BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 17 2012, pages 2242–2248
doi:10.1093/bioinformatics/bts369

Structural bioinformatics

Advance Access publicaion July 13, 2012

Bayesian sampling of evolutionarily conserved RNA secondary structures with pseudoknots

Gero Doose[1,2,3] and Dirk Metzler[1,*]

[1]Department of Biology, LMU Biocenter, Ludwig-Maximilians-Universität München, Großhaderner Str. 2, D-82152 Planegg-Martinsried, [2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig and [3]Transcriptome Bioinformatics group, LIFE—Leipzig Research Center for Civilization Diseases, University of Leipzig, Philipp-Rosenthal-Strasse 27, D-04107 Leipzig, Germany

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Today many non-coding RNAs are known to play an active role in various important biological processes. Since RNA's functionality is correlated with specific structural motifs that are often conserved in phylogenetically related molecules, computational prediction of RNA structure should ideally be based on a set of homologous primary structures. But many available RNA secondary structure prediction programs that use sequence alignments do not consider pseudoknots or their estimations consist on a single structure without information on uncertainty.

Results: In this article we present a method that takes advantage of the evolutionary history of a group of aligned RNA sequences for sampling consensus secondary structures, including pseudoknots, according to their approximate posterior probability. We investigate the benefit of using evolutionary history and demonstrate the competitiveness of our method compared with similar methods based on RNase P RNA sequences and simulated data.

Availability: PhyloQFold, a C++ implementation of our method, is freely available from http://evol.bio.lmu.de/_statgen/software/phyloqfold/

Contact: gero@bioinf.uni-leipzig.de, metzler@bio.lmu.de

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

Received on February 8, 2012; revised on June 18, 2012; accepted on June 26, 2012

1 INTRODUCTION

Since the first discovery of catalytic RNA activity in the early 1980s a great number of structural and enzymatic RNAs have been found, which are involved in a large variety of tasks concerning highly topical fields of research including gene regulation (Roth and Breaker, 2009), tumor biology (Poliseno *et al.*, 2010) and epigenetic phenomena (Bonasio *et al.*, 2010). A fundamental condition for the biological function of RNA is the ability to adopt complex tertiary structures. These molecule comformations are encoded to some extent by the sequence of the molecule's nucleotides, i.e. the primary structure (Onoa and Tinoco, 2004). Since structure determination by experimental techniques like X-ray crystallography and NMR spectroscopy are time consuming and costly (Fürtig *et al.*, 2003), researchers are interested in computational methods to predict the molecule's conformation based on its primary structure.

In cases when several similar functional RNA sequences of evolutionarily related organisms are available, their structure can be inferred by comparative sequence analysis. Since the biological function of ncRNAs is determined by their molecular structure, RNA sequence evolution is constrained by the secondary structure. Changes in the molecules primary structure have less influence on selection pressure as long as the secondary and tertiary structure is maintained. This is the case when compensatory mutations conserve complementarity between base pairings (Durbin *et al.*, 1999; Wheeler and Honeycutt, 1988). Thus, in a multiple sequence alignment where structurally similar parts are grouped together, pairs of columns can be found that provide these co-varying signals. Comparative structure prediction methods use information like this to predict the structure that a group of related RNA sequences share (i.e. the *consensus* secondary structure). If a structurally correct or at least a high quality alignment is given, methods of this category, e.g. Pfold (Knudsen and Hein, 2003) and RNAalifold (Hofacker *et al.*, 2002), achieve very good results (Gardner and Giegerich, 2004). (Knudsen and Hein, 2003) describe the algorithms implemented in Pfold in terms of Stochastic Context-Free Grammars (SCFGs), which generate secondary structures with nested base pairings. This means that if position i is paired with position $i' > i$, and j is paired with $j' > j$, then either $i < j < j' < i'$ or $j < i < i' < j'$ must hold. Also many other RNA secondary structure reconstruction methods, including RNAalifold, focus on secondary structures that fulfill this assumption.

An remaining problem is given by structural elements called *pseudoknots*, which contain non-nested base pairings. Even if they occur rather rarely, pseudoknotted RNAs are known to be essential for a large variety of functional processes in the cell (Staple and Butcher, 2005). Several methods have been proposed to reconstruct secondary structures with pseudoknots (see, for example, Chen *et al.*, 2008; Reeder and Giegerich, 2004; Rivas and Eddy, 1999; Ruan *et al.*, 2004). The program IPKnot of (Sato *et al.*, 2011) reconstructs pseudoknotted consensus structures from homologous RNA sequences. Benchmarks like CompaRNA (Puton *et al.*, 2011) reveal that we cannot expect to find for every primary structure the correct secondary structure with any of the available prediction approaches, even if pseudoknots are neglected. For this reason, it is important to assess the reliability of a prediction. One possibility is to use Bayesian sampling, see (Eddy, 2004) for a primer.

*To whom correspondence should be addressed.

Following this strategy, (Metzler and Nebel, 2008) presented the RNA secondary structure prediction program McQFold. Instead of computing only the most likely folding of a given RNA primary structure, it samples possible secondary structures according to their approximate posterior probability, which makes the uncertainty of the predictions explicitly assessable. They developed an extended grammar model, permitting pseudoknots of arbitrary type, and a Metropolis Hastings strategy for a Markov chain Monte Carlo (MCMC) method to sample possible pseudoknot configurations. In the program SimulFold (Meyer and Miklós, 2007) implemented an MCMC method for the simultaneous Bayesian sampling of multiple sequence alignments, phylogenies and RNA secondary structures. The latter are first sampled without pseudoknots, but the consensus structures that are subsequently computed from the samples are so-called bi-secondary structures, which can contain a certain type of pseudoknots.

Here we present PhyloQFold, a new version of McQFold that makes use of the evolutionary history of a set of homologous RNA sequences. PhyloQFold is able to accept a multiple sequence alignment and a phylogenetic tree as an input and benefits from comparative sequence analysis. We suggest an extended model and show how the information that is contained in the alignment is weighted by the phylogeny of the sequences and how this information is integrated into the algorithms. We exemplify the potential of our method and investigate the improvement introduced by the new functionality. Furthermore, we compare the performance of our method to that of similar approaches on RNase P RNA sequence alignments and on data simulated according to our model.

2 MODEL

To define a prior probability distribution on the set of folded alignments we combine the SCFG model of (Knudsen and Hein, 2003), which emits alignment columns, with the capability of the model underlying McQFold to generate pseudoknots (Metzler and Nebel, 2008). This means, we assume that the first step to generate a set of aligned RNA sequences is to generates a secondary structure, which may contain pseudoknots, according to the extended SCFG model of (Metzler and Nebel, 2008), which is based on the grammar of (Knudsen and Hein, 1999) (see online Supplementary Material 1). Compared with the SCFG of (Knudsen and Hein, 2003), our grammar has an additional terminal symbol q, which can generate pseudoknots after the SCFG has generated the sequence of terminal symbols. For this, the q symbols occuring in the SCFG output are randomly arranged into pairs that generate stem regions, so-called 'q-stems'. The q symbols are then replaced by the corresponding sequence sections that are matched in the q-stem. The q-stems are usually not nested with the paired regions ('stems') generated by the SCFG.

After the grammar has generated the secondary structure, positions of the structure that are not paired emit unpaired alignment columns, whereas position pairs involved in stems emit pairs of alignment columns. For a given alignment, let $p_T(i)$ be the probability of the i-th column, given that it is emitted from an unpaired position, and let $p_T(j,k)$ be the probability of columns j and k, given that they are emitted as a pair of columns. These probabilities depend on the given phylogenetic tree T along which the sequences have evolved. For the emission probability $p_T(i)$ of unpaired alignment columns we assume that all unpaired sites evolve independently of each other along the phylogeny. Given a substitution rate matrix for loop regions, we compute $p_T(i)$ by Felsenstein's pruning algorithm (Felsenstein, 1981). We account for the stochastically dependent evolution of alignment columns pairs by modelling the substitution process of doublets (Schöniger and von Haeseler, 1994), which in our case are nucleotides at opposite positions of a stem. Thus, the underlying transition dynamic $P_{f\to h}(t)$ defines the probability of the observed change of doublet $f \in \{a,g,c,u\}^2$ into doublet $h \in \{a,g,c,u\}^2$ after time t. We compute the emission probabilities $p_T(j,k)$ of paired alignment columns with Felsenstein's pruning algorithm, in which we use the 16×16 substitution model of (Knudsen and Hein, 1999).

3 ALGORITHMS AND IMPLEMENTATION

The general flow of our new program PhyloQFold is similar to that of McQFold (Metzler and Nebel, 2008). The program generates a set of candidates for the q-stems, i.e. segment pairs of the alignment which may be involved in pseudoknots. From this pool of candidates, configurations of q-stems are proposed in each step of the MCMC sampling process. To decide whether a q-stem configuration is accepted as the next state of the Markov chain, it is rated by the probability of the sequence alignment conditioned on the q-stem configuration, which is computed by the inside algorithm (Lari and Young, January 1990). After a burn-in phase of 100 Metropolis Hastings steps, every 50th state of the following 1000 steps is used to sample one consensus structure and to approximate the pairing probabilities (these default values can be changed by the user). Each structure is computed by the inside algorithm with randomized backtracking. For each possible pair of positions, the posterior pairing probabilities are calculated with a variant of the inside–outside algorithm and averaged over 20 sampled q-stem configurations (Lari and Young, January 1990; Metzler and Nebel, 2008). The search for the most probable folding is based on simulated annealing (Kirkpatrick *et al.*, 1983). In the first 1000 steps of this heuristic, the proposal chain is optimized using the inside algorithm to rate the states of the proposal chain, whereas in the second phase (also 1000 steps) the CYK algorithm (Cocke, 1969; Kasami, 1965; Younger, 1967) is used to rate each state. Finally, the most probable parse tree together with the q-stem configuration of the last state of the Markov chain is assumed as the most probable consensus secondary structure of the alignment.

In PhyloQFold, the SCFG algorithms (inside, outside, CYK) use the emission probabilities $p_T(i)$ and $p_T(i,j)$ of alignment columns and of pairs of alignment columns. Thus, PhylQFold uses, like PFold, the phylogenetic information contained in $p_T(.)$ and $p_T(.,.)$. This is the main difference to McQFold, which uses only emission probabilities of single nucleotides and pairs of nucleotides.

The runtime of the SCFG algorithms is cubic in the length ℓ of the input alignment. The number n of sequences only affects the runtime of the Felsenstein pruning algorithm, which is linear in n and is performed only once for each of the ℓ alignment columns and ℓ^2 pairs of alignment columns, see also online Supplementary Material 4.1 and Table 1.

The current C++ implementation of PhyloQFold takes as input a multiple sequence alignment in FASTA format and a phylogeny in rooted Newick tree format. The program is based on the prerequisite that each header line of the FASTA file corresponds to one leaf label

of the Newick tree file. The output is given by three result files. One file consists of 20 sampled consensus secondary structures, another file contains the pairing probabilities for all position pairs and the third file contains the most probable consensus folding.

3.1 Detecting q-stem candidates

For a given alignment, the pool of q-stem candidates is a set of interval pairs $\xi = \{(i, i+1, \ldots, i+k), (j-k, \ldots, j-1, j)\}$. To decide during the MCMC procedure which ξ are proposed to be involved in a pseudokont, we apply two criteria. First, a q-stem candidate ξ should provide complementary nucleotides in the corresponding alignment columns. Second, ξ are preferred if they have a small probability to form a regular non-pseudoknot stem.

To find segment pairs ξ that comply with the first criterion, we search for *high-scoring segment-pairs* (HSP; Altschul *et al.*, 1990) between the input alignment and its complement, which means that the order of positions is inverted and nucleotides match if they are complementary. As alignment score for a pair of alignment columns, we use a log-scaled likelihood ratio, here defined as the ratio of the probability $p_T(i,j)$ of the two columns i and j, under the assumption that they are paired, to the probability $p_T(i) \cdot p_T(j)$ of the two columns, under the assumption that they are independent, where both probabilities are conditioned on the phylogenetic tree T. We define the score σ_ξ of a segment pair ξ as given above by

$$\sigma_\xi = \sum_{\ell=0}^{k} \ln\left(\frac{p_T(i+k, j-k)}{p_T(i+k) \cdot p_T(j-k)}\right).$$

To test if a detected ξ is in accordance with the second criterion, the SCFG inside–outside algorithm is applied and both criteria are used to rate q-stem candidates as described by (Metzler and Nebel, 2008).

3.2 Treatment of gaps

In the current implementation, gaps are treated as missing data, which means that the likelihood of a tree, given one column with gaps, is the same as the likelihood of the pruned tree, given the pruned column. This strategy is straightforward for the calculation of the probabilities $p_T(i)$ of unpaired alignment columns, but during the calculation of the probability $p_T(i,j)$ of column pairs situations may occur in which only one of the two doublet positions is a gap symbol. In such cases the corresponding branch is not pruned, but the remaining doublets that are still possible are assumed to have a probability of 1. This procedure can cause high probabilities for column pairs with many gaps, resulting in a prediction where gaps are paired with nucleotides. To avoid this, we penalize gaps by multiplying $p_T(i,j)$ with a factor $(1 - \epsilon \cdot \mu / n) < 1$, where n is the number of aligned sequences, μ is the number of column positions at which one of the two columns provides a gap symbol, and the parameter ϵ can be chosen by the user. In the analyses reported below we used $\epsilon = 0.7$. Simulations indicate, however, that changing ϵ to a different value between 0.5 and 1.0 has little effect on the performance of our method (data not shown).

4 RESULTS AND DISCUSSION

Though PhyloQFold is primarily designed for sampling secondary structures in accordance with their posterior probability, the following evaluation relies on the most probable folding that the heuristic optimization process attempts to find. Thereby a comparison to other approaches, which mostly produce one optimized folding, became possible. The production rules of the extended SCFG are those as given by (Metzler and Nebel, 2008), see also online Supplementary Material 1, while the rate matrices of the evolutionary model are those of (Knudsen and Hein, 1999).

Although there does not exist a perfect method to describe a confusion matrix of a classifier by a single value, a measure well established in bioinformatics is *Matthews correlation coefficient* (MCC; Matthews, 1975). It can be approximated by the geometric mean of sensitivity and selectivity (Gorodkin *et al.*, 2001): $\mathrm{MCC} \approx \sqrt{\frac{\mathrm{TP}}{\mathrm{TP+FN}} \cdot \frac{\mathrm{TP}}{\mathrm{TP+FP}}}$. Here, sensitivity is the fraction of predicted position pairs among all pairs of positions given in the reference folding, whereas selectivity is the fraction of correctly predicted position pairs among all pairs of positions that are predicted to be paired with each other. To obtain these frequencies, we compare the predicted consensus structure to the correct consensus structure, neglecting all alignment positions that contain gaps (see online Supplementary Material 2).

4.1 Example for Illustration: RNase P RNA

The RNase P Database by (Brown, 1999) provides a high quality structural alignment of archaeal RNase P RNAs, which is examined by extensive comparative analyses (Harris *et al.*, 2001). We extracted the structural alignment of the four sequences of *Acidianus ambivalens* (Sequence 1), *Acidianus brierleyi* (Sequence 2), *Sulfolobus shibatae* (Sequence 3) and *Sulfolobus solfataricus* (Sequence 4). We estimated the corresponding phylogenetic tree by a maximum-likelihood (ML) approach applying a standard 4×4 DNA substitution model. Results of applying PhyloQFold to an alignment of all 41 available archaeal RNase P RNA sequences are shown in online Supplementary Material 4.1.

Bayesian Sampling: The 20 consensus secondary structures produced by the Bayesian Sampling procedure of PhyloQFold indicate that many different consensus structures are possible for this set of molecules (see online Supplementary Fig. S4).

Posterior pairing probabilities: Figure 1 shows that all of the column pairs predicted as highly probable $(0.8-1)$ are indeed also given in the database structure, while most of the stems predicted as unlikely $(0.05-0.4)$ are not existent.

Most probable folding: the predicted consensus structure obtained by the simulated annealing procedure together with the correct structure were put into graphs with the help of the program *VARNA* (Darty *et al.*, 2009). Figure 2 reveals a strong similarity between both foldings, where most of the base pairs were correctly predicted. In fact all of the predicted stems are also present in the reference structure including the stem (positions 20–25 and 220–225) that is responsible for the pseudoknot. Yet some of the base pairs of the correct structure were not predicted. Since the applied grammar model only allows for stems consisting of at least three base pairs, the stems of length two in the reference structure are not provided by the predicted structure. The performance evaluation of this prediction led to an MCC value of 0.905.

4.2 Benefit of accounting for the evolutionary history

As an example for the benefit of the extension presented in this work, the influence of phylogenetic information on the prediction

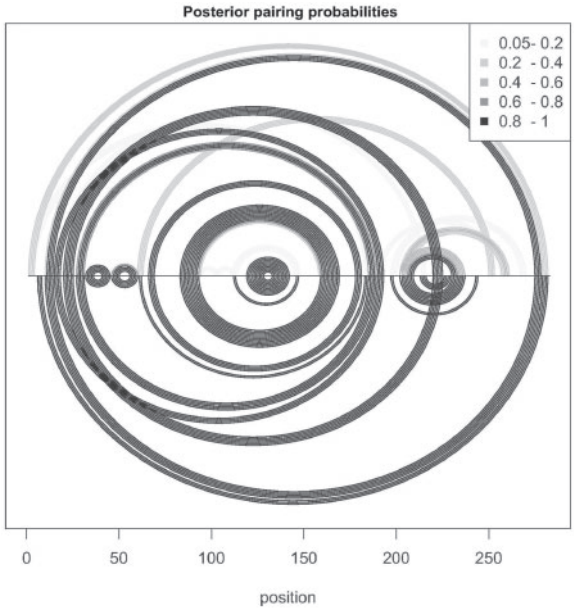

Fig. 1. Alignment positions that are paired with each other in the consensus secondary structure are connected by arcs. The upper half displays the grey tones code for the posterior pairing probabilities estimated by PhyloQFold's MCMC procedure. The lower half shows the structure given in the RNase P Database

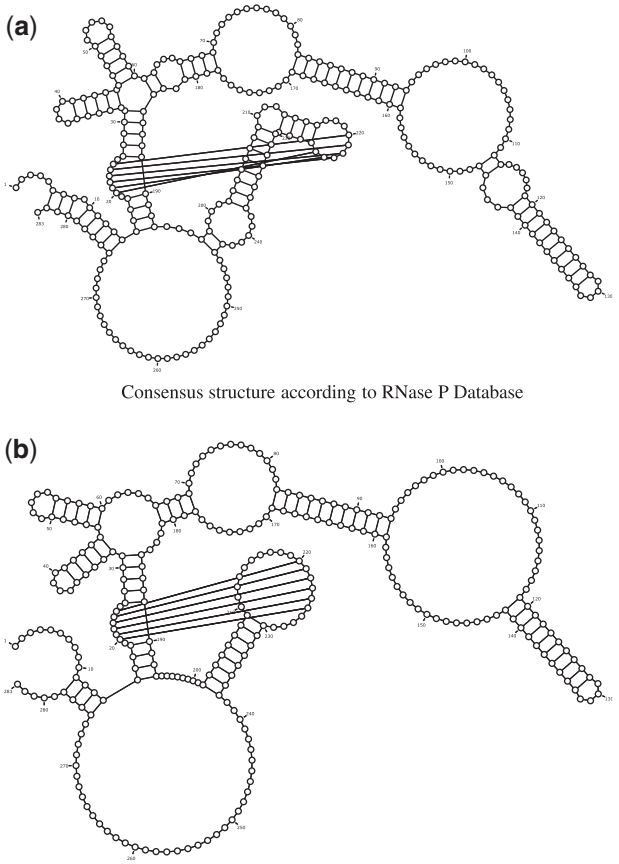

Fig. 2. Graph comparison of the structure according to the database and the predicted consensus structure found by the simulated annealing procedure of PhyloQFold

quality was evaluated with the same sequences as in Section 4.1. The single lengths of the four sequences range between 259 and 267 nucleotides and their pairwise sequence identities range from 56.3% to 94.6%.

We extracted the structural alignments from the RNase P Database. With the present analysis we want to treat the case that a reliable phylogeny is given, but in fact we had to reconstruct the phylogeny from the given data. To keep the risk of using an erroneous phylogeny as small as possible, we took the structural information given in RNase P Database into account in the phylogeny reconstruction. For this we used the software *MrBayes* (version 3.1.2; Huelsenbeck *et al.*, 2001), which allowed us to apply the standard 4-state nucleotide substitution model for the loop regions and a 16-state doublet model for the stem regions. Corresponding to the *F81* model (Felsenstein, 1981) each partition model assumes the same rate for all of its substitutions. Convergence of the MCMC processes was examined with the program *tracer* (Rambaut A, 2007). Consensus trees were derived from sampled phylogenies, describing the evolutionary relationship of the subsets of the RNase P RNA molecules. This dataset made it possible to investigate the performance, provided that the correct alignment and phylogeny are known.

The prediction results for this scenario are given in Table 1. We observed a performance improvement, which correlates with the amount of employed sequences. The first row shows the results examined by McQFold, whereas the other rows show the results determined by PhyloQFold. The performances, achieved by using only single sequences, differ between MCC values of 0.599 and 0.786 with arithmetic mean 0.707. Predictions based on two sequences achieved an arithmetic mean of 0.809, while the range is rather large with 0.226. In this category the lowest prediction result is 0.699, which is still 33.22% superior to the worst single-sequence prediction. When three sequences were used, the lowest prediction result is given by a MCC value of 0.778, which is nearly as good as the best single-sequence prediction. The arithmetic mean is 0.848. The performance evaluation of the prediction based on all four sequences led to a MCC value of 0.905, which is an improvement of 28% compared with the arithmetic mean of the single-sequence prediction performances.

Table 1. MCCs for McQFold and PhyloQFold results for test set with given structural alignments

Number of sequences	MCC		
	Min	Max	Mean
1	0.599	0.786	0.707
2	0.699	0.925	0.809
3	0.778	0.932	0.848
4	0.905	0.905	0.905

4.3 Performance comparison

We compare PhyloQFold to other methods for inferring a consensus secondary structure from an alignment of RNA sequences. In the

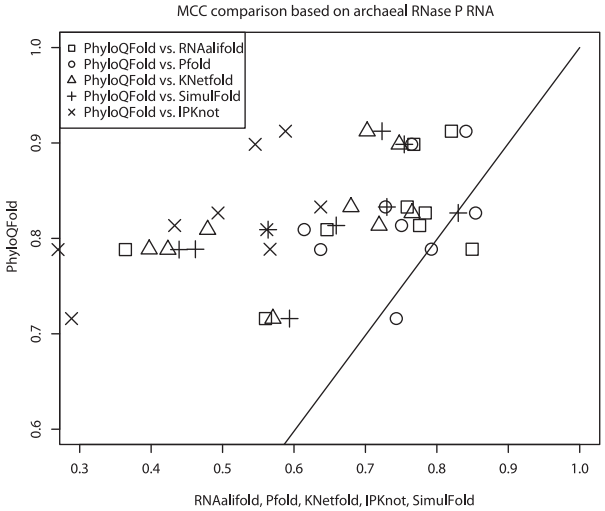

Fig. 3. MCC values reached by RNAalifold, Pfold, KNetfold, SimulFold and IPKnot, each compared with the MCC value that PhyloQFold achieved for the same test instance. Points below the $X = Y$ line represent test instances for which PhyloQFold achieved poorer prediction quality, whereas points above this line indicate greater prediction quality reached by PhyloQFold

field of approaches following this strategy, RNAalifold (version with RIBOSUM scoring; Bernhart *et al.*, 2008; Hofacker *et al.*, 2002) and Pfold (Knudsen and Hein, 2003) are widely used programs. The program Pfold applies a SCFG model for secondary structures combined with an evolutionary model of RNA sequences, whereas RNAalifold implements a variant of the well-established dynamic-programming algorithm from (Zuker and Stiegler, 1981). As these two programs are not able to deal with pseudoknots we also chose KNetfold (Bindewald and Shapiro, 2006) because it can also predict base pairs that violate the nested structure of other pairings. It uses k-nearest neighbor classifiers, mutual information and the fraction of complementary nucleotides of column pairs in a machine learning framework. The newest program in our comparison is IPKnot (version 1.2.1; Sato *et al.*, 2011). Moreover, we included SimulFold (Meyer and Miklós, 2007) because, like PhyloQFold, it follows a Bayesian MCMC approach.

This comparison is based on the archaeal structure alignment from the RNase P Database by Brown (1999). The available 41 sequences were divided into 10 disjunct groups of size 4, while 1 sequence remained unaffected. For each group the corresponding structural alignment was extracted. Nine of these structural alignments form the test set, while one of them was not used, as it possesses 677 columns and is therefore too large for the current version of Pfold. Since PhyloQFold requires in addition a phylogenetic tree as input data, for each of these nine alignments the program RAxML was employed to estimate the phylogeny (Stamatakis *et al.*, 2008). In SimulFold we chose the options -S -T to sample structures together with phylogenies while keeping the given alignment fixed. Up to this exception, all programs were applied with their default parameter settings. It should be noted that all the caveats mentioned by (Gardner and Giegerich, 2004) also hold for this comparison.

Figure 3 displays a direct performance comparison between PhyloQFold and the other programs for each test instance. In 40 out of the 45 comparisons PhyloQFold reached a higher prediction quality. Pfold supports for three test instances slightly better MCC values than those reached by PhyloQFold, whereas RNAalifold and SimulFold achieved for only one test instance a better MCC value. For all nine test instances KNetfold's and IPKnot's predictions are poorer compared with those of PhyloQFold.

We also compared the performance of the different programs on alignments of RNAs whose secondary structures (according to the public databases) do not contain pseudoknots. For 35 alignments of nuclear yeast RNA sequences from the RNase P database, PhyloQFold achieved the best average MCC values. On four RNA P and tRNA alignments that had been used in the study of (Gardner and Giegerich, 2004) only PFold and RNAalifold, which both assume the absence of pseudoknots, achieved better MCMC values than PhyloQFold, and for seven alignments of SPR RNA sequences (Rosenblad *et al.*, 2009) only RNAalifold and KNetfold showed a better average performance than PhyloQFold. Details on these comparisons are given in Section 5.2 of the online Supplementary Material.

4.4 Simulation study

We performed a simulation study to investigate the qualification of our model and the applied model parameters. By using the same model parameters for data generation as for the analysis of the prediction results, an impression of the performance potential of our method was provided. Furthermore, we applied the same programs used in Section 4.3 on the simulated data and compared their prediction results.

For this purpose we simulated 50-folded alignments based on 6 arbitrary trees (see online Supplementary Fig. S15). Although each phylogeny compasses five sequences, they vary in their pairwise taxa distances from 0.006 to 1.8 time units. The length of the simulated alignments differ between 68 and 410 columns. Since the applied substitution models do not cover insertion–deletion modelling, the alignments were generated devoid of gap symbols. To take the ability of handling gaps into consideration, gaps were included in a post-simulation process where different probabilities (0.1 and 0.02) were applied to loop and stem regions in order to change each nucleotide into a gap symbol. We employed PhyloQFold, RNAalifold, Pfold, KNetFold SimulFold and IPKnot to the simulated alignments and compared the quality of their predictions. Again, the phylogenetic trees for PhyloQFold were estimated with RAxML and SimulFold was allowed to sample the phylogeny together with the secondary structures. We also applied PhyloQFold and SimulFold with the phylogeny fixed to the true tree used in the sequence simulation, and the results were only marginally different from those obtained with estimated trees (data not shown).

Figure 4 shows for PhyloQFold, RNAalifold, PFold, KNetFold, SimulFold and IPKnot the MCC distribution based on simulated data. PhyloQFold reached the highest median (0.739) which is 11.8–28.7% greater than the medians of the other programs and showed the least standard deviation in this value (0.112). The mean MCC value of PhyloQFold (0.712) is significantly higher than those of RNAalifold (0.603), PFold (0.624), KNetFold (0.562), SimulFold (0.6) and IPKnot (0.528) (all p-values below 10^{-5}, one-tailed paired t-tests with df $= 49$).

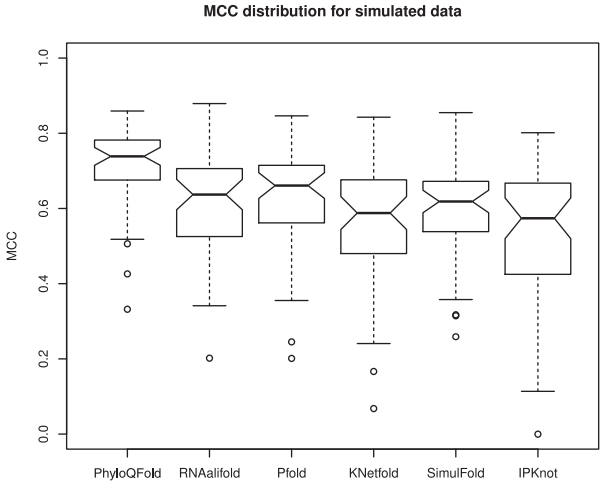

Fig. 4. MCC distributions of the six programs. The predictions were ascertained on a test set, which was simulated according to our model. All programs were used with their default parameter settings

5 CONCLUSION AND OUTLOOK

We have presented a method to sample consensus RNA secondary structures with pseudoknots according to their posterior probability. PhyloQFold, the implementation of our method, is able to incorporate the information given by the evolutionary history of a set of homologous RNA primary structures. We combined an extended SCFG with an explicit evolutionary model to define the prior probability distribution of structures.

The results of the performance comparison for RNase P RNA data are very encouraging and suggest that PhyloQFold is indeed competitive with widely used approaches. But even MCC values around 0.9, which we observed in some of our analyses, imply that there is still some uncertainty in the inferred secondary structures. On synthetic data PhyloQFold performed significantly better than the other programs, which may partly be due to the fact that we generated the folded alignments according to our model. The mean of PhyloQFold's MCC values (0.712) together with the standard deviation in these values (0.112) indicate the need to assess the uncertainty of each prediction. For this purpose, PhyloQFold's Bayesian Sampling procedure calculates posterior pairing probabilities which give vital information about the reliability of the structure predictions.

In Section 4.1 we have shown examples where PhyloQFold was capable of predicting pseudoknots (Fig. 2). This example has also shown, however, that the minimum length of stems of three, which is dictated by our grammar rules, can be obstructive. This issue could be avoided by using a refined grammar model. Moreover, as all our results were obtained without any parameter adjustment, there may be some potential for improvement by fitting the grammar parameters to available data. For SCFG models it is possible to compute ML estimators for model parameters from RNA sequences of unknown secondary structure by combining expectation–maximization with inside–outside algorithms (Eddy and Durbin, 1994). For our grammar, ML estimation is more difficult even if trusted secondary structures are given. The reason for this is the ambiguity of our grammar. Each stem in a structure could either be a q-stem or a stem generated by the SCFG. One possibility to compute (and then numerically optimize) the likelihood of the grammar parameters would be to apply MCMC for sampling the contributions of many possible q-stem configurations. However, the fact that the grammar used in McQFold and PhyloQFold has only four free parameters makes it possible to estimate the parameters from summary statistics, as for example the numbers and lengths of loops, stems and pseudoknots in trusted secondary structures. We provide an example of such a summary-statistic-based parameter estimation procedure using Approximate Bayesian Computation (Beaumont *et al.*, 2002; Csilléry *et al.*, 2010) on the PhylpQFold homepage. However, given the good performance of PhyloQFold on RNase P RNA data, it may surprise in the simulation study in Section 4.4 that PhyloQFold was only slightly better than the other programs even though it was the only program that 'knew' the model and parameters underlying the simulated data. This indicates that the benefit of fitting the grammar parameters to data may be limited.

As observed in Section 4.2, the information contained in the alignment and weighted by the phylogeny of the sequences seems to improve the quality of our structure estimations. But the alignment is of crucial importance and the dependency of the alignment quality implies the conceptual problem shared by all comparative approaches that are based on a fixed input alignment. In an application-orientated scenario, where no structural alignment is at hand, the use of structure-enhanced alignment programs like R-coffee (Wilm *et al.*, 2008) may be a good choice to estimate the input alignment. However, a more general solution to this problem would be to combine PhyloQFold with the Gibbs algorithm (Geman and Geman, 1984) of SimulFold (Meyer and Miklós, 2007) to sample from the joint distribution of alignments and secondary structures.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their comments, which helped us to substantially improve our manuscript.

Funding: This publication is supported by LIFE—Leipzig Research Center for Civilization Diseases, Universität Leipzig. This project was partly funded by means of the European Social Fund and the Free State of Saxony.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
Beaumont,M.A. *et al.* (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
Bernhart,S.H. *et al.* (2008) Rnaalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
Bindewald,E. and Shapiro,B.A. (2006) RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, **12**, 342–352.
Bonasio,R. *et al.* (2010) Molecular signals of epigenetic states. *Science*, **330**, 612–616.
Brown,J.W. (1999) The ribonuclease P database. *Nucleic Acids Res.*, **27**, 314.
Chen,X. *et al.* (2008) Flexstem: improving predictions of RNA secondary structures with pseudoknots by reducing the search space. *Bioinformatics*, **24**, 1994–2001.
Cocke,J. (1969) *Programming Languages and Their Compilers: Preliminary Notes*. Courant Institute of Mathematical Sciences, New York University.
Csilléry,K. *et al.* (2010) Approximate Bayesian computation (abc) in practice. *Trends Ecol. Evol.*, **25**, 410–418.
Darty,K. *et al.* (2009) Varna: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Durbin,R. *et al.* (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge UK.

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

Eddy,S.R. (2004) What is Bayesian statistics? *Nat. Biotechnol*., **22**, 1177–1178.

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

Fürtig,B. *et al.* (2003) NMR spectroscopy of RNA. *Chembiochem*, **4**, 936–962.

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

Geman,S. and Geman,D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, **6**, 721–741.

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

Harris,J.K. *et al.* (2001) New insight into Rnase P RNA structure from comparative analysis of the archaeal RNA. *RNA*, **7**, 220–232.

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

Huelsenbeck,J.P. *et al.* (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Kasami,T. (1965) An efficient recognition and syntax analysis algorithm for context-free languages. *Technical Report AFCRL-65-758*, Air Force Cambridge Research Laboratory, Bedford, MA.

Kirkpatrick,S. *et al.* (1983) Optimization by simulated annealing. *Science*, **220, 4598**, 671–680.

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.

Lari,K. and Young,S.J. (1990) The estimation of stochastic context-free grammars using the inside-outside algorithm. *Comput. Speech Lang*., **4**, 35–56.

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

Metzler,D. and Nebel,M.E. (2008) Predicting RNA secondary structures with pseudoknots by MCMC sampling. *J. Math. Biol*., **56**, 161–181.

Meyer,I.M. and Miklós,I. (2007) Simulfold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol*., **3**, e149.

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

Poliseno,L. *et al.* (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, **465**, 1033–1038.

Puton,T. *et al.* (2011) A Server for Continuous Benchmarking of Automated Methods for RNA Structure Prediction. http://comparna.amu.edu.pl/. (last accessed 15. December 2011)

Rambaut,A, Drummond,A. (2007) *Tracer v1.4*, http://beast.bio.ed.ac.uk/tracer. (last accessed 6. June 2012)

Reeder,J. and Giegerich,R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, **5**, 104.

Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol*., **285**, 2053–2068.

Rosenblad,M.A. *et al.* (2009) Kinship in the SRP RNA family. *RNA Biol*., **6**, 508–516.

Roth,A. and Breaker,R.R. (2009) The structural and functional diversity of metabolite-binding riboswitches. *Annu. Rev. Biochem*., **78**, 305–334.

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

Sato,K. *et al.* (2011) Ipknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**, i85–i93.

Schöniger,M. and von Haeseler,A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.*, **3**, 240–247.

Stamatakis,A. *et al.* (2008) A rapid bootstrap algorithm for the RaxML web servers. *Syst. Biol.*, **57**, 758–771.

Staple,D.W. and Butcher,S.E. (2005) Pseudoknots: RNA structures with diverse functions. *PLoS Biol*., **3**, e213.

Wheeler,W.C. and Honeycutt,R.L. (1988) Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implications. *Mol. Biol. Evol*., **5**, 90–96.

Wilm,A. *et al.* (2008) R-coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res*., **36**.

Younger,D.H. (1967) Recognition and parsing of context-free languages in time n3. *Inform. Control*, **10**, 189–208.

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