1)} + S_{--,kl}, \\ D_{(i+1)j,(k+1)l} + S_{i-,k-}, \\ D_{i(j-1),k(l-1)} + S_{-j,-l}, \\ D_{(i+1)j,k(l-1)} + S_{i-,-l}, \\ D_{i(j-1),(k+1)l} + S_{-j,k-}, \\ D_{(i+1)j,kl} + S_{i-,--}, \\ D_{i(j-1),kl} + S_{-j,--}, \\ D_{ij,(k+1)l} + S_{--,k-}, \\ D_{ij,k(l-1)} + S_{--,-l} \end{array} \right\}$$

The maximal $D_{ij,kl}$ provides the maximal similarity between the two sequences. One difference between this recursion and that of plain sequence alignment, is there is no zero value. This is because the matrix $\underline{\underline{D}}$ indicates both ends of the alignment, a_i with b_k and a_j with b_l, and also because we have to allow for negative values to occur within the complete alignment.

By considering a structural alignment as one consensus sequence with preassigned structure, the dynamic programming algorithm can be extended to aligning two collections of alignments against each other. However, in that alignment process each of the sequences is aligned to every other sequence, and the structure component of the $\underline{\underline{S}}$ matrix is applied for aligned positions for which both collections contain basepairs.

With this general scheme of aligning any size collection to any size collection, we may build up greedy algorithms like those used for CLUSTALW (Thompson, Higgins, & Gibson 1994) and CONSENSUS (Hertz, Hartzell,III, & Stormo 1990). The basic idea of building up these kinds of comparisons is to pairwise compare all individual sequences to each other, then compare all the pairwise alignments to all the individual sequences, such that no sequence appears more than once in each comparison. The next greedy step would be to align all the triplet alignments to the individual sequences, and compare all the pairwise alignments to themselves, still such that no sequence appears more than once in each final alignment. The algorithm may then be continued until all sequences have been compared in a final alignment. Clearly this approach explodes in time. Thus we need efficient procedures to eliminate alignments that are not useful. For the results presented below we only include comparisons between single sequences and $r-1$ sequences to obtain alignments of size r. And we use a threshold s (typically 30) of the best alignments after each "round" r. In this approach, the time complexity at each round is $O(L^4 r^2 s(N-r))$, for a total complexity of $O(L^4 N^4 s)$. In selecting the "best" alignment of $M \leq N$ sequences, there is a problem that scores increase with the number of sequences. As a preliminary analysis we look at a few best alignments of each round r, and compare score versus alignment length, and find a trade off from which the final alignment is chosen.

Results

We select four published data sets for investigation, all from SELEX experiments and for which a consensus structure has been proposed. The first set (H1) of RNA was found to bind to the human immunodeficiency virus type 1 rev protein (Tuerk *et al.* 1992). The second set (H2) contains a pseudoknot with specific affinity for HIV-1-RT (Tuerk, MacDougal, & Gold 1992). The third set (THEO) of RNA binds to the bronchodilator theophylline (Jenison *et al.* 1994). The fourth set (R17) is RNA ligands for the bacteriophage R17 coat protein (Schneider, Tuerk, & Gold 1992). The length of all the sequences is in the range of 30 to 50 nucleotides, and the data set sizes range from 13 to 36 sequences.

Using other approaches

To compare with our scheme presented above we first tried a few other methods which are publicly available.

First, we performed multiple alignment of the SELEX data by using the program CLUSTALW with default parameters. Since this program performs multiple alignment based on sequence conservation alone, we do not expect it to identify structurally conserved regions. As expected, on data sets with significant amounts of sequence conservation it does fairly well at identifying those. In the R17 data, the conserved hairpin loop with A bulge is aligned, but it does not provide a consistent alignment of the conserved stem region. For the H2 data set, which contains a pseudoknot, where the one basepair region is conserved in sequence, but the other is not, only the sequence con-

served region is found. The THEO data contains sequence conservation, but it is misaligned in the two subclasses. The H1 data has less sequence conservation than the others and is not aligned well.

Next, we applied the COVE program (Eddy & Durbin 1994) to structurally align the data sets. This stochastic context–free grammar approach which considers mutual information is similar to the work of (Sakakibara *et al.* 1994). COVE performs global alignment on a collection of sequences and has been shown to perform well on tRNA for which a global (consensus) structure is defined. It is well known that it is hard to find local features using a global alignment procedure, so we did not expect this package to perform well on the SELEX data sets, and it did not. Using the same data sets as for CLUSTALW we did not find any strong signals for consensus structures, not even if we used the CLUSTALW alignments as input to the program. The multiple structural alignment method presented here can be used to construct a core model which can then be used by COVE to extend and refine that model.

The Vienna RNA package (Hofacker *et al.* 1994) (see references therein) includes a program RNAfold to find the minimum free energy structures, and a program RNAdistance to find the distance between two structures in terms of the smallest cost along the editing path when representing the structures as trees. We folded each sequence in the data sets and found that many of the structures resembled the published structures, when neglecting the H2 pseudoknot. Using RNAdistance to pairwise compare the structures and appropriate cut–off scores to select reasonable comparisons, we found a number of sequences with similar structures (in the respective data sets) were clustered together. However, we did not find a strongest common structure, or the consensus structure for the respective data sets.

FOLDALIGN

For each of the investigated data sets we *did* find the subset with the strongest common motif in sequence and structure, matching what has been published.

The data set H1 has been assigned three structural classes. The classes all contain the same structural elements, the largest class consisting of 10 sequences. However, three of them use the constant SELEX regions in the basepairing (which were not included in our data sets), so we would not expect to find more than seven of them with conserved structure. For those seven sequences FOLDALIGN finds the published alignment as shown in Table 1. Furthermore FOLDALIGN succeeds in merging the two strongest classes, getting only one sequence wrong in the alignment. (All of the sequences in the third class utilize part of the constant region in their structures, and so cannot be aligned properly with the other sequences.) The additional structure identified by FOLDALIGN (Table 1), but not included in the published consensus structure, is included in the consensus structure for the largest class (Figure 3 of (Tuerk *et al.* 1992)).

Table 1: The strongest aligned class of the H1 data set. The parentheses indicate predicted basepairing, the underscore complete matching for a column. The numbers refers to the published sequence labels. Only the aligned part of the sequences are shown.

```
GGAUUUGAGAUACAC-GGAA-GUGGACUCUCC  17
GCC-UUGAGAUACACUAUAUAGUGGAC-CGGC  5
GGC-UGGAGAUACAAACUAU-UUGG-CUCGCC  4a
AUU---GAGAAACAC-GUUU-GUGGACUCGGU  6b
ACC-UUGAGGUACUC-UUAA-CAGG-CUCGGU  11
GCA-UUGAGAAACAC-GUUU-GUGGACUCUGU  6a
GAA-UUGAGAAACAC--UAA-CUGGCCUCUUU  14
(((.........((........)).....)))  (publ.)
(((...(_(.._((........))_.).))))  (FOLDALIGN)
```

Even better results are obtained for the data set H2. As mentioned this data set contains a pseudoknot, two overlapping stem-loop regions, and therefore violates the knot constraint in dynamic programming. One stem region is highly conserved in sequence, and the other has almost no sequence conservation. FOLDALIGN aligns the sequence conserved regions based on their sequence alignment, but at the same time aligns the other stem region which only is conserved in structure (see Table 2). We anticipate that other folding approaches, given this alignment, easily will predict the pseudoknot (Cary & Stormo 1995).

Table 2: The strongest aligned class of the H2 data set. The parentheses indicate basepairing (and the square brackets for pseudoknot), the underscore complete alignment for a column. The numbers refers to the published sequence labels. Only the aligned part of the sequences are shown.

```
CCAGAGGCCCAACUGGUAAACGGGC  1.17
CCG-AAGCUCAACGGGAUAAUGAGC  2.4a
CCG-AAGCCGAACGGGAAAACCGGC  1.3a
CC-CAAGCGC-AGGGGAGAA-GCGC  1.6
CCG-ACGCCA-ACGGGAGAA-UGGC  1.8
CCGUUUUCAG-UCGGGAAAAACUGA  1.1
CCGUUACUCC-UCGGGAUAAAGGAG  2.11
CCGUAAGAGG-ACGGGAUAAACCUC  2.7a
CCG-UAGGAG-GCGGGAUAU-CUCC  2.10
CCG--UGCCG-GCGGGAUAU-CGGC  1.9b
CCG-AACUCG-ACGGGAUAA-CGAG  2.1b
CCG--ACUCG--CGGGAUAA-CGAG  2.12
[[....((((....]].....))))  (publ.)
_....((((...._._.._..))))  (FOLDALIGN)
```

The third data set, THEO, consists of two structural classes which are circular permutations of each other. FOLDALIGN identifies the proper motif from the largest class, getting the alignment exactly right for six of the eight sequences. The two remaining sequences contain the shortest stems, only two basepairs in one case, and require the most gaps for proper alignment. The second class could not be aligned with the

first due to the circular permutation, but their common structure should be identifiable if they are treated as a separate class (not tested).

The final set, R17, consists of 36 sequences. For many of these part of the structural motif is contained in the constant SELEX region of the sequence, and so not available to the program for alignment. However, we obtained a perfect alignment for the subset of nine sequences that have at least six basepairs in the stem. We also found a perfect alignment for 12 of 16 sequences with at least five basepair stems. Alignment of larger subsets are nearly correct, although they do contain a few misaligned sequences.

Conclusion

We have presented a method to structurally align a set of RNA sequences, as well as selecting the subsets containing the most significant alignments. The method was able to fully find the published alignments of conserved motifs. The complete structure was not always obtained, as in the case of the pseudoknot, due to the dynamic programming limitation. But the core alignment that is obtained can be used by existing methods to complete the motif identification. For this type of problem the method clearly outperforms the other commonly used methods applied, and we conclude that our method, to a very large extent, can replace the alignments currently made by hand, or provide significant hints to assist with "hands-on" methods.

Acknowledgements

Thanks to B. Javornik and NexStar for providing the data electronically. We thank S. Eddy for advice on the use of COVE. This work was sponsored in part by NIH grant HG00249 to GDS. JG was supported by the Danish National Research Foundation.

References

Cary, R. B., and Stormo, G. D. 1995. Graph–theoretic approach to RNA modeling using comparative data. In *Proceedings of the third International Conference on Intelligent Systems in Molecular Biology*, 75–80. AAAI/MIT Press.

Eddy, S., and Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Research* 22:2079–2088.
(http://genome.wustl.edu/eddy/#cove).

Gorodkin, J.; Heyer, L. J.; and Stormo, G. D. 1997. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *submitted*.

Hertz, G. Z.; Hartzell,III, G. W.; and Stormo, G. D. 1990. Identification of consensus patterns in unaligned dna sequences known to be functionally related. *CABIOS* 6(2):81–92.

Hofacker, I. L.; Fontana, W.; Stadler, P. F.; Bonhoeffer, L. S.; Tacker, M.; and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie* 125:167–188. (http://www.tbi.univie.ac.at/~ivo/RNA/).

Jenison, R. D.; Gill, S. C.; Pardi, A.; and Polisky, B. 1994. High-resolution molecular discrimination by RNA. *Science* 263:1425–1429.

Nussinov, R., and Jacobson, A. B. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Procedings of the National Academy of Science USA* 77:6309–6313.

Sakakibara, Y.; Brown, M.; Underwood, R. C.; Mian, I. S.; and Haussler, D. 1994. Stochastic context-free grammars for modeling RNA. In Hunter, L., ed., *Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences: Biotechnology Computing, vol. V*, 284–293. IEEE Computer Society Press.

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

Schneider, D.; Tuerk, C.; and Gold, L. 1992. Ligands to the bacteriophage R17 coat protein. *Journal of Molecular Biology* 228:862–869.

Smith, T. F., and Waterman, M. S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147:195–197.

Thompson, J. D.; Higgins, D. G.; and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673–4680.

Tuerk, C., and Gold, L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505–510.

Tuerk, C.; MacDougal, S.; Hertz, G. Z.; and Gold, L. 1992. *In vitro* evolution of functional nucleic acids: High-affinity RNA ligands of the HIV-1 *rev* protein. In Ferré, F.; Mullis, K.; Gibbs, R.; and Ross, A., eds., *The Polymerase Chain Reaction*. Birkhauser, Springer–Verlag NY.

Tuerk, C.; MacDougal, S.; and Gold, L. 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Procedings of the National Academy of Science USA* 89:6988–6992.