BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 24 2006, pages 2988–2995
doi:10.1093/bioinformatics/btl514

Sequence analysis

Measuring covariation in RNA alignments: physical realism improves information measures

S. Lindgreen*, P. P. Gardner[1] and A. Krogh

Bioinformatics Centre and [1]Molecular Evolution Group, Institute of Molecular Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen Ø, Denmark

Received on June 28, 2006; revised and accepted on October 4, 2006
Advance Access publication October 12, 2006
Associate Editor: Golan Yona

ABSTRACT

Motivation: The importance of non-coding RNAs is becoming increasingly evident, and often the function of these molecules depends on the structure. It is common to use alignments of related RNA sequences to deduce the consensus secondary structure by detecting patterns of co-evolution. A central part of such an analysis is to measure covariation between two positions in an alignment. Here, we rank various measures ranging from simple mutual information to more advanced covariation measures.

Results: Mutual information is still used for secondary structure prediction, but the results of this study indicate which measures are useful. Incorporating more structural information by considering e.g. indels and stacking improves accuracy, suggesting that physically realistic measures yield improved predictions. This can be used to improve both current and future programs for secondary structure prediction. The best measure tested is the RNAalifold covariation measure modified to include stacking.

Availability: Scripts, data and supplementary material can be found at http://www.binf.ku.dk/Stinus_covariation

Contact: stinus@binf.ku.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In recent years, it has become increasingly evident that many RNAs play a more active role in the cell than just being the mediator of information between the DNA and protein levels. These non-coding RNAs (ncRNAs) have a functional role, and their function is frequently tied to the structure of the molecule. Methods for predicting the secondary structure of functional RNAs are therefore becoming increasingly important. A benchmarking study of various structure prediction programs can be found in the somewhat dated study by Gardner and Giegerich (2004).

Secondary structure prediction has been pursued in various ways: folding a single sequence using energy minimization is implemented in e.g. RNAfold (Hofacker *et al*., 1994), Mfold (Zuker and Stiegler, 1981; Zuker, 2003) and RNAstructure (Mathews *et al*., 2004), while a method for simultaneously performing sequence alignment and structure prediction was described by Sankoff (1985). Although there is no published implementation of the full Sankoff algorithm, different simplifications exist e.g. FOLDALIGN (Gorodkin *et al*., 1997; Havgaard *et al*., 2005), Dynalign (Mathews and Turner, 2002) and PMmulti (Hofacker *et al*., 2004).

The most widely used methods for the analysis of ncRNA are based on comparative analysis of related sequences. All such methods are based on some measure of covariation making an in-depth study of different measures important. A number of these programs use the standard mutual information (MI) or variants hereof [e.g. KNetFold (Bindewald and Shapiro, 2006), COVE (Eddy and Durbin, 1994), ILM (Ruan *et al*., 2004), MatrixPlot (Gorodkin *et al*., 1999), Construct (Lück *et al*., 1999)] while others use more advanced measures [e.g. RNAalifold (Hofacker *et al*., 2002), RNAz (Washietl *et al*., 2005), MSARI (Coventry *et al*., 2004)] or are based on stochastic context-free grammars [e.g. Pfold (Knudsen and Hein, 2003), QRNA (Rivas and Eddy, 2001), EvoFold (Pedersen *et al*., 2006)]. In this study, we analyze a number of measures ranging from simple MI to more advanced measures. Although it has been suggested that measures including phylogenetic information are the most powerful (Akmaev *et al*., 2000), these methods have not been included in this study.

In the comparative approach, the goal is to predict the common structure for a set of aligned RNA sequences by using the evolutionary information in the alignment. The secondary structure is constituted by base pairing interactions between nucleotides. These are mainly the Watson–Crick base pairs (C • G and A • U) and the wobble base pair (G • U), although other base pairing interactions have been shown to be more important than previously anticipated (Leontis *et al*., 2002; Lee and Gutell, 2004). During evolution, RNA sequences can mutate while retaining the same structure. If a base pair is disrupted by a mutation at one position, evolution favours mutations that correct this. It is these compensatory mutations that comparative methods use when optimizing the structure of a sequence alignment.

The comparative approach relies on two conflicting properties: information and alignment quality. Structure can only be inferred for positions that have actually mutated. Fully conserved columns have no covariance information, and thus highly diverged sequences carry most covariance information. On the other hand, highly diverged sequences are difficult to align correctly, and therefore the comparative methods are expected to be best for sequences that are diverged to the point where one can still obtain a good multiple alignment (without using covariance information). A recent survey showed that sequences below ∼65% identity were inaccurately

*To whom correspondence should be addressed.

aligned, thus destroying secondary structure information (Gardner *et al*., 2005). It is also clear that it may be advantageous to use covariance information for the alignment, and thus align and predict structure at the same time, but current methods for that are limited to small alignments.

As more focus is being given to ncRNAs, interest in locating these in genomic sequences is growing. Genefinders for this problem are appearing: Using stochastic context-free grammars is employed in the programs QRNA (Rivas and Eddy, 2001) and EvoFold (Pedersen *et al*., 2006). In these, the grammars are designed to detect signal from base pairing interactions using an implicit covariation measure. MSARI (Coventry *et al*., 2004) uses a combination of base pairing probabilities (McCaskill, 1990) and a sliding window for finding complementary subsequences while allowing for small misalignments. In RNAz (Washietl *et al*., 2005) a sliding window is used to fold subalignments using RNAalifold. The consensus structure is compared to the minimum free energy structures of the individual sequences. The covariation measure used in this approach is therefore the same as in RNAalifold.

To measure covariation between two sites in the RNA molecule, where the sequence is changed while base pairing interactions are preserved, mutual information is the textbook example, e.g. (Durbin *et al*., 1998) because it measures the information in base pairs which cannot be explained from single base frequencies. The idea is to find sites where the degree of co-occurring mutations is higher than one would expect by chance. Sites showing a high degree of covariation are seen as likely base pairs. Several methods include an MI component in their scoring scheme, e.g. COVE (Eddy and Durbin, 1994), ILM (Ruan *et al*., 2004), MatrixPlot (Gorodkin *et al*., 1999), KNetFold (Bindewald and Shapiro, 2006) and Construct (Lück *et al*., 1999).

Although MI and other measures of covariation are widely used there has, to the best of our knowledge, never been a thorough analysis of the discriminative power of these measures. In this study, we analyze a number of RNA datasets using different measures of covariation. The datasets cover both different classes of structural RNAs and different degrees of overall sequence identity. We show that the standard MI is not very discriminative, and that extending the measure with additional structural information yields a more powerful measure. The best covariation measure tested in this study is a new formulation of the measure used in RNAalifold (Hofacker *et al*., 2002) where stacking of base pairs is taken into account. The most discriminative measure, though, is averaged base pairing probability matrices calculated using the partition function (McCaskill, 1990), which uses energy terms and is independent of covariation information.

2 APPROACH

We have evaluated the discriminative power of different covariation measures. The following measures described previously in the literature were implemented:

- Standard MI (Shannon, 1948; Chiu and Kolodziejczak, 1991; Gutell *et al*., 1992)
- MI summing only Watson–Crick and wobble base pairs (Gorodkin *et al*., 1999) (MI^W)
- Normalized MI (Martin *et al*., 2005) (MI^N)
- The covariation measure used in RNAalifold (Hofacker *et al*., 2002) (B)

A number of novel measures based on the preceding list were implemented and evaluated as well:

- MI using gap penalties (MI^P)
- MI summing base pairs and including stacking ($MI^{S\diamond W}$)
- MI summing base pairs, including stacking and using gap penalties ($MI^{S\diamond W\diamond P}$)
- MI summing base pairs and using gap penalties ($MI^{W\diamond P}$)
- The B measure including stacking (B^s)

Note that for simplicity some of the measures are denoted MI although they are technically not mutual information measures. This will be elaborated in the following. The performance of the above measures is compared to using averaged base pairing probability matrices calculated using the partition function (McCaskill, 1990). We used the implementation from the Vienna package (Hofacker *et al*., 1994). The different measures are described in detail under methods.

The datasets were compiled by making three random samplings from each of three large structural alignments (tRNA, 5S rRNA and U5) (Griffiths-Jones *et al*., 2003; Szymanski *et al*., 2002; Zwieb, 1997) yielding a total of nine datasets. The alignments used in this study are known to be of high quality, and they have previously been used in the benchmark by Gardner *et al*. (2005), but other datasets could in principle have been used. The sampling was performed in such a way that the overall identity of the subalignments was controlled. The subalignments are of low (40–60%), medium (60–80%) and high (80–100%) overall identity, where the % ID is calculated using the reference alignments. In each subalignment, any all-gap columns were removed but otherwise the correct alignment and reference structure is preserved. The sizes of the alignments are summarized in Table 1.

Table 1. The sample sizes

RNA family	Low % ID	Medium % ID	High % ID
tRNA	29	16	14
5S rRNA	21	25	30
U5	20	20	20

The number of sequences in each dataset used in the analysis.

The nine individual datasets were analyzed separately, but in the following the results reported are averages for each of the identity intervals (i.e. containing an alignment of 5S rRNA, tRNA and U5). This is done to avoid bias from the composition of the individual families while measuring the performance as a function of sequence similarity.

The measures are compared using the Matthew's correlation coefficient (MCC, see methods), which is maximal (= 1) if, for a given threshold γ, all true base paired columns are above and all other pairs are below γ. Thus, the higher the MCC, the better the discrimination. Each measure was evaluated using 100 threshold values evenly distributed between the minimum and maximum value for that particular measure. Since the range and distribution

of the different measures vary, the actual threshold values are not comparable. Other binning strategies were tested but did not affect the results (data not shown).

3 METHODS

The covariation measures used in this study are described in the following along with the evaluation scheme used to benchmark the measures.

3.1 Evaluating the measures

Each covariation measure is evaluated based on how well it discriminates between true and false base pairs. Given a sequence alignment of structural RNAs, every possible base pair (i, j) can be evaluated using each of the suggested measures. Since the reference structure is available, these scores can be divided into true base pairs BP_T and false base pairs BP_F.

Each measure calculates a score for a possible base pair, but the question is how large a score has to be to best discriminate true base pairing interactions from false. For this, a threshold value has to be used. For a given threshold value, γ, the number of true positives, TP, can be found as the number of scores in BP_T greater than or equal to γ. Similarly, the number of false negatives, FN, is the number of scores in BP_T that are smaller than γ. The number of false positives, FP and true negatives, TN, can be found in a similar manner from the scores in BP_F.

To evaluate the different measures, 100 thresholds evenly distributed between the minimum and maximum score were used and the numbers TP, FP, TN and FN calculated. For example, for the standard mutual information, 100 numbers between 0 and $\log_2(4)$ were tried. For each threshold, one could use the sensitivity and positive predictive value to evaluate the discriminative power of the measure. Instead of selecting thresholds based on either of these, however, a balance between them is sought by using the MCC (Matthews, 1975):

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Using this, MCC as a function of the threshold values was analyzed for each covariation measure (see Supplementary material). The actual values of the thresholds vary between the measures due to the different ranges, but an equal number of possible thresholds was used for each.

The datasets analyzed contained either tRNA, 5S rRNA or U5 sequences. For each of these families, three datasets were generated having an overall identity of 40–60%, 60–80% and 80–100% (in the following these are referred to as low, medium and high identity, respectively). For each identity interval the average over the three datasets was used, thus showing the dependency on the sequence similarity while being independent of the specific family.

3.2 Standard MI

Functional RNA molecules are under selective pressure to preserve their secondary structure. This evolutionary pressure can lead to a change in the primary sequence that keeps the base pairing interactions intact. An example of such a chain of mutations is:

$$\mathrm{A \bullet U \rightarrow G \bullet U \rightarrow G \bullet C}$$

Thus, related RNAs can differ in sequence while having the same structure. This makes pure sequence alignment of structural RNAs difficult. As shown in Gardner *et al*. (2005), most sequence alignment tools fail when the sequence identity is below 50–60%. To use this evolutionary information in the prediction of secondary structure, one needs to find pairs of columns in the alignment that show a higher degree of covariation than expected by chance. This might indicate base pairing interaction between the two positions. The classic measure for this is the MI content (Shannon, 1948; Chiu and Kolodziejczak, 1991; Gutell *et al*., 1992; Durbin *et al*., 1998).

In the context of RNA alignment, this score is defined on pairs of columns, i and j, in the multiple alignment. Let a be a letter from column i and let b be a letter from column j. The frequency $f_{i,j}(ab)$ at which a base pair of type ab is observed is compared to the number of times one would expect the pair to occur by chance. The latter is calculated using the frequencies of the two single nucleotides in the two columns, $f_i(a)$ and $f_j(b)$. For two columns i and j, the mutual information is given by the following expression:

$$MI_{i,j} = \sum_{ab} f_{i,j}(ab) \log_2 \frac{f_{i,j}(ab)}{f_i(a) f_j(b)}, \quad (1)$$

where the sum is over all 16 possible pairs of bases in the two columns. Keep in mind that $0 \cdot \log_2(0) = 0$. This is the relative entropy of the joint distribution relative to the product distribution, also known as the Kullback–Leibler divergence (Kullback and Leibler, 1951). The MI gives the amount of information obtained about one position in the alignment if one knows what the other position is. The higher the MI, the more information is gained.

There are some problems with this measure: if one or both positions are conserved in sequence the MI is zero. Another problem is the level of noise from covarying columns that cannot form base pairs and thus should not contribute to the measure in the context of classical canonical pairing. Furthermore, structurally neutral mutations, such as A • U → G • U do not contribute to the MI. Finally, there is the question of how to deal with gaps in the alignment.

Since gap characters symbolize insertion/deletion events, it makes little sense to deduce covariation from them. It is therefore necessary to subtract the number of gaps when calculating the MI. Note, however, that the number of gaps in the two columns i and j may vary. But if the frequencies are based on two different numbers of observations, the MI score fails. Instead, all the pairs that contain at least one gap character are disregarded when calculating frequencies.

3.3 MI with gap penalty

As mentioned, the standard formulation of MI does not penalize gaps. Instead, positions with gaps are disregarded. This presents a new problem: consider a pair of columns containing many gaps but where the few remaining positions display a high degree of covariation. Since the gaps are disregarded, this column pair receives a good score. This is not necessarily the desired behaviour: it would be interpreted as a high degree of structural conservation, but that does not correspond with the large number of indels.

Instead, a large number of gaps at a given position implies that the region is variable and the MI score is less certain. This means that less weight should be placed on these positions. A simple way to incorporate gap penalties into the MI score is to define a gap penalty β and let it influence the MI score as a function of the number of gap positions. Let $N^G_{i,j}$ be the number of positions containing at least one gap, and let $MI_{i,j}$ be defined as in Equation (1). A variation of the MI score with gap penalties is:

$$MI^P_{i,j} = MI_{i,j} - N^G_{i,j} \cdot \beta. \quad (2)$$

As can be seen, the MI score of a column pair with no gaps will not be penalized, while more gaps give a larger penalty. If all positions contain a gap, the column pair will receive a negative MI score of $N \cdot \beta$. In the experiments we used $\beta = \frac{1}{N}$.

3.4 MI using only canonical base pairs

The standard MI is sensitive to noise from unwanted, non-canonical base pairs. A possible variation of MI that might limit the noise is to focus on the acceptable base pairs alone and ignore the rest (Gorodkin *et al*., 1999). While the 4 nucleotides give rise to 16 possible base pairs, only 6 of these are considered structurally important. If this distinction is incorporated into the MI score, it would only gather information from positions that actually display structural covariation.

Let the set of the six canonical base pairs be denoted BPs. Using Equation (1) over members of the set BPs instead of all 16 possible pairs would not

yield a mutual information since only a subset of the possible occurrences is used. Instead, a relative entropy measure (Cover and Thomas, 1991) is introduced, but for consistency the MI terminology is used.

Let $b_{i,j} \in \{0,1\}$ indicate whether there is a canonical base pair between columns i and j, and let $p(b_{i,j} = 1)$ be the probability of such a base pair. Then, the probability of having a canonical base pair between columns i and j becomes:

$$p(b_{i,j} = 1) = \sum_{ab \in BPs} p_{i,j}(ab).$$

The probabilities of the specific base pairs can be estimated from the actual frequencies. This probability is compared to the probability of observing a base pair by chance given the single nucleotide probabilities. Let this background be denoted $q(b_{i,j} = 1)$:

$$q(b_{i,j} = 1) = \sum_{ab \in BPs} p_i(a) p_j(b).$$

This incorporates only the six canonical base pairs. Using these definitions, the probability for not having a base pair is given as $p(b_{i,j} = 0)$ with background $q(b_{i,j} = 0)$. In combination, this yields the following relative entropy:

$$MI^{W_a}_{i,j} = \sum_{x \in \{0,1\}} p(b_{i,j} = x) \log_2 \frac{p(b_{i,j} = x)}{q(b_{i,j} = x)}.$$

Note that $p(b_{i,j} = 0)$ and $q(b_{i,j} = 0)$ are equal to $1 - p(b_{i,j} = 1)$ and $1 - q(b_{i,j} = 1)$, respectively. Although this is the most rigorous measure, experiments showed that better discrimination was achieved if the term with $x = 0$ was dropped from the measure (data not shown). The final scoring function therefore only considers the structural base pairs, thus filtering noise from non-pairing interactions:

$$\mathrm{MI}^{W}_{i,j} = p(b_{i,j} = 1) \log_2 \frac{p(b_{i,j} = 1)}{q(b_{i,j} = 1)} \tag{3}$$

Intuitively, this measure avoids some of the problems with the standard MI by explicitly focusing on the canonical base pairs. A measure using only the log-odds score $\left(\log_2 \frac{p(b_{i,j}=1)}{q(b_{i,j}=1)} \right)$ was also tried but it did not perform well (data not shown).

A combined measure, that includes an explicit gap penalty, can easily be defined in a manner similar to Equation (2):

$$\mathrm{MI}^{W \diamond P}_{i,j} = MI^{W}_{i,j} - N^{G}_{i,j} \cdot \beta \tag{4}$$

3.5 MI with stacking

The stacking of adjacent base pairs—also known as nearest-neighbour interactions—is a common feature in RNA secondary structure (Borer *et al*., 1974; Onoa and Tinoco, 2004). Therefore it is reasonable to extend the MI measure and incorporate stacking. Using column pair (i,j) as a reference, if $\mathrm{MI}_{i,j}$ is large, stacking implies that the adjacent column pair $(i+1, j-1)$ might also give a good score. By combining the two expressions, stacking would be considered explicitly by the MI.

As mentioned for the standard MI, there is a problem with noise, and this is dramatically increased when considering adjacent columns. In the standard formulation of MI, 16 pairs are considered of which 10 are non-canonical. In the stacking formulation, the combination of two columns lead to a summation over 256 terms of which 220 contain at least one non-canonical pair. Thus, the signal-to-noise ratio decreases.

Instead, a measure incorporating stacking but only considering the canonical base pairs is used. This drastically reduces the number of variables to estimate, which makes it possible to calculate the measure based on most alignments. The measure uses relative entropy and is an extension of the MI^{W} measure. Let $b_{i,j} \in \{0,1\}$ be defined as before, and let $b_{i+1,j-1} \in \{0,1\}$ indicate a base pair at the internal positions. The relative entropy is a sum over the four possible combinations of $b_{i,j}$ and $b_{i+1,j-1}$ (corresponding to pair/pair, pair/not pair, not pair/pair and not pair/not pair):

$$\mathrm{MI}^{S}_{i,j} = \sum_{\substack{x \in \{0,1\} \\ y \in \{0,1\}}} p(b_{i,j} = x, b_{i+1,j-1} = y) \log_2 \frac{p(b_{i,j} = x \mid b_{i+1,j-1} = y)}{q(b_{i,j} = x \mid b_{i+1,j-1} = y)}.$$

The probabilities are estimated from the aligned sequences. For a given column pair in an alignment, let c_{xy} count the number of times that $b_{i,j} = x$ and $b_{i+1,j-1} = y$. For instance, c_{11} is the number of times a canonical pair is observed both between (i,j) and $(i+1,j-1)$. To calculate the relative entropy, only these four numbers are necessary making it a practical measure to use. The joint probability is simply found using the corresponding count:

$$p(b_{i,j} = x, b_{i+1,j-1} = y) = \frac{c_{xy}}{N},$$

where N is the number of sequences. The conditional probability is found using the relation:

$$p(a \mid b) = \frac{p(a,b)}{p(b)}$$

which in the present case gives:

$$p(b_{i,j} = x \mid b_{i+1,j-1} = y) = \frac{\frac{c_{xy}}{N}}{\frac{c_{0y}+c_{1y}}{N}} = \frac{c_{xy}}{c_{0y} + c_{1y}}$$

The background q describes the probability of the observations occurring by chance. The dinucleotide probabilities are estimated from the single nucleotide frequencies. Since q is the null–model, the columns are considered independent. Therefore, the conditional probability $q(a \mid b)$ is simply the probability of the first random variable:

$$q(b_{i,j} = x \mid b_{i+1,j-1} = y) = q(b_{i,j} = x)$$

Practical experiments showed that the measure performed best when only considering positions containing a canonical base pair between columns i and j. This corresponds to always demanding $b_{i,j} = 1$, effectively removing the outer summation. This corresponds well to the final formulation of MI^{W}. The stacking measure using canonical base pairs is:

$$\mathrm{MI}^{S \diamond W}_{i,j} = \sum_{y \in \{0,1\}} p(b_{i,j} = 1, b_{i+1,j-1} = y) \log \frac{p(b_{i,j} = 1 \mid b_{i+1,j-1} = y)}{q(b_{i,j} = 1)}. \tag{5}$$

This measure can be extended with an explicit gap penalty as in Equation (2):

$$\mathrm{MI}^{S \diamond W \diamond P}_{i,j} = \mathrm{MI}^{S \diamond W}_{i,j} - (N^{G}_{i,j} + N^{G}_{i+1,j-1}) \cdot \beta'. \tag{6}$$

Since the number of gap positions is found from two column pairs, the gap penalty β' is modified to fit the possibly larger number of gaps by using $\beta' = \beta/2$.

3.6 Normalized MI

Martin *et al*. (2005) argue that normalizing the standard MI score by the joint entropy of the same random variables yields a more discriminative measure. Given a multiple alignment, let a be a character from column i and let b be a character from column j. The joint entropy of the two columns is given as:

$$H_{i,j} = -\sum_{ab} P_{i,j}(a,b) \log P_{i,j}(a,b),$$

where $P_{i,j}(a, b)$ is the joint probability of observing character a in column i and character b in column j. The joint probability can be estimated from the

frequencies of the dinucleotides. This yields the following expression for the normalized MI:

$$\mathrm{MI}^N_{i,j} = \frac{\mathrm{MI}_{i,j}}{H_{i,j}}, \tag{7}$$

using the definition of $\mathrm{MI}_{i,j}$, from Equation (1).

It has been argued that this normalization removes some of the noise in the MI score (Martin *et al*., 2005): a column pair with a large entropy score will also receive a (not necessarily warranted) large MI score. This skew is removed by normalizing with the entropy. However, this argument is based on protein alignments and assumes that most of the pairs do not show significant MI due to structural and functional constraints. This does not necessarily hold for RNA alignments, where it is known that structure is more conserved than sequence. Furthermore, the number of possible pairs is much smaller in RNA than in proteins, so the MI signal might be better in the present setting.

3.7 RNAalifold measure

In the RNAalifold program an alternative measure of covariation is used (Hofacker *et al*., 2002). Let N be the number of aligned sequences, let α and β denote sequences, $\alpha,\beta = 1,2,\ldots,N$, and let a^α_i denote the character at position i in sequence α. As before, we only consider base pairs in BPs, i.e. the set of Watson–Crick basepairs and the G • U wobble base pair. For each sequence α, the matrix Π^α describes the possible base pairs. Thus, $\Pi^\alpha_{i,j} = 1$ if base pair $(a^\alpha_i, a^\alpha_j) \in BPs$, and $\Pi^\alpha_{i,j} = 0$ otherwise.

Let $\delta(a_i a_j, b_i b_j)$ be the Hamming distance between two base pairs at positions i and j in the alignment, i.e. $\delta = 0$ if the 2 base pairs are identical, $\delta = 1$ if the base pairs vary at exactly one position (consistent substitution), and $\delta = 2$ if the 2 base pairs are different (compensatory mutations). The more mutations observed that retain the base pairing interaction, the more evidence that the base pair is correct:

$$C_{i,j} = \frac{1}{\binom{N}{2}} \sum_{\alpha<\beta} \delta(a^\alpha_i a^\alpha_j, b^\beta_i b^\beta_j)\Pi^\alpha_{i,j}\Pi^\beta_{i,j}$$

To penalize inconsistent pairs (i.e. all non-WC pairs and pairs between a gap and a nucleotide), let $I(a^\alpha_i a^\alpha_j)$ be an indicator variable denoting whether a pair in a sequence α is inconsistent:

$$I(a^\alpha_i a^\alpha_j) = \begin{cases} 0 & \text{if } \Pi^\alpha_{i,j} = 1 \\ 1 & \text{otherwise.} \end{cases}$$

The penalty for the base pair under consideration is then found as:

$$q_{i,j} = \frac{1}{N} \sum_{\alpha=1}^{N} I(a^\alpha_i a^\alpha_j).$$

The combined covariation measure is then given as:

$$B_{i,j} = C_{i,j} - \phi q_{i,j}, \tag{8}$$

where ϕ is a scaling factor for the penalty term. We used $\phi = 1$ as in the original paper.

3.8 The RNAalifold measure including stacking

The RNAalifold covariation already includes gap penalties and treatment of canonical base pairs, and the idea of including stacking information can also be extended to this measure. Originally, we used an asymmetric version that only considered the base pair internal to (i,j) cf. Equation (5). An anonymous referee suggested the symmetric version described here which further improved the performance.

For a pair of columns (i,j), we also consider the neighbouring pairs $(i-1,j+1)$ and $(i+1,j-1)$. Let $\delta(a_i a_j, b_i b_j)$, $\Pi^\alpha_{i,j}$ and $I(a^\alpha_i a^\alpha_j)$ be defined as above. The covariation for a pair $a^\alpha_i a^\alpha_j$ in a sequence α now also depends on $a^\alpha_{i-1} a^\alpha_{j+1}$ and $a^\alpha_{i+1} a^\alpha_{j-1}$. The calculation of the covariation is, as before, found by considering all possible sequence pairs, but now neighbouring nucleotide pairs are considered. The inconsistency penalty $q_{i,j}$ is found in a similar manner.

By normalizing the stacking version to give a score in the same range as the original measure, it becomes clear that the covariation measure B^S can be found as a weighted average of the original RNAalifold score of the three pairs under consideration. Thus, since the covariation between (i,j) and $(i-1,j+1)$ and between (i,j) and $(i+1,j-1)$ is considered, the final formulation becomes:

$$B^S_{i,j} = \frac{B_{i-1,j+1} + 2 \cdot B_{i,j} + B_{i+1,j-1}}{4}. \tag{9}$$

3.9 Base pair probabilities

The partition function described by McCaskill (1990) calculates the probability $P^s(i,j)$ of seeing a base pair between positions (i,j) in sequence s given the nearest-neighbour energy model. The base pairing probabilities are based on the ensemble of all possible structures for the given sequence weighted by the free energy of the individual structures. This information can be used to assign probabilities to proposed base pairs in an alignment of N sequences as follows: first, a probability matrix M^s is calculated for each ungapped sequence s. When a base pair in the alignment is proposed between columns i and j, it has some probability of occurring in each of the N sequences. If a sequence s contains a gap at either of the two positions in the alignment the probability is 0, otherwise the corresponding entry in M^s is used.

Given positions (i,j) in the alignment, let (i^s, j^s) be the original positions in the ungapped sequence s. The partition score for a given base pair is then given as:

$$P_{i,j} = \frac{1}{N} \sum_{s=1}^{N} M^s(i^s, j^s). \tag{10}$$

The score is thus the mean probability assigned to a base pair by the partition function. If a base pair is undefined in a number of sequences, the score is lower due to the lack of evolutionary evidence for that particular base pair. We use the partition function as implemented in the RNAfold program from the Vienna package (Hofacker *et al*., 1994).

4 RESULTS AND DISCUSSION

4.1 Information decreases with overall similarity

The performance of the different measures was analyzed using MCC as a function of threshold values. Graphs showing MCC versus threshold for the individual datasets as well as graphs for the different identity intervals can be found in the Supplementary material. The performance of the different measures is summarized in Figure 1. For each measure, the maximum MCC for all the thresholds we used is shown for each of the three identity intervals. This makes comparison of the relative performance both between measures and between the identity intervals easy.

We also analyzed the performance of the different measures as a function of the number of sequences in the dataset by using sets containing from 2 to 20 sequences (see Supplementary material). All measures performed best with many sequences, which is not surprising, but the relative rating of each measure did not change dramatically.

As expected, all covariation based measures perform best on the low identity datasets with a mean MCC of 0.56. As the overall similarity of the sequences increase, the methods perform worse. This is due to the fact that all these measures rely on a signal from sequence variation. If the sequences are too similar, there is no signal in the alignment and the measures fail. However, the drop in performance varies between the different methods. Some methods are drastically affected (e.g. MI^P and B), while the standard

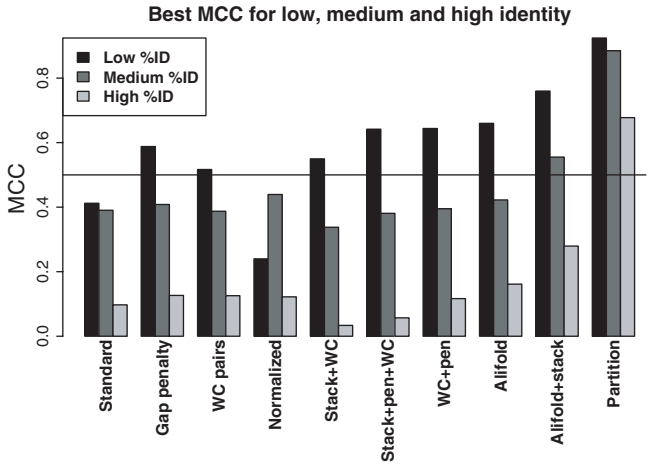

Fig. 1. For each measure, the maximum MCC obtained for the thresholds used in this study is shown for the low, medium and high identity datasets.

mutual information measure is only slightly worse. The best measure is B^S with $MCC = 0.76$.

While most of the sequence variation based measures gave MCC scores greater than 0.5 on the low identity interval, only B^S exceeded 0.5 on the medium identity interval with MCC = 0.56 (mean MCC of 0.41). The normalized mutual information (MI^N) is the only measure that shows an increase in MCC from the low to the medium overall identity datasets. A likely explanation is that the normalized measure depends on the MI measure, which is only slightly affected by the increase in sequence identity. At the same time, the normalization is based on the entropy of the alignment, which decreases as the alignment becomes more 'ordered' with the larger overall identity. This is the second best measure on this identity interval with MCC = 0.44.

When the overall similarity is further increased to the 80–100% interval all measures perform significantly worse, and the mean MCC drops to 0.12. It is clear that using the information content is futile when the sequences are almost identical. The best method is again B^S with MCC = 0.28.

For comparison, the partition function is included in the evaluation. This measure is widely used in RNA structure prediction methods but it is based on a completely different approach. All the other measures use evolutionary information from a sequence alignment, while the partition function uses experimentally obtained energy terms to determine base pairing interactions. Since the partition function does not rely on covariation it is not affected by sequence similarity to the same extent as the other measures. For the purpose of discriminating between true and false base pairs the partition function performs very well on all identity intervals.

It is interesting to see that the partition function also performs better on the low identity datasets. This is not due to covariation of the sequences, but due to the benefits of averaging over a divergent set of structure predictions—an effect that is known from ensemble methods (Krogh and Vedelsby, 1995).

On the low identity datasets the partition function obtains an MCC of 0.92, which is significantly better than the covariation measures. For the 0.60–0.80% interval, the MCC is decreased to 0.88. Finally, the lowest MCC obtained using the partition function (0.68 for the high identity dataset) is only slightly lower than the highest MCC obtained by any other measure (0.76 for B^S on the low identity dataset). Based on these results, it is clear that the partition function is an excellent measure for discriminating between true and false base pairs, as it was also shown in Mathews (2004). Combining the partition function with one of the covariation measures makes obvious sense, such as it is done in e.g. RNAalifold (Hofacker *et al*., 2002).

Table 2. Summary of performance

Measure	Low identity (γ,*MCC*)	Medium identity (γ,*MCC*)	High identity (γ,*MCC*)
MI	(1.01,0.41)	(1.01,0.39)	(0.57,0.10)
MI^P	(0.91,0.59)	(0.88,0.41)	(0.58,0.13)
MI^W	(0.60,0.52)	(0.52,0.39)	(0.08,0.13)
MI^N	(0.30,0.24)	(0.46,0.44)	(0.28,0.12)
$MI^{S\diamond W}$	(0.68,0.55)	(0.78,0.34)	(0.66,0.03)
$MI^{S\diamond W\diamond P}$	(0.50,0.64)	(0.45,0.38)	(0.77,0.06)
$MI^{W\diamond P}$	(0.45,0.64)	(0.41,0.40)	(0.09,0.12)
B	(0.61,0.66)	(0.85,0.42)	(0.58,0.16)
B^S	(0.36,0.76)	(0.52,0.56)	(0.42,0.28)
P	(0.22,0.92)	(0.26,0.88)	(0.19,0.68)

The table shows the optimal pairs of threshold (γ) and MCC obtained in this study for each measure used on the three identity intervals.

4.2 Extending the basic measure

Since the performance is so dependent on the sequence similarity only the results from the low identity datasets will be discussed in the following. Referring to Figure 1 and Table 2 , it can be seen that the standard MI only achieves an MCC of 0.41. The measure is therefore a poor choice for discriminating between true and false base pairs. Normalizing by the entropy (MI^N) as suggested by Martin *et al*. (2005) does not help, on the contrary the performance decreases (MCC = 0.24). A partial explanation for the poor behaviour of these MI scores is that they rely on many frequency estimates, which results in in a poor signal-to-noise ratio unless an unrealistically high number of sequences is available.

Constraining the measure to only consider canonical base pairs (MI^W) gives an increase in MCC to 0.52, and further using the stacking formulation in $MI^{S\diamond W}$ yields MCC = 0.55. This is the expected behaviour since the new formulations should limit the noise and improve the true structural signal. The simple extension of adding a gap penalty (MI^P) gives a considerable improvement of the MCC to 0.59. It should be noted that this improvement in MCC can be due to the alignments used: since gaps in stems are actively avoided in the hand-curated alignments, the bonus from the gap penalty is boosted. Nevertheless, since gaps are unwanted in structurally important stretches, this result is promising.

Combining the use of canonical base pairs with a gap penalty ($MI^{W\diamond P}$) gives a good MCC of 0.64, which is significantly better than either one alone. Adding a gap penalty to the stacking measure ($MI^{S\diamond P\diamond W}$) increases the MCC to 0.64 compared to stacking alone. The gap penalty thus shows good improvements when used in combination with the different measures. It is possible that the gap penalty could be further optimized, since it is a simple formulation that is used.

The measure used in RNAalifold also limits the allowed base pairs and explicitly penalizes gaps and inconsistent pairs. This gives a good MCC of 0.66 which is better than any of the MI based measures. The stacking version of the RNAalifold measure (B^S) is even better giving an MCC of 0.76. The RNAalifold measure and its stacking extension are comparable to $MI^{W \diamond P}$ and $MI^{S \diamond P \diamond W}$, respectively: they all include gap penalties and only focus on canonical base pairs. The main difference is the explicit penalty for inconsistencies in the RNAalifold measure, which gives a significant improvement in MCC.

The advantage of using the symmetric stacking is that the level of noise is smoothed while the signal is only slightly affected (mainly in stem ends), thus improving the signal-to-noise ratio. In parallel with the B^S measure, a symmetric version of the $MI^{S \diamond W}$ measure was tested. However, experiments showed that the performance decreased which is possibly due to the decreased signal-to-noise ratio as a result of the increase in variables to estimate (data not shown). We also tested a next-to-nearest-neighbour version of the B^S measure which resulted in a slight improvement. However, the weighting scheme was rather *ad hoc* and could not be justified (the fourth row of Pascal's triangle; data not shown). It might be worth doing further investigations into this stacking model.

5 CONCLUSION

The standard MI is not well suited for the task of secondary structure prediction, and not all the proposed extensions can remedy this. Adding a simple gap penalty, though, greatly increases the performance as shown above. Likewise, counting only the most common base pairs also gives better results. It is also seen that combining different extensions in general improve the MCC.

The covariation measure used in RNAalifold (Hofacker *et al*., 2002) performs very well and is a good choice due to its simplicity. The simple extension to include stacking in this measure also shows potential and might be worth exploiting in the future. Of the covariation based measures evaluated in this work, this was the most discriminative. The partition function used for comparison was the most powerful, though. The performance of the individual measures for the three similarity classes is summarized in Table 2 together with the threshold values.

In this study, we have analyzed a number of measures to distinguish true and false base pairs. Standard MI is still widely used, but our evaluation indicates that this is not the best measure. The results presented here should give other researchers an idea of useful information measures for RNA secondary structure prediction and—just as importantly—an idea of which measures are not useful.

ACKNOWLEDGEMENTS

The authors would like to thank Jonathan P. Bollback for discussions regarding inclusion of stacking in mutual information, and Ivo L. Hofacker for information on the RNAalifold covariation measure. The authors also thank four anonymous reviewers for helpful comments, especially reviewer 4 for suggesting the symmetric version of RNAalifold with stacking. S.L. was supported by a Novo Scholarship. P.P.G was supported by a Carlsberg Foundation Grant (21-00-0680).

Conflict of Interest: none declared.

REFERENCES

Akmaev,V.R. *et al*. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics*, **16**, 501–512.

Bindewald,E. and Shapiro,B. (2006) RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, **12**, 342–352.

Borer,P. *et al*. (1974) Stability of ribonucleic acid double-stranded helices. *J. Mol. Evol*., **86**, 843–853.

Chiu,D.K. and Kolodziejczak,T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci*., **7**, 347–352.

Coventry,A. *et al*. (2004) MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 12102–12107.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. John Wiley & Sons, Inc.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Eddy,S. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*., **22**, 2079–2088.

Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.

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

Gorodkin,J. *et al*. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*., **25**, 3724–3732.

Gorodkin,J. *et al*. (1999) Matrixplot: visualizing sequence constraints. *Bioinformatics*, **15**, 769–770.

Griffiths-Jones,S. *et al*. (2003) Rfam: an RNA family database. *Nucleic Acids Res*., **31**, 439–441.

Gutell,R.R. *et al*. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*., **20**, 5785–5795.

Havgaard,J.H. *et al*. (2005) Pairwise local structure alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.

Hofacker,I.L. *et al*. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, **125**, 167–188.

Hofacker,I.L. *et al*. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol*., **319**, 1059–1066.

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

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

Krogh,A. and Vedelsby,J. (1995) *Advances in Neural Information Processing Systems. Chapter Neural Network Ensembles, Cross Validation and Active Learning*. MIT Press, pp. 231–238.

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat*., **22**, 79–86.

Lee,J. and Gutell,R. (2004) Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. *J. Mol. Biol*., **344**, 1225–1249.

Leontis,N. *et al*. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res*., **30**, 3497–3531.

Lück,R. *et al*. (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res*., **27**, 4208–4217.

Martin,L.C. *et al*. (2005) Using information theory to search for co–evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.

Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Bio*., **317**, 191–203.

Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.

Mathews,D. *et al*. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.

Matthews,B. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, **405**, 442–451.

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

Onoa,B. and Tinoco,I.,Jr (2004) RNA folding and unfolding. *Curr. Opin. Struct. Biol*., **14**, 374–379.

Pedersen,J. *et al*. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*., **2**.

Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**.

Ruan,J. *et al*. (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58–66.

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

Shannon,C.E. (1948) A mathematical theory of communication. *The Bell Syst. Tech. J*., **27**, 379–423, 623–656.

Szymanski,M. *et al*. (2002) 5S ribosomal RNA database. *Nucleic Acids Res*., **30**, 176–178.

Washietl,S. *et al*. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*., **9**, 133–148.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.

Zwieb,C. (1997) The uRNA database. *Nucleic Acids Res*., **25**, 102–103.