Predicting RNA Structure Using Mutual Information

ARTICLE *in* APPLIED BIOINFORMATICS · FEBRUARY 2005

DOI: 10.2165/00822942-200504010-00006 · Source: PubMed

3 AUTHORS, INCLUDING:

Paul P. Gardner
University of Canterbury
52 PUBLICATIONS **2,429** CITATIONS

Approval for publication Signed _____ Date _____ Number of amended pages returned _____

ORIGINAL RESEARCH ARTICLE

Appl Bioinformatics 2005; 4 (1): 1
1175-5636/05/0001-0001/$34.95/0

Predicting RNA Structure Using Mutual Information

Eva Freyhult,[1] *Vincent Moulton*[1] and *Paul Gardner*[2]

1 The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden
2 Department of Evolutionary Biology, University of Copenhagen, Copenhagen, Denmark

Abstract

Background: With the ever-increasing number of sequenced RNAs and the establishment of new RNA databases, such as the Comparative RNA Web Site and Rfam, there is a growing need for accurately and automatically predicting RNA structures from multiple alignments. Since RNA secondary structure is often conserved in evolution, the well known, but underused, mutual information measure for identifying covarying sites in an alignment can be useful for identifying structural elements. This article presents MIfold, a MATLAB® toolbox that employs mutual information, or a related covariation measure, to display and predict conserved RNA secondary structure (including pseudoknots) from an alignment.

Results: We show that MIfold can be used to predict simple pseudoknots, and that the performance can be adjusted to make it either more sensitive or more selective. We also demonstrate that the overall performance of MIfold improves with the number of aligned sequences for certain types of RNA sequences. In addition, we show that, for these sequences, MIfold is more sensitive but less selective than the related RNAalifold structure prediction program and is comparable with the COVE structure prediction package.

Conclusion: MIfold provides a useful supplementary tool to programs such as RNA Structure Logo, RNAalifold and COVE and should be useful for automatically generating structural predictions for databases such as Rfam.

Availability: MIfold is freely available from http://www.lcb.uu.se/~evaf/MIfold/

As the secondary structure of functional RNA is integral in determining function,[1,2] automatic RNA structure prediction is an important problem. Until recently, RNA secondary structure was usually predicted from a single sequence using free-energy minimisation (MFE) methods. Although such prediction methods are relatively fast, prediction is often not very accurate.[3] Moreover, MFE structure prediction does not take advantage of the growing number of sequenced RNAs in databases such as the European ribosomal RNA database,[4] the Comparative RNA Web Site[5] and Rfam.[6]

Since RNA structure is often conserved in RNA evolution, structure prediction directly from multiple sequence alignment has proven to be a powerful tool.[7-9] Various methods have been proposed for automating this process. These methods include RNAalifold,[10] which uses averaged free-energy values plus bonuses for sites with structure-neutral mutations, ConStruct,[11] which computes consensus structure based on base-pair probability matrices, and COVE,[12] which uses a covariance model to create a multiple alignment of RNA sequences but can also be used to predict a secondary structure from an existing alignment.

A well known, but underused, method for identifying conserved secondary structures from a multiple alignment of RNA sequences is the mutual information measure.[13] This measure can be used to detect sites that are covarying.[14-16] Programs such as RNA Structure Logo[17] allow the user to display mutual information content for an alignment of RNA sequences, although they do not use this information to actually predict structure (RNA Structure Logo requires the structure *a priori*). In addition, COVE is a program that can predict a consensus secondary structure from a multiple alignment using a covariation model.[12]

In this article, we present a method for RNA structure prediction from a multiple alignment using mutual information and a related covariation measure to infer secondary structures. This method also allows the prediction of simple pseudoknots (non-nested secondary structures). It is implemented in the MATLAB®

toolbox MIfold, which is freely available from http://www.lcb.uu.se/~evaf/MIfold/.

Methods

As is well known, mutual information can be employed to detect compensatory mutations in an alignment.[15,16] To compute the mutual information $H(i,j)$ of columns i and j of an alignment, the following statistics are required: the frequency $f_k(X)$ of base X in column k, for $k = i,j$, where X is A, C, G, U or - (note that gaps and non-canonical bases are treated as a fifth base); and the joint frequency $f_{i,j}(X \bullet Y)$ of complementary bases $X \bullet Y$, where $X \bullet Y$ is $G \bullet C$, $C \bullet G$, $A \bullet U$, $U \bullet A$, $G \bullet U$ or $U \bullet G$. The mutual information is then given by (equation 1):

$$H(i,j) = \sum_{X\ Y} f_{i,j}(X\ \ Y) \log_2 \frac{f_{i,j}(X\ \ Y)}{f_i(X) f_j(Y)}$$

(Eq. 1)

Note that in the classical definition of mutual information, the above sum is taken over all $X \bullet Y$, with X, Y any base. However, we found that this measure has a high signal-to-noise ratio (data not shown).

In Hofacker et al.,[10] a covariation score was employed instead of the mutual information measure, which for columns i and j of an alignment is defined as (equation 2):

$$C(i,j) = \sum_{X \bullet Y, X' \bullet Y'} f_{i,j}(X \bullet Y) D(X \bullet Y, X' \bullet Y') f_{i,j}(X' \bullet Y')$$

(Eq. 2)

where $D(X \bullet Y, X' \bullet Y')$ is the Hamming distance between $X \bullet Y$ and $X' \bullet Y'$, and the sum is taken over all complementary base pairs as for $H(i,j)$ in equation 1. This method has the advantage of giving wobble base pairs ($G \bullet U$) different weights to canonical ones. In addition, Hofacker et al.[10] introduced an inconsistent sequences penalty (equation 3):

$$q(i,j) = 1 - f_{i,j}^{comp} - f_{i,j}(- \bullet -)$$

(Eq. 3)

where $f_{i,j}^{comp}$ is the frequency of complementary base pairs in columns i and j. This penalty, $q(i,j)$, is subtracted from either $H(i,j)$ or $C(i,j)$ to increase the signal-to-noise ratio. Users of MIfold can predict secondary structures from RNA alignments using mutual information or covariation measures with or without the inconsistent sequences penalty. The inconsistent sequences penalty has been observed to improve the MIfold structure predictions (data not shown) for a small number of sequences, but for larger numbers of sequences the penalty does not significantly affect predictivity. By default, the inconsistent sequences penalty is used.

Once the mutual information (or covariation) is computed, a secondary structure is inferred by MIfold using a dynamic programming algorithm, similar to the Nussinov algorithm,[18] that maximises the sum of mutual information taken over all base pairs. The following recursion is used to compute the maximal sum of mutual information on the interval i to j (equation 4):

$$D(i,j) = 0 \qquad \text{for } |i-j| \le 3$$

$$D(i,j) = \begin{cases} D(i+1,j) \\ D(i,j-1) \\ M(i,j) + D(i+1,j-1) \\ \max\limits_{i+4 \le k \le j-5} \left[D(i,k) + D(k+1,j) \right] \end{cases} \qquad \text{for } |i-j| > 3$$

(Eq. 4)

where $M(i,j)$ is one of the measures $H(i,j)$ or $C(i,j)$ [with or without the penalty $q(i,j)$] or zero if the value is below a user-defined threshold $0 < \tau < 2$. The threshold τ is used to further reduce the signal-to-noise ratio. An optimal secondary structure is computed with a usual traceback procedure.[14]

Once an optimal structure is computed, MIfold predicts simple pseudoknots by setting all $M(i,j)$ values to zero with i or j included in some base pair of the secondary structure, and running the dynamic programming algorithm again (this approach is mentioned in Ruan et al.[19]). This procedure predicts simple pseudoknots (of the first order). The method of maximum weight matching has previously been used to predict pseudoknots from a score matrix. This method, however, was shown to have a low accuracy,[19] whereas the method of iterated loop matching (ILM)[19] was shown to more accurately predict pseudoknots of any order. However, since most known pseudoknots are not higher order, we decided to search for only first-order pseudoknots to reduce the number of falsely predicted base pairs.

Predicted secondary structures are displayed in a mountain plot[20] (see figure 1a). A mountain plot is a graph whose x-coordinate indicates nucleotide positions and y-coordinate indicates the number of base pairs enclosing each base. The MIfold mountain plot is not incremented by base-pair count, but by mutual information or covariation (similar to the mountain plots produced by RNAalifold[10]). By weighting the mountains with mutual information, secondary structures that are well supported by the mutual information will be more prominent.

A histogram of the sequence information content I[21] is displayed below the mountain plot (see figure 1a). For column i, the information content of base X is defined as (equation 5):

$$I_i(X) = f_i(X) \log_2 \frac{f_i(X)}{g(X)}$$

(Eq. 5)

where $g(X)$ is a background frequency for base X (the background frequency can be set by the user in MIfold, or computed as the total frequency of the base in the alignment). In the histogram in

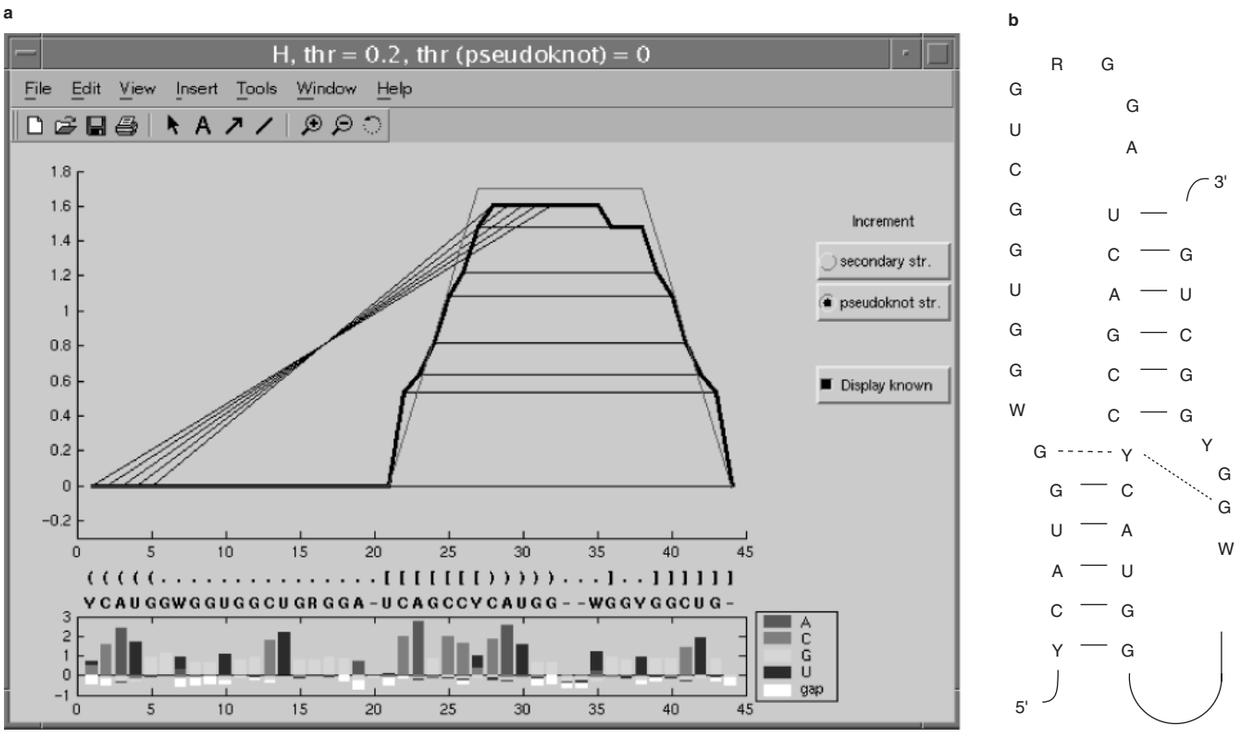

Fig. 1. (**a**) Screenshot of a MIfold prediction of the prion protein mRNA pseudoknot. The predicted 3′ helix is represented by a black mountain plot and the known 3′ helix by a grey mountain plot. Base pairs in the 3′ helix are represented by horizontal black lines, whereas base pairs in the 5′ helix are represented by sloping black lines. The spacing between the black horizontal lines is proportional to the mutual information of the base pairs. The sequence information content is shown below the mountain plot in a histogram. The most informative sequence and the predicted secondary structure in dot-bracket notation are shown between the mountain plot and the histogram. (**b**) The pseudoknot structure in squiggle format. The dashed line indicates a base pair in the reference structure, the dotted line indicates a MIfold predicted base pair that conflicts with the reference structure. Solid lines indicate base pairs that are present in both the consensus and predicted structures.

figure 1a, negative information contents corresponding to under-represented nucleotides (nts) appear below the x-axis and positive information contents (over-represented nucleotides) are shown above the x-axis.

The information content is used to compress the alignment into a single sequence, the most informative sequence, by representing column i in the alignment by the character in International Union of Pure and Applied Chemistry (IUPAC) code[22] corresponding to the set of nucleotides with $I_i(X) > t$, where t is a user-defined threshold. The most informative sequences are used in the structure shown in figure 1b.

Both the most informative sequence and the predicted secondary structure, in dot-bracket notation, can be output in the MATLAB® command window for further analysis with, for example, the Vienna package.[23]

Results

To demonstrate some features of MIfold, we first present an analysis of a prion protein mRNA pseudoknot alignment. Then, to test its performance, we apply MIfold to some sequence alignments obtained from the Rfam database.[6] Finally, we investigate the performance of MIfold and the related programs RNAalifold[10] and COVE[12] for alignments of increasing numbers of sequences.

The prion protein mRNA pseudoknot is a small pseudoknot with two knotted stems of six base pairs each (see figure 1b).[24] We used the alignment presented in Barrette et al.,[24] with identical sequences removed. The alignment has length of 44 nts and contains 42 sequences with an average pairwise sequence identity of 68%. MIfold correctly predicts 11 of the 12 consensus base pairs. Figure 1a shows a screenshot of the MIfold prediction for the pseudoknot. The height of the mountain representing the pseudoknot indicates high mutual information support for the predicted base pairings. However, note that the base pair between

nucleotides 27 and 36 (represented by the topmost horizontal line in the mountain plot) is falsely predicted, although it has a relatively low mutual information value. We compared this MIfold prediction with the pseudoknot prediction by the ILM program.[19] Following the ILM authors' recommendations, we used the default values for an alignment of ten sequences or more. ILM predicts three stems, or 18 base pairs in total, eight of which are correctly predicted. Thus, MIfold is both more sensitive and more selective than ILM on this dataset, at least if the default parameters are used. Although the average sequence identity is 68%, many sequences have a very high sequence identity, with a few sequences that are very different. Therefore, we also tried to use the parameters recommended for an alignment of less than ten sequences. This prediction was better, as all 12 true base pairs were predicted, but also a third stem with six base pairs was incorrectly predicted.

To investigate the utility of MIfold we analysed some datasets obtained from the Rfam database.[6] We selected a representative dataset of alignments with different lengths, numbers of sequences and divergences. We do not present datasets with high sequence identity, since identical sequences will give mutual information values of zero. None of the selected datasets from Rfam contain RNAs known to have pseudoknot structures, so this analysis illustrates the utility of MIfold without pseudoknots prediction. The selected datasets are: the full Rfam alignment of type III hammerhead ribozyme (Hh3), length 84 nts, 209 sequences, 74% identity; the full Rfam alignment of iron response element (IRE), length 30 nts, 175 sequences, 66% identity; the full Rfam alignment of U2 spliceosomal RNA (U2), length 357, 418 sequences, 50% identity; the full Rfam alignment of U4 spliceosomal RNA (U4), length 312 nts, 268 sequences, 53% identity; and the Rfam seed alignment of transfer RNA (tRNA), length 119 nts, 1163 sequences, 32% sequence identity. As measures of performance, we compute the percentage of true positives (*TP*, i.e. the percentage of base pairs in the consensus Rfam structure that MIfold predicts) and the percentage of false positives (*FP*, i.e. the percentage of base pairs in the structure predicted by MIfold that are not in the consensus Rfam structure). High *TP* values indicate sensitivity, and low *FP* values selectivity. A method that is both sensitive and selective would, of course, be optimal. Note that a strict definition of 'consensus structure' is the intersection of the base-pair sets for all sequences in the alignment. For practical purposes, the consensus structure is generally considered to be the set of base pairs preserved by the majority of known homologues. For testing, we use alignments and consensus structures provided by the Rfam database.

Structures were predicted using MIfold with both the mutual information and the covariation measure with varying threshold values, τ, that can be set in order to increase the signal-to-noise ratio. Results are displayed in table I. The tRNA alignment, which has the lowest pairwise sequence identity, shows the most accurate structure prediction. However, for the other four alignments, there is no obvious correlation between sequence identity and *TP* and *FP* values. For all alignments, the *TP* and *FP* values decrease with increasing τ (both for the mutual information and for the covariation measure).

To illustrate in more detail the effect of varying τ, we present a further analysis of the Rfam tRNA seed alignment (from table I) in figure 2. This alignment has 1163 sequences with an average percentage identity of 43%. Results in figure 2 confirm those in table I: that high/low threshold values result in low/high *TP* and *FP* values, respectively. Thus, τ will determine the type of prediction obtained. If a sensitive prediction is required, then τ should be decreased, whereas τ should be increased for more selective predictions. In practice, τ can be selected by inspection of the mutual information matrix (for more detail see MIfold tutorial at http://www.lcb.uu.se/~evaf/MIfold/).

We also investigated RNAalifold[10] and COVE[12] structure predictions for the five datasets in table I. For the IRE alignment,

Table I. MIfold selectivity and sensitivity on Rfam full alignments of type III hammerhead ribozyme (Hh3), iron response element (IRE), U2 spliceosomal RNA (U2) and U4 spliceosomal RNA (U4), and the seed alignment of transfer RNA (tRNA). The table contains the number of sequences in the alignment (# seqs), length of the alignment (n), percentage pairwise identity (% ID), threshold value (τ) and percentage true and false positives (*TP* and *FP*, with the Rfam consensus structure used as a reference) for the mutual information (*H*) and covariation (*C*) measures (both with the inconsistent sequences penalty, *q*)

RNA	# seqs	n	% ID	τ	H		C	
					TP	FP	TP	FP
Hh3	209	84	77	0.0	87	48	87	43
				0.2	67	55	53	33
				0.4	33	38	33	17
IRE	175	30	68	0.0	80	20	80	20
				0.2	80	11	80	0
				0.4	60	0	70	0
U2	418	357	60	0.0	67	40	51	47
				0.2	58	24	36	24
				0.4	42	5	29	0
U4	268	312	60	0.0	72	54	69	41
				0.2	34	27	44	26
				0.4	25	0	16	17
tRNA	1163	119	43	0.0	100	30	100	28
				0.2	95	20	95	9
				0.4	90	17	90	0

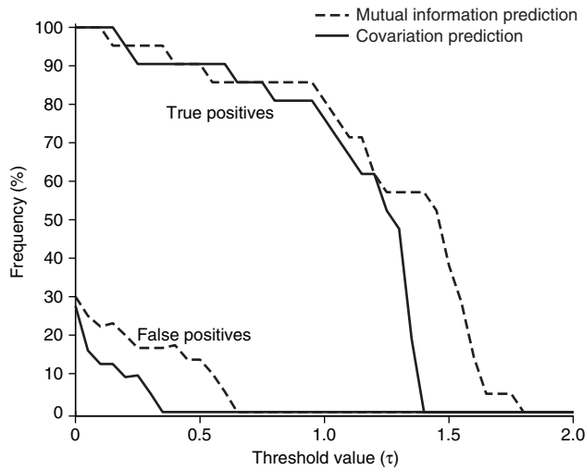

Fig. 2. The percentage of true and false positives for MIfold structure predictions with increasing threshold values for the Rfam transfer RNA (tRNA) seed alignment in table I.

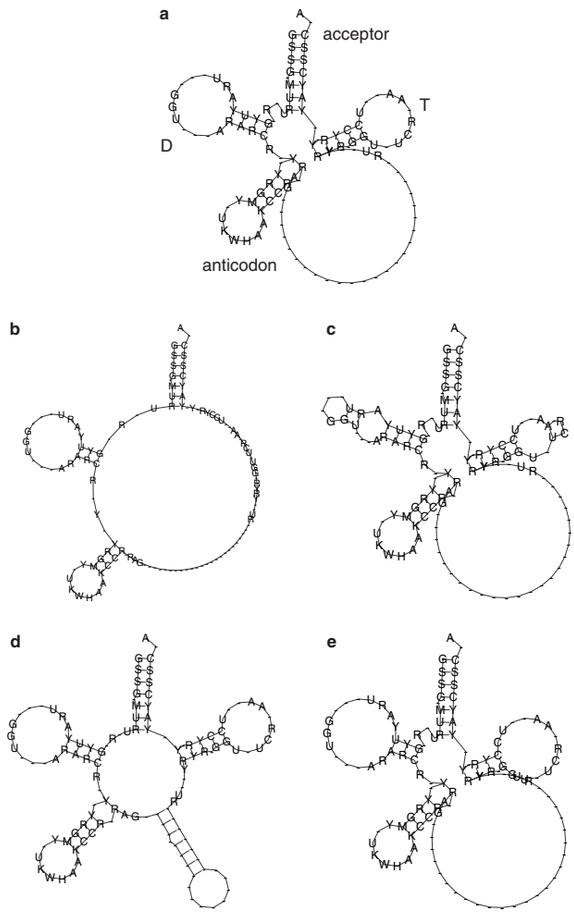

Fig. 3. Rfam transfer RNA (tRNA) seed alignment structures in squiggle format. Displayed are the Rfam consensus structure (**a**), RNAalifold prediction (**b**), COVE prediction (**c**) and MIfold predictions (mutual information measure (**d**) and covariation measure (**e**) with thresholds values 0.05 and 0.35, respectively). The plots were generated using RNAplot, part of the Vienna RNA package.[23] The most informative sequence with threshold value zero forms the backbone. The anticodon, acceptor, T- and D-stems are marked in the Rfam structure (**a**).

the RNAalifold, COVE and MIfold predictions are similar. For the Hh3 alignment, the RNAalifold prediction is the best (*TP* = 100, *FP* = 12), whereas the COVE prediction has fewer base pairs, both true and false ones, than both the RNAalifold and the MIfold predictions. The fact that RNAalifold outperforms both COVE and MIfold for these alignments is to be expected: the alignments have a relatively high percentage sequence identity, and both COVE and MIfold rely on covariations in the alignment, whereas RNAalifold also takes into account minimal free energy. RNAalifold did not predict a structure for the more divergent sequence alignments of U2 and U4, whereas both MIfold and COVE gave predictions with approximately equal *TP* values, and the *FP* values for COVE were lower than for MIfold. Differences in the performance of COVE and MIfold can be partially explained by the treatment of gaps. COVE filters gap-rich sites whereas MIfold includes these sites, and hence may predict structure in gap-rich regions (see figure 3). In addition, COVE does not restrict loop size, which biological evidence suggests should be no less than three nucleotides. COVE performance was generally comparable with MIfold in terms of sensitivity, but it was slightly more selective in base-pair assignments. However, MIfold was found to outperform COVE on low-diversity datasets and when the number of sequences was small (see figure 4).

Since RNAalifold, COVE and MIfold gave reasonable predictions for the tRNA seed alignment, we analysed this dataset in more detail. We randomly selected subcollections from the seed alignment containing different numbers of sequences and predicted their structure using RNAalifold, COVE and MIfold. For each number of sequences we repeated this selection 50 times and computed the average *TP* and *FP* values. In figure 4, we see that *TP* increases with number of sequences for both the MIfold prediction methods. For mutual information, *FP* values are both constant and relatively high, which is due to the fact that some sequences conserve a fifth stem that does not appear in the Rfam consensus structure (see figure 3). However, for covariation, *FP* values decrease with number of sequences. For small numbers of sequences (less than ten), the COVE prediction cannot be trusted as the *FP* value is very high and the *TP* value low.

TP and *FP* values for COVE predictions increase with the number of sequences; however, the high *FP* cannot be explained by the prediction of a fifth stem (see figure 3). For a large number of sequences (ten or more), the *TP* and *FP* values for RNAalifold

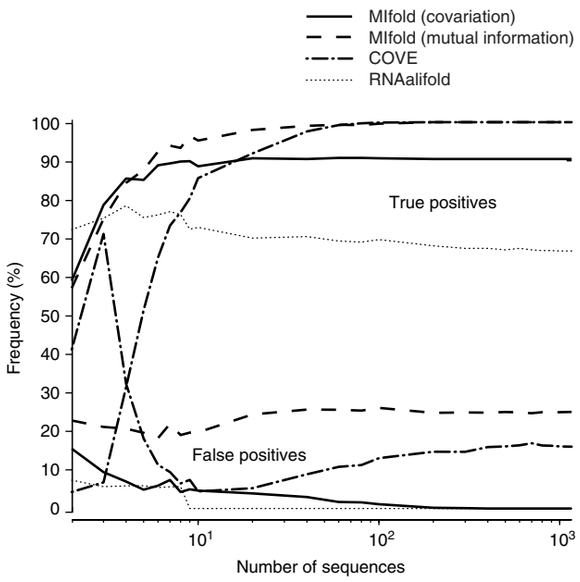

Fig. 4. The average percentage of true and false positives for RNAalifold, COVE and MIfold structure predictions for an increasing number of sequences in the Rfam transfer RNA (tRNA) alignment. Sequences were randomly selected and structures predicted; this was repeated 50 times. The following threshold values were used for the MIfold structure predictions: mutual information, 0.05; covariation, 0.35. The high false-positive value for MIfold prediction when using mutual information measures is due to the prediction of the tRNA variable loop, a loop that is not in the Rfam consensus structure.

are fairly constant (70% and 0%, respectively). Thus, for small numbers of sequences (less than four), we see that RNAalifold tends to make better predictions than both MIfold and COVE, but for larger numbers of sequences MIfold is optimal.

In figure 3, structure predictions from RNAalifold, COVE and MIfold for the full Rfam tRNA seed alignment are shown. Structures in figure 3 indicate that RNAalifold does not predict the T-stem. The structures predicted by MIfold using the covariation measure are both selective and sensitive, predicting all four stems in the consensus tRNA structure. The mutual information structures, however, are the most sensitive: they contain the possible fifth stem in the variable loop region (for example, *Escherichia coli* leucine-tRNA and selenocysteine-tRNA),[25] which is not present in the Rfam consensus structure.

Discussion and Conclusion

We present a tool, MIfold, that uses mutual information and a related covariation measure to infer secondary structure, including simple pseudoknots, from RNA alignments. The algorithm is efficient, and it has the same time and memory complexity as the Nussinov algorithm,[18] which for a sequence of length n are $O(n^3)$ and $O(n^2)$, respectively.

MIfold can also be a useful alternative to the RNA Structure Logo software.[17] MIfold does not require a structure in advance, and it both displays and predicts a secondary structure. Moreover, if the secondary structure is known in advance it can still be input to MIfold, which will display base pairs that are supported by compensatory mutations.

Using various datasets, we show that MIfold can be used to predict simple pseudoknots, and that the performance can be adjusted to make it either more sensitive or more selective. In addition, we demonstrate that the overall performance of MIfold improves with the number of aligned sequences for certain RNAs, which in view of the ever-increasing number of sequenced RNAs indicates that it should be a useful tool for automatic RNA structure prediction. For these sequences, we also show that MIfold is more sensitive than the related RNAalifold structure prediction program, although RNAalifold is somewhat more selective, and that the performance of MIfold and COVE is comparable. A comparison of MIfold and ILM for pseudoknot prediction showed that the MIfold prediction was both more sensitive and more selective than the ILM prediction using the ILM defaults.

A limitation of MIfold, shared by many comparative structure analysis programs, is that it depends upon the quality of the input alignment. In this article, we have used Rfam alignments for comparing MIfold predictions with the Rfam consensus structures. The Rfam alignments are structural alignments, and if the sequences are realigned using ClustalW[26] some structural information is lost, resulting in a lower number of true positives and a larger number of false positives (data not shown). However, it is not obvious how to evaluate the predictions for such realignments. Moreover, in MIfold, only covarying sites will provide signal, i.e. base pairs that are conserved on the sequence level will not be predicted. This is why MIfold tends to not predict structure well from alignments of sequences with low divergency; too many base pairs are conserved.

In conclusion, MIfold provides a useful supplementary tool to programs such as RNA Structure Logo, RNAalifold and COVE and should be useful for automatically generating structural predictions for databases such as Rfam.

Acknowledgements

EF and VM thank the Swedish Research Council for its support. PG thanks The Linnaeus Centre for Bioinformatics for hosting him, and the European Community for partially supporting him (under the programme 'Improving the Human Research Potential and the Socio-Economic Knowledge Base', grant number HPRI-CT-2001-00153). All authors thank Robert Giegerich for his support.

The authors have provided no information on conflicts of interest directly relevant to the content of this article.

References

1. Kim SH, Suddath GJ, Quigley GJ, et al. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. Science 1974; 185: 435-9
2. Shi H, Moore PB. The crystal structure of yeast phenylalanine tRNA at 1.93 a resolution: a classic structure revisited. RNA 2000; 6: 1091-105
3. Mathews D, Sabina J, Zuker M, et al. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J Mol Biol 1999; 288: 911-40
4. Wuyts J, Perriere G, Van De Peer Y. The European ribosomal RNA database. Nucleic Acids Res 2004; 32: D101-3
5. Cannone JJ, Subramanian S, Schnare MN, et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 2002; 3: 2
6. Griffiths-Jones S, Bateman A, Marshall M, et al. Rfam: an RNA family database. Nucleic Acids Res 2003; 31: 439-41
7. Rosenblad MA, Gorodkin J, Knudsen B, et al. SRPDB: Signal Recognition Particle Database. Nucleic Acids Res 2003; 31: 363-4
8. Espinosa de los Monteros A. Models of the primary and secondary structure for the 12S rRNA of birds: a guideline for sequence alignment. DNA Seq 2003; 14: 241-56
9. Vitreschak AG, Rodionov DA, Mironov AA, et al. Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. RNA 2003; 9: 1084-97
10. Hofacker I, Fekete M, Stadler P. Secondary structure prediction for aligned RNA sequences. J Mol Biol 2002; 319: 1059-66
11. Lück R, Gräf S, Steger G. ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. Nucleic Acids Res 1999; 27: 4208-17
12. Eddy S, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res 1994; 22: 2079-88
13. Wong AKC, Chiu DKY. An event-covering method for effective probabilistic inference. Pattern Recognit 1987; 20: 245-55
14. Durbin R, Eddy S, Krogh A, et al. Biological sequence analysis: probabilistic models of protein and nucleic acids. Chap 10. Cambridge: Cambridge University Press, 1998: 260-98
15. Chiu DK, Kolodziejczak T. Inferring consensus structure from nucleic acid sequences. Comput Appl Biosci 1991; 7: 347-52
16. Gutell RR, Power A, Hertz GZ, et al. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res 1992; 20: 5785-95
17. Gorodkin J, Heyer L, Brunak S, et al. Displaying the information contents of structural RNA alignments. Bioinformatics 1997; 13: 583-6
18. Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc Natl Acad Sci U S A 1980; 77: 6903-13
19. Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics 2004; 20: 58-66
20. Hogeweg P, Hesper B. Energy directed folding of RNA sequences. Nucleic Acids Res 1984; 12: 67-74
21. Schneider T, Stephens R. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990; 18: 6097-100
22. Cornish-Bowden A. IUPAC-IUB symbols for nucleotide nomenclature. Nucleic Acids Res 1985; 13: 3021-30
23. Hofacker IL, Fontana W, Bonhoeffer S, et al. Fast folding and comparison of RNA secondary structures. Monatsh Chem 1994; 125: 167-88
24. Barrette I, Poisson G, Gendron P, et al. Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. Nucleic Acids Res 2001; 29: 753-8
25. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997; 25: 955-64
26. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22: 4673-80

Correspondence and offprints: *Eva Freyhult*, The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, 751 24 Uppsala, Sweden.
E-mail: eva.freyhult@lcb.uu.se