BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 2 2011, pages 211–219
doi:10.1093/bioinformatics/btq634

Structural bioinformatics

Advance Access publication November 18, 2010

PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences

Stefan E. Seemann[1,2,†], Andreas S. Richter[3,†], Tanja Gesell[4], Rolf Backofen[1,3,5,*] and Jan Gorodkin[1,2,*]

[1]Center for non-coding RNA in Technology and Health, [2]IBHV, University of Copenhagen, Grønnegårdsvej 3, Frederiksberg C, 1870, Denmark, [3]Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, Freiburg, 79110, Germany [4]Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories (MFPL), University of Vienna, Medical University of Vienna and University of Veterinary Medicine, Dr. Bohr-Gasse 9, Vienna, 1030, Austria and [5]Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Albertstrasse 19, Freiburg, 79104, Germany

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Predicting RNA–RNA interactions is essential for determining the function of putative non-coding RNAs. Existing methods for the prediction of interactions are all based on single sequences. Since comparative methods have already been useful in RNA structure determination, we assume that conserved RNA–RNA interactions also imply conserved function. Of these, we further assume that a non-negligible amount of the existing RNA–RNA interactions have also acquired compensating base changes throughout evolution. We implement a method, PETcofold, that can take covariance information in *intra*-molecular and *inter*-molecular base pairs into account to predict interactions and secondary structures of two multiple alignments of RNA sequences.

Results: PETcofold's ability to predict RNA–RNA interactions was evaluated on a carefully curated dataset of 32 bacterial small RNAs and their targets, which was manually extracted from the literature. For evaluation of both RNA–RNA interaction and structure prediction, we were able to extract only a few high-quality examples: one vertebrate small nucleolar RNA and four bacterial small RNAs. For these we show that the prediction can be improved by our comparative approach. Furthermore, PETcofold was evaluated on controlled data with phylogenetically simulated sequences enriched for covariance patterns at the interaction sites. We observed increased performance with increased amounts of covariance.

Availability: The program PETcofold is available as source code and can be downloaded from http://rth.dk/resources/petcofold.

Contact: gorodkin@rth.dk; backofen@informatik.uni-freiburg.de

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

Received on April 11, 2010; revised on October 24, 2010; accepted on November 8, 2010

1 INTRODUCTION

Non-coding RNAs (ncRNAs) are receiving rapidly growing attention; they are present in large numbers as observed by The ENCODE Project Consortium (2007) and the FANTOM2 project (Ravasi *et al.*, 2006), and they appear in many unexpected cases (e.g. Mercer *et al.*, 2008). The potential of functional ncRNAs is further supported by the observation that the amount of non-protein-coding DNA increases with organismal complexity, whereas the amount of protein-coding DNA is relatively constant (Taft *et al.*, 2007). A substantial number of putative ncRNAs has emerged from several genomic *in silico* screens for RNA structure taking compensatory base pair changes into account (Torarinsson *et al.*, 2006, 2008; Washietl *et al.*, 2005; Weinberg *et al.*, 2007; Will *et al.*, 2007). Deep sequencing approaches are another growing source of ncRNAs (e.g. Sharma *et al.*, 2010).

One step towards assigning functions to these putative ncRNAs is to consider RNA interactions. An important class of these are RNA–RNA interactions; many ncRNAs base pair with other RNAs, such as the majority of characterized bacterial small regulatory RNAs (sRNAs) (Waters and Storz, 2009). Besides the well-known examples of microRNAs (miRNAs) and small interfering RNAs (siRNAs), there are also many longer eukaryotic ncRNAs that act via RNA–RNA interaction. Examples are small nucleolar RNAs (snoRNAs), of which the main function is to guide the editing of ribosomal RNA (Bachellerie *et al.*, 2002). However, there is growing evidence that other long ncRNAs might act via base pair interactions as well. For example, it is likely that certain long ncRNAs modulate the activity of miRNAs by forming RNA–RNA interactions (Wilusz *et al.*, 2009).

Even though most screens for ncRNA candidates involve mammalian genomes (Gorodkin *et al.*, 2010) we predominantly found with the exception of a few snoRNA examples, bacterial examples of verified RNA–RNA interactions involving the structure of both RNAs. This pattern might be the reason that many of the existing methods for the prediction of RNA–RNA interactions have been designed to work on bacterial sRNAs [see Backofen and Hess (2010) for a review]. Furthermore, extensive curated data for RNA–RNA interactions are limited because this effort is, just as

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

for RNA structural alignments, rarely acknowledged (Menzel *et al.*, 2009).

Recent interest in prediction of RNA–RNA interactions has led to four main classes of computational methods to solve this task. Initial approaches evaluated only base pairs involved in duplex formation (e.g. Rehmsmeier *et al.*, 2004; Tafer and Hofacker, 2008). The second class consists of methods that predict a joint secondary structure of two interacting RNAs by folding their concatenated RNA sequences (e.g. Andronescu *et al.*, 2005; Bernhart *et al.*, 2006; Dirks *et al.*, 2007). This concatenation approach considers only a restricted set of interaction types, and therefore, the third class evolved. Here, all interactions with one site are allowed and interaction site accessibility is calculated by the partition function of the structure ensemble (Bompfünewerer *et al.*, 2008; Busch *et al.*, 2008; Mückstein *et al.*, 2006, 2008; Richter *et al.*, 2010). The final class of methods handles more complex joint structures with multiple interaction sites (Alkan *et al.*, 2006; Chitsaz *et al.*, 2009a, b; Huang *et al.*, 2009, 2010; Pervouchine, 2004; Salari *et al.*, 2010a, b). However, these methods are very resource-demanding, which makes them not applicable for genome-wide scans. Furthermore, all aforementioned methods evaluate only interactions between a pair of single sequences.

For an approach that makes use of multiple alignments, our assumption is that a non-negligible amount of the existing RNA–RNA interaction contains compensatory changes across the binding sites. Whereas this is a motivating aspect, the general variation in the sequences with even a completely conserved interaction site might also yield prediction improvements. Such an improvement over single sequence-based methods can be obtained by taking the overall sequence variation into account for the joint framework of energy folding and covariation.

The literature contains only limited examples of conserved RNA–RNA interactions with likely compensating base pair changes, such as the MicA–*ompA* interaction, where base pairing is preserved by compensatory changes in several enterobacterial species (Udekwu *et al.*, 2005). These authors also developed a method for predicting interactions that takes phylogenetic conservation into account, but to our knowledge, they have not yet published it. One explanation for the limited amount of data containing covariance information might well be that most existing data have been found using sequence similarity-based methods such as BLAST (Altschul *et al.*, 1997) to find homologs of the interacting RNAs. Such an approach is expected to lead to a collection of RNAs that are highly conserved in the primary sequence rather than structure. Although in these cases it is likely that the interaction pattern is conserved as well, only a small number of compensatory base pair changes is expected. Therefore, we also include simulated data based on substitution statistics of interacting base pairs. Applying simulated data to make controlled tests or to supplement existing data has been done several times before (e.g. Kolbe and Eddy, 2009).

To summarize, there are three main problems for the prediction of RNA–RNA interactions: (i) the small number of examples with mapped interactions, (ii) most existing computational methods work on single sequences and therefore suffer from a low specificity and (iii) the search for complex types of interactions is expensive if a guaranteed maximum score is to be obtained. The last is discussed in more details in Seemann *et al.* (2010), where we propose an algorithm for searching for RNA–RNA interactions between two multiple RNA sequence (or sequence-structure) alignments. Our algorithm can compute a semi-optimal combination of *intra-* and *inter*-molecular base pairs to predict a joint secondary structure of two multiple alignments of RNA sequences, each representing its own evolutionary and structurally conserved RNA. The method makes use of the idea of a linker from RNAcofold (Bernhart *et al.*, 2006) by concatenating both RNA sequences, but employs this idea in the context of PETfold (Seemann *et al.*, 2008) along with a strategy for hierarchical folding (e.g. Gaspin and Westhof, 1995). PETfold is based on Pfold (Knudsen and Hein, 2003), which provides reliabilities for evolutionarily conserved base pairs, and unifies them with folding energies in one model. Hierarchical folding allows for the prediction of pseudoknots between *intra-* and *inter*-molecular base pairs but is still fast enough for genome-scale applications. Here we present its implementation, PETcofold, and an in-depth performance evaluation. A *key motivation* for implementing a method that can predict RNA–RNA interactions between two multiple alignments of RNA sequences is the existence of many ncRNA candidates identified by genome-wide *in silico* screens as mentioned before. These *de novo* predicted RNA structures exist already as multiple alignments and the PETcofold implementation presented here can readily be employed to further analyse these candidates for potential interaction partners.

2 METHODS

2.1 Algorithm and implementation

The algorithm presented in Seemann *et al.* (2010) is implemented here in the PETcofold program. The input consists of two RNA alignments $\mathcal{A}_1$ and $\mathcal{A}_2$, and the program performs hierarchical folding in two steps. This multi-step approach is motivated by the observation that interaction formation is often initiated at well-accessible *intra*-molecular structures such as hairpin loops (Brunel *et al.*, 2002). Furthermore, several existing methods are based on this observation and assume that the interaction sites are made accessible to allow for hybridization of the two RNAs (e.g. Busch *et al.*, 2008; Mückstein *et al.*, 2006). The workflow of the PETcofold pipeline is shown in Figure 1. Supplementary Figure S1 shows a flow chart of the different programs integrated in PETcofold.

In step 1, we search for highly reliable base pairs in both RNA alignments, these are interpreted as being not accessible for the RNA–RNA interaction. We apply the maximum expected accuracy approach of PETfold separately on the alignments $\mathcal{A}_1$ and $\mathcal{A}_2$ to calculate reliabilities unifying the thermodynamic probabilities from RNAfold (Hofacker *et al.*, 1994) and the evolutionary reliabilities from Pfold. The set of all base pairs with high base pair reliability $\mathcal{R}_{bp}$ in the folding of the individual alignments, i.e. $\mathcal{R}_{bp}$ is greater than a threshold δ, forms a partial structure σ^{p} and will be denoted as σ^{p}_1 and σ^{p}_2 for $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. To ensure that the highly reliable base pairs of the partial structure σ^{p} are also part of the final (consensus) structure, they should make up a significant portion of the probability $\Pr[\mathcal{E}(\sigma^{\mathrm{p}})]$ in either the thermodynamic or the evolutionary model, where $\mathcal{E}(\sigma^{\mathrm{p}})$ is the ensemble of structures that are compatible with the partial structure σ^{p}. A high value of $\Pr[\mathcal{E}(\sigma^{\mathrm{p}})]$ is guaranteed by the introduction of a second threshold γ. The threshold δ for highly reliable base pairs is now increased until the probability of $\mathcal{E}(\sigma^{\mathrm{p}})$ exceeds γ either in the evolutionary or the thermodynamic model or both. In addition, constrained stems in the partial structure σ^{p} are extended by inner and outer base pairs if the average reliability of the extended stem is larger than δ and if the partial structure probability $\Pr[\mathcal{E}(\sigma^{\mathrm{p}})]$ exceeds γ. This feature is enabled by the option *-extstem*.

In step 2, we concatenate the input alignments $\mathcal{A}_1$ and $\mathcal{A}_2$ with a linker symbol '&' to search for conserved interactions and structures of

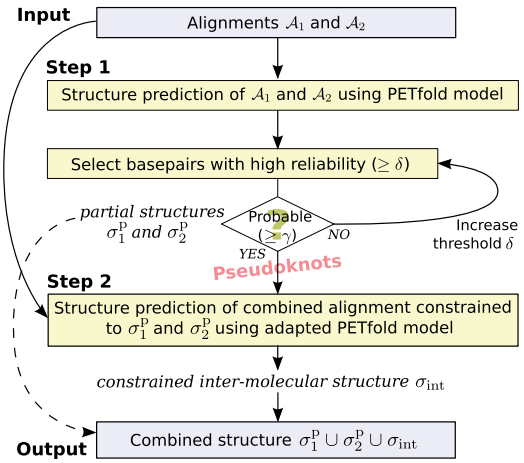

Fig. 1. The PETcofold pipeline consists of two steps: (1) *intra*-molecular folding by PETfold and selection of a set of highly reliable base pairs that only decreases the probability of the ensemble in some pre-defined range; (2) *inter*-molecular folding by adapted PETfold using constraints from step 1. In the end, partial structures and constrained *inter*-molecular structures are combined to form the joint RNA secondary structure including pseudoknots.

the sequences of the two alignments. On the concatenated alignment, we apply an adapted PETfold model that can handle fixed partial structures σ_1^{p} and σ_2^{p} from the first step by constrained expected accuracy scoring, which is an extension of PETfold's maximum expected accuracy scoring for constrained folding. We search for a joint structure σ of the combined alignment that extends both σ_1^{p} and σ_2^{p}, i.e. $\sigma \supseteq \sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}$. Although PETfold cannot handle pseudoknots and the linker in step 2 forbids pseudoknots in the concatenated sequences, i.e. the resulting structure has to be nested, the hierarchical folding strategy of PETcofold allows pseudoknots between *intra*-molecular base pairs from step 1 and *inter*-molecular (as well as *intra*-molecular) base pairs from step 2 by restricting the positions of the concatenated alignments that are covered by base pairs from σ_1^{p} and σ_2^{p} to be single stranded. Under these constraints, the thermodynamic probabilities are calculated with RNAcofold, and the evolutionary reliabilities are calculated with an adapted Pfold using constrained folding, which yields the structure σ_{int}. This constrained folding results in raw probabilities, which are then weighted by the product of partial structure probabilities $\Pr[\mathcal{E}(\sigma^{\mathrm{p}})]$ from step 1. However, to avoid underestimating the probabilities of step 2, we here replace the product with the geometric mean of partial structure probabilities. The constrained expected accuracy structure σ_{int} is calculated by a Nussinov-style algorithm and might contain both *intra*-molecular as well as *inter*-molecular base pairs. The final consensus structure including the interaction is then

$$\sigma = \sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}} \cup \sigma_{\mathrm{int}}.$$

In the final scoring, the reliabilities of all base pairs from σ_1^{p} and σ_2^{p} are equal to the partial structure probabilities $\Pr[\mathcal{E}(\sigma_1^{\mathrm{p}})]$ and $\Pr[\mathcal{E}(\sigma_2^{\mathrm{p}})]$, respectively.

The algorithm, which is described in more detail in Seemann *et al.* (2010), has a time complexity of $O(N \times I \times L^3)$, where N is the number of sequences, I is the number of iterations in the adaptation of δ to find probable partial structures and L is the sum of the sequence lengths of both alignments.

2.2 Datasets and processing

The evaluation of PETcofold was performed using datasets extracted from the literature combined with simulated data, in which the degree of compensating base changes is controlled.

First, a dataset of interactions between bacterial sRNAs and their target mRNAs was extracted. Starting from a set of experimentally verified interactions used by Busch *et al.* (2008), we included further sRNA–mRNA interactions with experimental support from the literature. The final dataset contained 13 different sRNAs and 32 interactions from *Escherichia coli* K12 (*E.coli*), *Salmonella typhimurium* LT2 and *Staphylococcus aureus* N315 (see Supplementary Table S1).

For each sRNA, the RNA family sequence alignment was downloaded from the Rfam database 9.1 (Gardner *et al.*, 2009). Orthologs of target genes were predicted in species with available complete genomes according to the Rfam annotation. The prediction of orthologous genes was carried out with OrthoMCL (Li *et al.*, 2003) using the default parameters. For each target gene, a 250 nt subsequence was extracted (150 nt upstream and 100 nt downstream of the annotated translational start sites) because all interactions occurred from −132 nt to +56 nt relative to the start codon. The sets of putative orthologous target sequences were locally aligned with MAFFT (using option E-INS-i for generalized affine gap costs) (Katoh and Toh, 2008).

The resulting dataset was processed by homology reduction and removal of sequences that were very distant to the reference, i.e. the organism in which the interaction was detected. This step aims to remove false positive predicted orthologs and was achieved by excluding all mRNA sequences with a low pairwise sequence identity ($\mathrm{PI}_{\mathrm{ref}}^{\mathrm{int}}$) at the interaction site and within 10 nt of the flanking sequences compared with the reference sequence. $\mathrm{PI}_{\mathrm{ref}}^{\mathrm{int}}$ thresholds of 40, 50 and 60% were applied. To avoid a bias caused by overweighting redundant sequence information, mRNA sequences were clustered with the BLASTClust tool (Altschul *et al.*, 1997) using a word size of 8 and a percent identity threshold of 100% over an area covering 90% of each sequence. The sequence with the lowest $\mathrm{PI}_{\mathrm{ref}}^{\mathrm{int}}$ was taken from each cluster. Further details on the preparation of the dataset are given in the Supplementary Material.

To evaluate the covariance of a dataset, the arithmetic mean of compensatory base pair (CBP) exchanges of all interactions was computed. For a specific interaction, the CBP is computed as the average number of compensatory base pairs in all interacting alignment columns, where a compensating base pair is distinct from the bases involved in the pairing of the reference sequence. Thus, the maximal value of the CBP is 5. Another covariance measure is the probability $\Pr[\sigma_{i,j}|\mathcal{A},T,M]$ of a base pair (i,j) calculated by Pfold. It is more accurate because it takes into account, by the tree T, the evolutionary distance of the sequences in the alignment $\mathcal{A}$. The model M describes the substitution rates of base pairs and unpaired bases and the probabilities of secondary structure production rules (Knudsen and Hein, 1999); however, the bias introduced by M can be ignored. To measure the covariance of a full interaction, we use the normalized Pfold reliability $\mathcal{R}(\mathrm{int})$, which is the mean of all base pair probabilities in the interaction.

2.3 Simulated data

The dataset described above is based on sRNA–mRNA interactions that were experimentally validated in one reference organism. The incorporation of homologous sequences from Rfam families and computational prediction of orthologous genes is often used. However, this approach could be limited in two respects. From a biological point of view, our method assumes that the homologous sRNAs regulate the same targets by the same mechanism in all organisms that are contained in the sequence sets. Although this assumption probably holds true for many interactions, it does not have to apply to all of them. Furthermore, the target orthologs might contain false positive predictions with different physiological functions and regulatory mechanisms. From a technical point of view, weak covariance at the interaction sites limits the full potential of PETcofold. For example, the most conserved region of the sRNA SgrS is involved in base pairing its target (Horler and Vanderpool, 2009).

Thus, we created a simulated dataset with increased covariance. We used SISSI (Gesell and von Haeseler, 2006) to simulate sequence data with site-specific interactions annotated by the sRNA–mRNA interactions of our dataset along phylogenetic trees. To be biologically relevant, we estimated each phylogenetic tree from the corresponding alignment using a maximum likelihood method with an independent model. For tree

reconstructions, IQPNNI 3.3.1 (Vinh and von Haeseler, 2004) was used with a F81 model (Felsenstein, 1981); otherwise, the default settings were used. To specify the rate matrices, we simply counted the frequencies of the nucleotides {A, C, G, U} for sites evolving independently and doublet frequencies for the distant RNA interaction pairs. For each of the 32 alignments, we performed 20 simulations with the same length and in the context of the annotations of the sRNA–mRNA interactions using a Markov model of nucleotide sequence evolution (Gesell and von Haeseler, 2006) with rate matrix types of the F81 model. We initially started simulations along the estimated phylogenetic trees. Then, to increase the covariance, we multiplied each branch length by a scaling factor. The simulation runs were repeated under the same parameters with eight different scaling values as shown in Table 1.

2.4 Performance evaluation

All predictions were evaluated by calculating their correlations to the structures from literature ignoring non-canonical base pairs. We used the Matthews correlation coefficient (Matthews, 1975) defined as

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

where TP is the number of mutual base pairs in the two assignments (true positives), TN are the true negatives, FP is the number of predicted base pairs not in the annotated assignment (false positives) and FN are the false negatives. For RNA secondary structures, the geometric mean of sensitivity ($\mathrm{SEN}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$) and positive predictive value ($\mathrm{PPV}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$), $\sqrt{\mathrm{SEN}*\mathrm{PPV}}$, is a good approximation of the MCC (Gorodkin *et al.*, 2001) and, here, is used to compare predicted joint secondary structures with the annotations. The MCC should be maximized to achieve the best possible trade-off between sensitivity and PPV.

The MCC of the prediction to the annotation is computed for each sequence in the alignment. The arithmetic mean of these single MCCs gives the mean MCC of a prediction. Arithmetic mean and median MCC of a whole dataset are calculated from the mean MCCs of all interactions contained in the dataset.

PETcofold is a comparative approach that detects conserved joint secondary structures. Hence, to compare with single sequence-based approaches, we also determined conserved consensus structures from the results of inteRNA, Pairfold, RactIP and RNAcofold. The consensus structure is defined by all base pairs that are conserved in a given percentage of single structures (here: 80 and 100%) of each of the sequences in the multiple alignment.

3 RESULTS

3.1 Parameter estimation on bacterial sRNA–mRNA interactions

The performance of PETcofold under various parameter settings was evaluated on a dataset of 32 bacterial sRNA–mRNA interactions. The PETcofold parameter δ sets the maximal *intra*-molecular base pair reliability of bases to be free for *inter*-molecular folding, and the parameter γ sets the minimal partial structure probability. For each of the two parameters, 11 values ranging from 0.0 to 1.0 in steps of 0.1 were tested yielding 121 parameter combinations. Furthermore, we tested the influence of the option *-noLP*, which disallows pairs that can only occur isolated in the thermodynamic part, i.e. it is used as an option for RNA(co)fold. Columns with more than 50% gaps were removed.

The influence of the data composition on PETcofold's prediction quality was analysed by varying the minimal pairwise interaction site sequence identity ($\mathrm{PI}^{\mathrm{int}}_{\mathrm{ref}}$).

PETcofold was applied to three datasets containing all 32 interactions, which were created by using $\mathrm{PI}^{\mathrm{int}}_{\mathrm{ref}}$ of 40, 50 or 60%. Figure 2 shows the 3D plot of the mean interaction MCCs for all parameter combinations tested for 60% $\mathrm{PI}^{\mathrm{int}}_{\mathrm{ref}}$ (without *-noLP*). Here, PETcofold yielded the best performance for $\delta=0.9$ and γ ranging from 0.0 to 0.5. The 3D plots for $\mathrm{PI}^{\mathrm{int}}_{\mathrm{ref}}$ of 40 and 50% are given in Supplementary Figure S2. Of all three datasets, the best mean MCC was 0.511 (median MCC: 0.583) for 50% $\mathrm{PI}^{\mathrm{int}}_{\mathrm{ref}}$, $\delta=\gamma=0.9$ and option *-noLP*. When taking all runs into account, the option *-noLP* had only marginal influence on the mean MCCs (see Supplementary Table S2).

The average number of compensatory interaction base pair exchanges (CBP) in the three datasets ranged from 0.1 to 0.2 (see Supplementary Table S2c). Since the best mean MCCs were very similar and differed by 0.02 at most, the variation in covariance observed here seemed to have no strong influence on the MCC.

The numerical experiments indicate that the use of *intra*-molecular constraints improves the prediction of interaction sites. Nevertheless, the probability threshold δ for base pairs in the partial structure σ^{p} has to be fairly high to achieve the best MCCs. This is in agreement with Seemann *et al.* (2010), in which we showed that the structural mass of *intra*-molecular partial structures has to be high to support the constrained expected accuracy scoring. When setting δ to 1, then no base will be constrained by base pairing for the *inter*-molecular folding step. Consequently, no loop–loop interactions between the two single structures are allowed. We found that a δ of 0.9 yields the best MCC. This setting forbids interactions in highly structured regions and, thus, accounts for the importance of interaction site accessibility.

The parameter γ adapts the probability of partial structures to cover a high mass of the entire ensemble of structures. This strategy avoids constraints that are not compatible with reliable alternative structures supporting the interaction site. In five out of six sets of input alignments, we observed the best performance for $\gamma=0.9$ (see Supplementary Table S2). For multiple alignments with $\mathrm{PI}^{\mathrm{int}}_{\mathrm{ref}}$ of 60%, we achieve the best performance for $0\leq\gamma\leq0.5$. In real applications, sequence-based alignments are often used as input, which perform satisfactorily for pairwise sequence

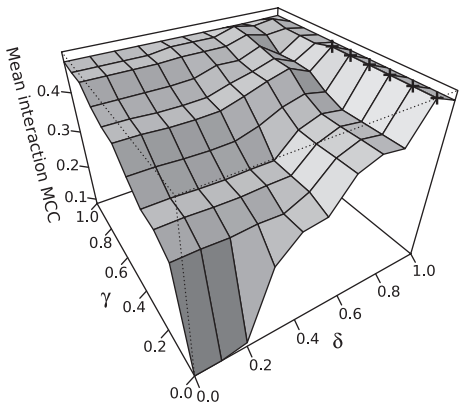

Fig. 2. The performance of PETcofold while varying the parameters δ (maximal *intra*-molecular base pair reliability) and γ (minimal partial structure probability). The 3D plot shows the mean MCC of 32 interactions using input sequences with a minimal pairwise mRNA interaction site sequence identity to the reference of 60%. Predictions were carried out without the option *-noLP*. The maximal MCC is marked with '+'.

identities above 60%. For the prediction of joint secondary structures, we use a rather restrictive γ of 0.1 to allow many constraints for *intra*-molecular base pairs. Furthermore, on the limited set of joint structures presented in Section 3.3, we achieved no performance increase by a higher value for γ (see Supplementary Table S3).

3.2 Simulated interactions with increased covariance

The average number of compensatory interaction base pair exchanges (CBP) in the 32 interactions from Section 3.1 was rather low with values of 0.1–0.2. Therefore, we created a simulated sequence dataset based on these interactions as described in Section 2.3. Here, covariance at the interaction sites was increased by scaling the branch lengths of the phylogenetic trees that guided the simulation. As shown in Table 1, the minimal CBP computed from the simulated data was 0.8, which was higher than the CBP of the real data. This was due to the fact that the simulations only covered a subset of the evolutionary constraints on the sequences. For example, only the interactions without thermodynamics were used as structural constraints. Other evolutionary pressures on the sequences were neglected. However, taking into account all aspects in the simulations with a corresponding maximum likelihood framework is beyond the scope of this article. At this point, we focused on evaluating PETcofold's performance on datasets with increased covariance.

We applied PETcofold to these datasets and computed the mean interaction MCC of all 20 simulation runs for 9 different phylogenetic scaling factors. Figure 3 shows the mean interaction MCC plotted over the phylogenetic scaling factor for PETcofold predictions with $\delta=0.9$, $\gamma=0.1$ and the option *-noLP*. The mean MCC of the predicted interactions was 0.505 for scaling factor 1. The prediction accuracy of PETcofold in terms of MCC increased with increasing scaling factors. A mean MCC of 0.821 was achieved for scaling factor 200. Table 1 shows for all scaling factors the mean interaction MCC and the covariance in the data as evaluated by $\mathcal{R}$(int) and CBP. It can be clearly seen that the performance of PETcofold correlated with the covariance at the interaction sites.

3.3 RNA joint secondary structures

In Seemann *et al.* (2010), we gave the theoretical background of PETcofold and presented the joint structure prediction of a prototype implementation on the well-known example OxyS-*fhlA* (Argaman and Altuvia, 2000). Additional examples are presented in the following and PETcofold's performance is explored more quantitatively using the MCC measure and a comparison to other joint structure prediction methods. The predictions for the examples MicA-*ompA*, OxyS-*fhlA*, RyhB-*sodB* and RyhB-*uof-fur* were compared with the interaction models based on structural mapping as described by Argaman and Altuvia (2000); Geissmann and Touati (2004); Udekwu *et al.* (2005) and Večerek *et al.* (2007). The interactions of RyhB-*sodB* and OxyS-*fhlA* involve one and two loop–loop interactions, respectively. Consequently, their joint structures contain pseudoknots between the single structure and the interaction.

For the structure prediction, the dataset with a minimal mRNA interaction site identity (PI^{int}_{ref}) of 60% was used. To make the predictions comparable with the annotated structures, an mRNA subsequence as given in the proposed interaction complex model was used instead of the 250 nt subsequence as described in Section 2.2. Regarding the sRNAs, the Rfam entry of RyhB (RF00057) missed the first 29 nt of the *E.coli* RyhB sequence. However, this subsequence is involved in forming the secondary structure. Thus, homologs of the RyhB sRNA were searched with the semi-global alignment tool GotohScan 1.3 (Hertel *et al.*, 2009) using defaults parameters. The genome-wide search was conducted with an *E*-value threshold of 1e-3 in all organisms contained in the RyhB alignment of the aforementioned dataset. Homologs were found in all of these organisms except *Vibrio cholerae* O395. A multiple sequence-structure alignment of the identified homologous RyhB sequences was computed with LocARNA 1.5a (Will *et al.*, 2007). The mRNA and sRNA alignments of the four examples were hand curated by removal of redundant sequences and of sequences that were very distinct from the reference organism *E.coli*, in which the interaction models were experimentally determined.

To predict the joint structures, PETcofold was called with parameters $\delta=0.9$, $\gamma=0.1$, option *-noLP* and optionally with option *-extstem*. Our evaluation included the following single sequence-based methods: the sparsified version of inteRNA (Salari *et al.*, 2010b), Pairfold (Andronescu *et al.*, 2005), RactIP (Kato *et al.*, 2010), all with default parameters, and RNAcofold (Bernhart *et al.*, 2006) using parameters *-d2 -noLP*. All methods apart from RactIP predict minimum free energy joint secondary structures. inteRNA is based on the model by Chitsaz *et al.* (2009b), whereas Pairfold and RNAcofold are

Table 1. Prediction performance of PETcofold on simulated sequence data

Scaling factor	Mean branch length	$\mathcal{R}$(int)	CBP	Mean MCC
1	0.03	0.486	0.832	0.505
5	0.15	0.665	1.818	0.677
10	0.3	0.699	2.182	0.717
25	0.75	0.732	2.550	0.743
50	1.5	0.775	2.755	0.773
75	0.75	0.791	2.858	0.783
100	3.0	0.801	2.908	0.790
150	4.5	0.822	2.988	0.802
200	6.0	0.839	3.040	0.821

Scaling factor multiplies each branch length of the phylogenetic tree. Mean branch length denotes the mean of the mean branch lengths for all 32 phylogenetic trees. $\mathcal{R}$(int) denotes the mean base pair probability at the interaction sites (calculated by Pfold). CBP denotes the average number of compensatory interaction base pair exchanges in the simulated sequence data. MCC evaluates only the interaction. PETcofold was called with parameters $\delta=0.9$, $\gamma=0.1$ and option *-noLP* to forbid lonely base pairs.

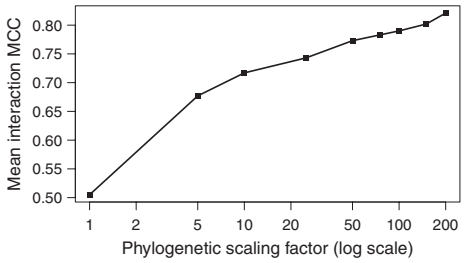

Fig. 3. Prediction performance of PETcofold on phylogenetically simulated sequence data for 32 interactions. Covariance of the data were increased by multiplying each branch length with a phylogenetic scaling factor. PETcofold was called with parameters $\delta=0.9$, $\gamma=0.1$ and option *-noLP*.

Table 2. Performance of PETcofold and other joint structure prediction methods on four sRNA–mRNA examples

sRNA-target pair	MCC										Run time (s)					
	PETcofold		inteRNA		Pairfold		RactIP		RNAcofold		PETcofold		inteRNA	Pairfold	RactIP	RNAcofold
		-extstem	80%	100%	80%	100%	80%	100%	80%	100%		*-extstem*				
MicA-*ompA*	0.87	0.83	0.49	0.51	0.86	0.74	0.57	0.57	0.80	0.67	28.7	28.4	69493.1	3.2	3.0	0.2
OxyS-*fhlA*	0.80	0.82	0.64	0.64	0.61	0.61	0.48	0.48	0.61	0.61	20.6	19.3	129636.7	1.9	2.0	0.2
RyhB-*uof-fur*	0.13	0.13	0.12	0.00	0.21	0.21	0.19	0.00	0.21	0.21	26.4	25.3	65599.2	2.6	2.7	0.2
RyhB-*sodB*	0.67	0.71	0.70	0.68	0.65	0.51	0.65	0.59	0.65	0.63	15.4	15.2	23579.3	1.7	2.0	0.1
Average	0.62	0.62	0.49	0.46	0.58	0.52	0.47	0.41	0.57	0.53	22.8	22.1	72077.1	2.4	2.5	0.2

The MCC evaluates the joint structure, i.e. both the interaction between the two RNAs and the secondary structure of each single RNA. PETcofold was called with parameters $\delta = 0.9$, $\gamma = 0.1$, option *-noLP* and optionally with option *-extstem*. inteRNA, Pairfold and RactIP were called with default parameters. RNAcofold was called with options *-d2 -noLP*. The columns 80% and 100% give the result for the consensus structure with base pairs that occur in 80 and 100%, respectively, of the single structures. The run time of all single sequence-based approaches is the sum for all input sequences.

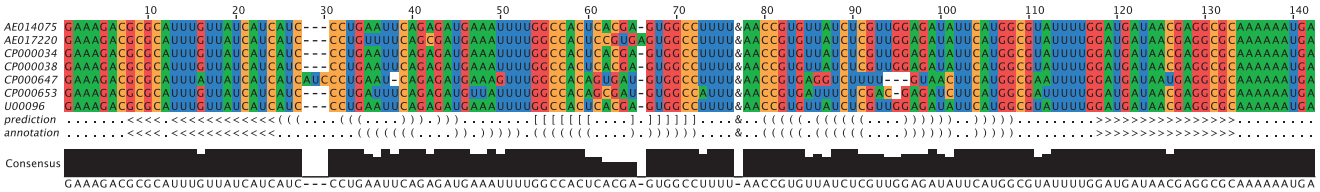

Fig. 4. Joint secondary structure of the sRNA MicA and the mRNA *ompA*. The alignment shows the two input alignments concatenated by the linker symbol '&', the joint structure predicted by PETcofold (with parameters $\delta = 0.9$, $\gamma = 0.1$ and option *-noLP*) and the model of the MicA–*ompA* complex proposed by Udekwu *et al.* (2005). Sequences are labelled with the genome accession numbers of the corresponding organisms. Angle brackets indicate *inter*-molecular base pairs. Round and square brackets indicate *intra*-molecular base pairs. Square brackets indicate positions that are constrained in step 1 of the PETcofold pipeline. For *ompA*, only 40 nt upstream of the interaction site are shown. The alignment was visualized with Jalview (Waterhouse *et al.*, 2009).

based on folding of the concatenated input sequences using the model of Zuker and Stiegler (1981). RactIP uses integer linear programming to maximize an objective function that is based on internal and external base pair probabilities. All computations were performed on a machine with AMD Opteron 2356 processor (2.3 GHz) and 16 GB RAM.

All resulting consensus structures were compared with the joint structure from the literature using the approximated MCC. Table 2 lists the MCCs of all four examples for PETcofold and the compared methods.

For the interactions of OxyS-*fhlA* and RyhB-*sodB*, the MCC of the PETcofold predictions was slightly higher when using the option *-extstem*. The opposite applies to MicA-*ompA*. On the non-curated alignments, the MCC was up to 0.2 lower (data not shown), which emphasized the importance of high-quality input alignments for our method. However, the option *-extstem* seemed to improve the prediction when using low-quality input alignments by extension of imperfectly (structurally) conserved stems.

When comparing PETcofold with the other methods, our method overall showed a better performance in predicting the joint structures. The prediction quality of the RyhB-*uof-fur* joint structure is very low for all compared approaches (maximal MCC of 0.21), which implies that the published interaction model is not predictable with the evaluated computational approaches. Thus, the prediction for this example is not reliable. When excluding RyhB-*uof-fur*, our approach gives consistently more reliable predictions than the single sequence-based methods (see Table 2.)

For comparison to the complex joint secondary structure prediction methods with high resource consumption (both high time and memory complexity), we were only able to compare to inteRNA, as this is currently the only method with a sufficiently low resource consumption. However, the evaluation shows that this (minimum free energy structure) prediction approach is not very reliable without homology information. Thus, one has to resort to the more complex partition function approaches, which, however, have drastically larger time and memory requirements. For example, it was not possible to obtain predictions from RNArip (Huang *et al.*, 2010) with reasonable resources. In comparison, PETcofold gave reliable predictions within seconds, which makes it also fast enough for genome-scale applications.

Figure 4 shows the annotation and PETcofold prediction for MicA-*ompA* together with the sequence alignments used as input. PETcofold correctly predicted all interaction base pairs from the annotation. The interaction site is highly conserved in both RNAs and contains only one compensatory mutation for the pairing between alignment positions 16 and 126. The *intra*-molecular structures contain compensatory mutations, for instance, between alignment positions 60 and 65.

Alignments and joint structures of OxyS-*fhlA*, RyhB-*sodB* and RyhB-*uof-fur* are given in Supplementary Figure S3. Two

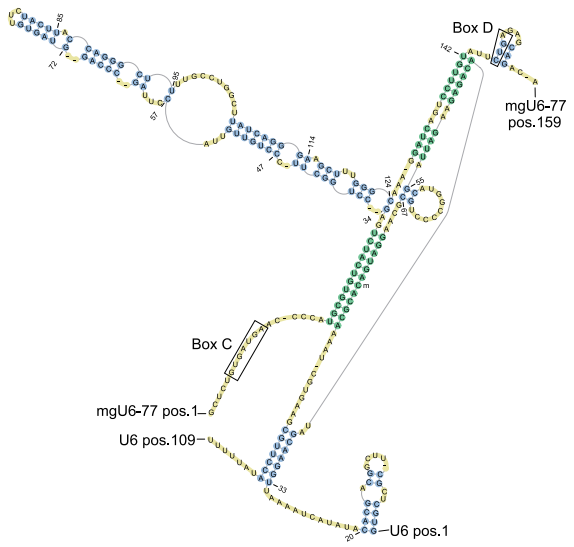

Fig. 5. Joint secondary structure of mgU6-77 snoRNA and *U6* snRNA as predicted by PETcofold using parameters $\delta = 0.9$, $\gamma = 0.1$ and options *-noLP* and *-extstem*. Mouse sequences are shown, but the numbering refers to the input alignment of human, chimp, mouse, cow, tenrec, dog and opossum. *Intra*-molecular base pairs are coloured in blue and *inter*-molecular base pairs in green. The cytosine at alignment position 78 (pos. 77 in mouse) is 2´-*O*-methylated and marked by 'm'. We used PseudoViewer3 for drawing (Byun and Han, 2009).

observations might explain the low prediction quality for RyhB-*uof-fur*. First, the interaction site given in the literature overlapped with the highly probable terminator stem of RyhB, which was constrained for *inter*-molecular folding. Second, a region of the interaction site that was subject to experimental validation (alignment positions 144 to 149) contained base pairs that were not supported by up to 3 out of 9 sequences of the alignment.

3.4 SnoRNA interactions in vertebrates

SnoRNAs have been experimentally detected in various vertebrates (e.g. Hüttenhofer *et al.*, 2001; Vitali *et al.*, 2003), but interactions between snoRNA and their target RNAs have only been predicted. The interactions are located in highly conserved regions for all examples of which we are aware. Despite of the lack of experimentally verified structures, we give one example showing that PETcofold is able to predict the joint structure of the known secondary structures and the predicted interaction.

Vertebrate spliceosomal snRNAs are highly modified by pseudouridylation and 2´-*O*-methylation. Tycowski *et al.* (1998) showed that depletion of the C/D box snoRNA mgU6-77 from *Xenopus* oocytes inhibits 2´-O-methylation of C77 in *U6* snRNA. Figure 5 shows the joint structure including a pseudoknot as predicted by PETcofold. The input consisted of the *mgU6-77* alignment from snoRNABase (Lestrade and Weber, 2006) and the *U6* seed alignment from Rfam. The joint structure consisting of the snoRNA structure, the interaction in mouse (Tycowski *et al.*, 1998) and the *U6* Rfam structure was predicted by PETcofold with an approximated MCC of 0.74. RNAcofold (parameters *-d2 -noLP*) performed with an approximated MCC of 0.51 for the consensus structure conserved in 80% as well as 100% of sequences. Both methods predicted further binding sites in addition to the reported highly conserved interaction, whereas PETcofold performed much better in predicting the long hairpin of the snoRNA (PETcofold MCC: 0.85, RNAcofold MCC: 0.0).

4 DISCUSSION

Here, we presented PETcofold, the first comparative method for the prediction of a joint secondary structure of two interacting RNAs. The method identifies evolutionary conserved structures and can exploit the information from compensating base changes in the *intra*-molecular structures of the two RNAs and the interactions between them. Furthermore, PETcofold allows for the prediction of pseudoknots between *intra*- and *inter*-molecular base pairs.

We have shown in controlled runs on simulated data that the covariance information improves the prediction ability for RNA–RNA interactions. We have also shown both for a vertebrate snoRNA and for four bacterial sRNAs that the addition of evolutionary information from multiple sequence alignments improves the performance in comparison to methods based on single sequences. This implies that single sequence-based methods could perform better if the comparative information is also taken into account. As for other RNA structure prediction methods making use of sequence-based alignments, it is well-documented that these work best with an average pairwise sequence identity above 60% (Gardner *et al.*, 2005; Washietl and Hofacker, 2004). In the process of cleaning up the (already sparse) bacterial data, we took this into account by removing more distant sequences to keep the balance of accurate assignment and covariance patterns of base pairs.

Many genomic screens for ncRNAs predict secondary structures with compensatory base pair changes on RNA alignments. These *intra*-molecular structures can be used as input of step 1 of the PETcofold pipeline to identify highly reliable substructures, which are then constrained for step 2. For example, PETcofold could be applied to predict RNA–RNA interactions on the CMfinder-generated structure-based alignments of *de novo* predicted RNA motifs in the ENCODE regions (Torarinsson *et al.*, 2008).

The sparse amount of known examples of sequences with RNA–RNA interactions and the paucity of covariations were the reasons why we introduced simulated data. Some of the known examples of RNA–RNA interactions, e.g. from bacterial sRNA–mRNA and eukaryotic miRNA–mRNA interactions, tend to be rather conserved. Even in cases with little or no compensating base pairs (in interaction sites or in the intra-molecular structure), any given variation will collectively contribute to the calculation of the reliabilities and thereby to the overall structure of the interaction complex. Hence, in the complete lack of covariation, PETcofold reduces to an improved energy folding approach, which also has impact on matching up the interacting base pairs. The sRNA–mRNA interactions studied here exhibit only a small number of compensatory base pair changes, which might be because sRNA sequences often show poor conservation across distant bacterial species. Thus, the regulators might be recently acquired and rapidly evolving (Waters and Storz, 2009). Nevertheless, many of the homologous ncRNAs and mRNAs have been found based on sequence similarity, which leads to highly identical sequences and thereby also to highly conserved interactions. PETcofold considers sequence conservation, but its full power is only revealed when the input data contains structural covariance. In the future, we expect that deep sequencing approaches will give rise to many more

characterized transcriptomes, which will increase the amount of data available for analysis including RNAs containing compensating base pair changes.

Certain RNA–RNA interactions, e.g. the functionally important interactions between the 16S and 23S rRNAs in the ribosome, involve non-canonical base pairs. Our method looks primarily for canonical, e.g. Watson–Crick, base pairs and G-U wobble base pairs. However, the Pfold model includes equilibrium distributions for the frequencies of all possible 16 bp and substitution rates for all possible base pair substitutions (16×16 matrix), which have been estimated from given trusted alignments of tRNA and rRNAs including non-canonical base pairs (Knudsen and Hein, 1999). The probabilities of non-canonical base pairs are low compared with Watson–Crick and Wobble base pairs and, thus, in practice, non-canonical base pairs are only found together with canonical base pairs. Nevertheless, the equilibrium distributions in the evolutionary model of PETcofold could be adapted to increase the impact of non-canonical base pairs by using a different training set.

Future improvements can include explicit handling of redundant sequence information. Currently, redundant sequences contribute equally to the scoring to directly reflect the input data instead of overweighting outliers in a dense evolutionary tree. Thus, datasets with highly redundant sequences should be cleaned prior to usage of the program.

Long RNA–RNA interactions form helices similar to the DNA double strand (Watson and Crick, 1953). These helices pose *inter*- and *intra*-molecular topological challenges. Due to the turn of the helix (i) the length of the interaction between the two RNAs is constrained and (ii) a certain number of unpaired bases on either side of the interaction are necessary to enable the first enclosing *intra*-molecular base pair (Pervouchine, 2004). In practice, this means that the computational methods can predict longer helices, because the outer bases are complementary, but in the three-dimensional topology these base pairings would be spatially unfeasible. However, taking topological constraints into account would be excessively time costly. Our method, and all other methods capable of genome-scale analysis, work only at the level of secondary structures. However, in the future it would be useful to accommodate this.

In some cases, the hierarchical folding approach predicts RNA–RNA interactions with reduced accuracy because the *intra*-molecular constraints overlap with the interaction sites. In these cases, the interaction site accessibility model of PETcofold is too strict and overestimates the stability of *intra*-molecular structures. As an example, one of the two OxyS–*fhlA* interaction sites is not predicted because the stem enclosing the second interaction site has a high reliability and, thus, gets constrained. In Seemann *et al.* (2010), the input alignment consisted of more distant sequences, which reduced the reliability of this stem and made the interaction site accessible. For avoiding too restrictive constraints, we introduced a parameter γ when reliable alternative structures exist. It is planned that a future version of PETcofold will take the cost of opening stems into account as, e.g. in RNAup and IntaRNA.

ACKNOWLEDGEMENTS

We thank Jakob Hull Havgaard for his assistance with the evaluation of the predictions. We further thank Sita Lange and Larry Croft for their comments on the manuscript.

Funding: Lundbeck Foundation (grant 374/06 to J.G.); Danish Research Council for Technology and Production Sciences (grant 274-09-0282 to J.G.); Danish Center for Scientific Computation (J.G.); German Federal Ministry of Education and Research (BMBF grant 0313921 to R.B.); German Research Foundation (DFG grants BA 2168/2-2 and BA 2168/4-1 to R.B.); Excellence Initiative of the German Federal and State Governments (grant EXC 294 to R.B.); Austrian GEN-AU project Bioinformatics Integration Network III (T.G. and Arndt von Haeseler); Wiener Wissenschafts-, Forschungs- and Technologiefonds (WWTF) to the CIBIV Institute (Arndt von Haeseler).

Conflict of Interest: none declared.

REFERENCES

Alkan,C. *et al.* (2006) RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, **13**, 267–282.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andronescu,M. *et al.* (2005) Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, **345**, 987–1001.

Argaman,L. and Altuvia,S. (2000) *fhlA* repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J. Mol. Biol.*, **300**, 1101–1112.

Bachellerie,J.P. *et al.* (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.

Backofen,R. and Hess,W.R. (2010) Computational prediction of sRNAs and their targets in bacteria. *RNA Biol.*, **7**, 33–42.

Bernhart,S.H. *et al.* (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3.

Bompfünewerer,A.F. *et al.* (2008) Variations on RNA folding and alignment: lessons from Benasque. *J. Math. Biol.*, **56**, 129–144.

Brunel,C. *et al.* (2002) RNA loop-loop interactions as dynamic functional motifs. *Biochimie*, **84**, 925–944.

Busch,A. *et al.* (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.

Byun,Y. and Han,K. (2009) Pseudoviewer3: generating planar drawings of large-scale rna structures with pseudoknots. *Bioinformatics*, **25**, 1435–1437.

Chitsaz,H. *et al.* (2009a) biRNA: fast RNA-RNA binding sites prediction. In Salzberg,S. and Warnow,T. (eds). *Proceedings of the 9th Workshop on Algorithms in Bioinformatics (WABI)*, Vol. 5724 of *Lecture Notes in Computer Science*, pp. 25–36. Springer, Berlin/Heidelberg.

Chitsaz,H. *et al.* (2009b) A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, **25**, i365–i373.

Dirks,R.M. *et al.* (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, **49**, 65–88.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

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

Gardner,P.P. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.

Gaspin,C. and Westhof,E. (1995) An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *J. Mol. Biol.*, **254**, 163–174.

Geissmann,T.A. and Touati,D. (2004) Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J.*, **23**, 396–405.

Gesell,T. and von Haeseler,A. (2006) In silico sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics*, **22**, 716–722.

Gorodkin,J. *et al.* (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res.*, **29**, 2135–2144.

Gorodkin,J. *et al.* (2010) De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol.*, **28**, 9–19.

Hertel,J. *et al.* (2009) Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Res.*, **37**, 1602–1615.

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

Horler,R.S.P. and Vanderpool,C.K. (2009) Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Res.*, **37**, 5465–5476.

Huang,F.W.D. *et al.* (2009) Partition function and base pairing probabilities for RNA-RNA interaction prediction. *Bioinformatics*, **25**, 2646–2654.

Huang,F.W.D. *et al.* (2010) Target prediction and a statistical sampling algorithm for RNA-RNA interaction. *Bioinformatics*, **26**, 175–181.

Hüttenhofer,A. *et al.* (2001) Rnomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.

Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.

Kato,Y. *et al.* (2010) RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*, **26**, i460–i466.

Knudsen,B. and Hein,J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.

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

Kolbe,D.L. and Eddy,S.R. (2009) Local RNA structure alignment with incomplete sequence. *Bioinformatics*, **25**, 1236–1243.

Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.

Li,L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem. Biophys. Acta*, **405**, 442–451.

Menzel,P. *et al.* (2009) The tedious task of finding homologous noncoding RNA genes. *RNA*, **15**, 2075–2082.

Mercer,T.R. *et al.* (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl Acad. Sci. USA*, **105**, 716–721.

Mückstein,U. *et al.* (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.

Mückstein,U. 0 (2008) Translational control by RNA-RNA interaction: improved computation of RNA-RNA binding thermodynamics. In Elloumi,M. *et al.* (eds), *Bioinformatics Research and Development*, volume 13 of *Communications in Computer and Information Science*, Springer, Berlin Heidelberg, pp. 114–127.

Pervouchine,D.D. (2004) IRIS: intermolecular RNA interaction search. *Genome Inform.*, **15**, 92–101.

Ravasi,T. *et al.* (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.*, **16**, 11–19.

Rehmsmeier,M. *et al.* (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.

Richter,A.S. *et al.* (2010) Seed-based IntaRNA prediction combined with GFP-reporter system identifies mRNA targets of the small RNA Yfr1. *Bioinformatics*, **26**, 1–5.

Salari,R. *et al.* (2010a) Fast prediction of RNA-RNA interaction. *Algorithms Mol. Biol.*, **5**, 5.

Salari,R. *et al.* (2010b) Time and space efficient RNA-RNA interaction prediction via sparse folding. In Berger,B. (ed), *Proceedings of RECOMB 2010*, Vol. 6044 of *Lecture Notes in Computer Science.* Springer Berlin/ Heidelberg, pp. 473–490.

Seemann,S.E. *et al.* (2008) Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.*, **36**, 6355–6362.

Seemann,S.E. *et al.* (2010) Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions. *Algorithms Mol. Biol.*, **5**, 22.

Sharma,C.M. *et al.* (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, **464**, 250–255.

Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, **24**, 2657–2663.

Taft,R.J. *et al.* (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, **29**, 288–299.

The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Torarinsson,E. *et al.* (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.*, **16**, 885–889. (Erratum in: *Genome Res.* 2006, **16**, 1439).

Torarinsson,E. *et al.* (2008) Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions. *Genome Res.*, **18**, 242–251.

Tycowski,K.T. *et al.* (1998) Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol. Cell*, **2**, 629–638.

Udekwu,K.I. *et al.* (2005) Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev.*, **19**, 2355–2366.

Večerek,B. *et al.* (2007) Control of Fur synthesis by the non-coding RNA RyhB and iron-responsive decoding. *EMBO J.*, **26**, 965–975.

Vinh,L.S. and von Haeseler,A. (2004) IQPNNI: moving fast through tree space and stopping in time. *Mol. Biol. Evol.*, **21**, 1565–1571.

Vitali,P. *et al.* (2003) Identification of 13 novel human modification guide RNAs. *Nucleic Acids Res.*, **31**, 6543–6551.

Washietl,S. and Hofacker,I.L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, **342**, 19–30.

Washietl,S. *et al.* (2005) Genome-wide mapping of conserved RNA secondary structure structures predicts thousands of functional non-coding RNAs in human. *Nat. Biotechnol.*, **23**, 1383–1390.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Waters,L.S. and Storz,G. (2009) Regulatory RNAs in bacteria. *Cell*, **136**, 615–628.

Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids. a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–741.

Weinberg,Z. *et al.* (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res.*, **35**, 4809–4819.

Will,S. *et al.* (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Comput. Biol.*, **3**, e65.

Wilusz,J.E. *et al.* (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.*, **23**, 1494–1504.

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