# Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments

Seemann, Ernst Stefan; Gorodkin, Jan; Backofen, Rolf

*Published in:*
Nucleic Acids Research

*DOI:*
10.1093/nar/gkn544

*Publication date:*
2008

*Document version*
Publisher's PDF, also known as Version of record

*Citation for published version (APA):*
Seemann, E. S., Gorodkin, J., & Backofen, R. (2008). Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Research*, *36*(20), 6355-6362. https://doi.org/10.1093/nar/gkn544

Published online 4 October 2008 Nucleic Acids Research, 2008, Vol. 36, No. 20 6355–6362
doi:10.1093/nar/gkn544

# Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments

**Stefan E. Seemann**[1], **Jan Gorodkin**[1] **and Rolf Backofen**[2,*]

[1]Division of Genetics and Bioinformatics, IBHV and Center for Applied Bioinformatics, University of Copenhagen, Groennegårdsvej 3, DK-1870 Frederiksberg C, Denmark and [2]Chair for Bioinformatics, Institute of Computer Science, Albert-Ludwigs-Universität, Georges-Koehler-Allee, Geb. 106, D-79110 Freiburg, Germany

Received June 16, 2008; Revised August 8, 2008; Accepted August 9, 2008

## ABSTRACT

**Computational methods for determining the secondary structure of RNA sequences from given alignments are currently either based on thermodynamic folding, compensatory base pair substitutions or both. However, there is currently no approach that combines both sources of information in a single optimization problem. Here, we present a model that formally integrates both the energy-based and evolution-based approaches to predict the folding of multiple aligned RNA sequences. We have implemented an extended version of Pfold that identifies base pairs that have high probabilities of being conserved *and* of being energetically favorable. The consensus structure is predicted using a maximum expected accuracy scoring scheme to smoothen the effect of incorrectly predicted base pairs. Parameter tuning revealed that the probability of base pairing has a higher impact on the RNA structure prediction than the corresponding probability of being single stranded. Furthermore, we found that structurally conserved RNA motifs are mostly supported by folding energies. Other problems (e.g. RNA-folding kinetics) may also benefit from employing the principles of the model we introduce. Our implementation, PETfold, was tested on a set of 46 well-curated Rfam families and its performance compared favorably to that of Pfold and RNAalifold.**

## INTRODUCTION

With the recent focus on nonprotein coding RNA (ncRNA) genes, interest in detecting novel ncRNAs has rapidly emerged. Since the structure of RNA is evolutionarily more conserved than its sequence, predicting the RNA's secondary structure is one of the most important steps towards its functional analysis. There are two fundamentally different approaches to predicting RNA secondary structures, namely free-energy minimization and probabilistic approaches, which often use phylogenetic information given by a multiple sequence alignment. This *duality* of approaches represents two different types of RNA structure information, the former relies on the physical properties of single sequences, while the latter uses evolutionary information in the form of compensatory base pair substitutions.

The advantage of energy minimization is that it relies on experimentally determined parameters. On the other hand, it is known that methods such as RNAfold (1) and Mfold (2) achieve an overall sensitivity of about 70% (3). For example, H/ACA snoRNAs usually do not produce the characteristic two-stem when the energy minimization alone is used to determine the structure. Instead, information on the functional sites have to be used as additional constraints to the thermodynamic folding (4–6). In general, there are several reasons for the inaccuracy of energy minimization. First, the energy model is incomplete, especially for multi-loop structures. Second, there can be alternative structures with similar free energies. And third, true structures are often stabilized by bound molecules. The only feasible way to determine the effects of these inaccuracies is to include phylogenetic information by looking for conserved substructures that cannot be accounted for by thermodynamic folding alone.

Hence, it is desirable to combine this duality of RNA structure information into a single optimization problem. This was probably addressed for the first time in 1985 by David Sankoff (7), who introduced an algorithm that solved the problem of simultaneously aligning and folding a set of unaligned RNA sequences. While, this is a kind of gold standard for comparative RNA secondary structure prediction, the 'algorithm requires extreme amounts of memory and time' (8). Thus, practical implementations of the Sankoff algorithm like FOLDALIGN (9–12), Dynalign (13,14), PMcomp (15), LocARNA (16,17) and PARTS (18), published more than 20 years later, introduce different constraints and somewhat arbitrary choices in their scoring schemes to make the approach tractable for realistic input sizes.

*To whom correspondence should be addressed. Tel: +49 761 203 7461; Fax: +49 761 203 7462; Email: backofen@informatik.uni-freiburg.de

Probabilistic approaches to the problem of simultaneously aligning and folding a set of RNA sequences, e.g. Stemloc (19) and Consan (20), are usually based on stochastic context-free grammars (SCFG). Unlike the aforementioned Sankoff-like methods where the energy is explicitly reflected in the scoring scheme, these approaches rely purely on statistical learning methods to determine their parameters. A mixed approach is employed by CMfinder (21) that implicitly combines energy contributions with an SCFG. As a seed CMfinder uses energetically folded structures from which a covariance model (SCFG) is constructed in successive rounds of optimization. Another approach is SimulFold (22), which simultaneously infers structures (including pseudoknots), alignments *and* trees. There are no sequence-dependent energy contributions, but an energy term that depends on the topology of the consensus structure.

The main disadvantage of the Sankoff-like approaches is their high-computational cost. For this reason, a class of methods was developed that saves computational resources by predicting the optimal structure from given RNA alignments, which are usually produced by multiple sequence alignment methods. This approach has proven useful in genomic screens for ncRNAs (23,24), despite their limitations in finding RNA structures in more divergent sequences (25). It has been shown that the quality of the predictions breaks down in cases where sequence identity is <60% (26). On the other hand, these methods can also be applied to improve the consensus structure prediction in Sankoff-like approaches. The reason is simply that the Sankoff-like approaches usually apply a progressive strategy by combining pairwise alignments to build the final multiple alignment. Thus, consensus structure prediction can be improved when considering the complete phylogenetic information.

In this article, we extend Pfold (27) to simultaneously use evolutionary and energetic information while searching for the common structure in a set of prealigned sequences. The probabilistic approach underlying Pfold combines an explicit evolutionary model of the RNA sequences with a probabilistic model of the secondary structures. When trying to extend this approach to incorporate thermodynamic folding, the first idea was to develop a combined probabilistic model for evolution *and* folding, using the partition function approach of McCaskill (28) as a probabilistic model for thermodynamic folding.

There are two main problems for a combined probabilistic model. First, there is no simple way to weight the different information sources in such a combined model. And second, the structure prediction would be based on a maximum likelihood or maximum *a posteriori* (MAP) approach. Recent work by Carvalho and Lawrence (29) has shown that this approach often does not yield to the desired result. Basically, the implicit assumption of a maximum likelihood or MAP approach is that the structure with the highest probability is also the structure where the ensemble of close neighbors has the highest probability mass, which is often not the case. Hence, Carvalho and Lawrence proposed different new classes of estimators, which included estimators already used in sequence alignment and RNA secondary structure prediction. One example is the maximum expected accuracy (MEA) method originally introduced by ref. (30) and successfully applied to sequence alignment in ProbCons (31) and to RNA structure prediction in CONTRAfold (3). A partition function version of the Sankoff algorithm (which is related to MEA scoring) was introduced in ref. (32). Another example is Pfold itself, since the current implementation does not use the maximum likelihood approach originally introduced in ref. (33), but is based on reliability scores (27), which can be interpreted as a variant of MEA scoring.

Hence, our approach is based on a combined MEA score. The combined score is based on both base pair and single-strand probabilities as calculated by RNAfold and the reliabilities of base pairs and single-stranded positions as extracted from Pfold. The combined MEA scoring has the advantage that it allows for an explicit weighting of the contribution of phylogenetic and thermodynamic information and smoothens the effects of incorrectly predicted base pair probabilities. The latter is due to the fact that the algorithm searches for a structure that shares as many base pairs as possible with all alternative structures. We implemented this Probabilistic Evolutionary and Thermodynamic folding algorithm (PETfold) for multiple RNA sequence alignments.

Recently, several methods for finding the common structure in a set of sequences that are either unaligned or aligned have been published (34–37). Essentially, they all carry out a global alignment or predict a common structure on a set of globally aligned sequences. RNAalifold (38) basically applies energy minimization to a complete alignment. In addition, it introduces an evolutionarily motivated score to measure sequence covariation for base pairs. Since it is widely used and applies a thermodynamic-based scoring scheme, it was chosen as a benchmark method in addition to Pfold.

## MATERIALS AND METHODS

### The Pfold model

We now recall the Pfold (33) model. In the following, a secondary structure $\sigma$ is a set of base pairs that do not cross. Let $A$ be a multiple alignment of the sequences $s_1 \ldots s_n$. Furthermore, let $T$ be a given evolutionary tree and $M$ be a prior model for the secondary structures. Thus the model provides a probability distribution on the structures, given the data (i.e. the alignment $A$) and the background information (i.e. the secondary structure background model $M$ and the tree $T$):

$$\begin{aligned}\Pr[\sigma|A, T, M] &= \frac{\Pr[A, \sigma|T, M]}{\Pr[A|T, M]} \\ &= \frac{\Pr[A|T, \sigma, M]P[\sigma|T, M]}{\Pr[A|T, M]}\end{aligned} \qquad \mathbf{1}$$

Since $\Pr[A|T, M]$ is independent from the structure $\sigma$, we need to optimize only $\Pr[A|T, \sigma, M]P[\sigma|T, M]$.

Here, $P[\sigma|T, M]$ is the prior distribution over all secondary structures. $M$ is given by the following simple SCFG, which is taken from Pfold:

$$S \to LS|L \qquad F \to dFd|LS \qquad L \to s|dFd. \qquad \mathbf{2}$$

It has been shown by Dowell and Eddy (39) that this grammar performs best on a given benchmark data set. Furthermore, this grammar is unambiguous, i.e. every structure is generated in a unique way. For each structure $\sigma$, let $\tau_M(\sigma)$ be the associated parse tree that produces the structure $\sigma$ using the grammar $M$. With $r(\sigma)$ we denote the root node of $\tau_M(\sigma)$.

The term $\Pr[A|T, \sigma, M]$ provides a distribution over alignments. Here, only the sequences evolve and the common structure $\sigma$ (adopted by all sequences) determines the mode of evolution. The alignment probability is calculated as the product of alignment column probabilities, which are calculated using Felsenstein's dynamic programming for phylogenetic trees (see Supplementary Material for details).

The full term $\Pr[A|T, \sigma, M]P[\sigma|T, M]$ now can be calculated with a combined SCFG by multiplying the probabilities for the secondary structure rules in Equation (2) with the probabilities for generating the associated alignment columns (see Supplementary Material again). In the following, we will denote it with $\Pr(r(\sigma), A)$.

Originally, Pfold (33) used a MAP approach for calculating the consensus structure $\sigma$. Thus, $\sigma$ was defined to be the structure that maximizes $\Pr(r(\sigma), A)$. This maximization problem can be solved using the CYK algorithm. Later, this was replaced in Pfold by a more successful MEA approach discussed in the next section.

### Maximum expected accuracy

The basic idea of MEA scoring is to consider $\Pr(r(\sigma), A)$ as a probability distribution over consensus structures, where higher probabilities denote a better fit to the alignment and its associated evolutionary history. We write position $i \notin \sigma$ as short for 'position $i$ is not involved in a base pair in $\sigma$' [i.e. $\forall j : ((i,j) \notin \sigma \wedge (j,i) \notin \sigma)$]. Given the structure $\sigma$ and an alignment $A$ with $m$ columns, the set of all unpaired positions in the consensus structure is denoted as $\sigma = \{i \notin \sigma \mid 1 \leq i \leq m\}$. Then, we can compute the expected overlap ex-over$^{\text{evo}}(\sigma)$ of a specific consensus structure $\sigma$ with all possible consensus structures, weighted according to their probabilities. This can be done as follows:

$$\begin{aligned}
&\text{ex-over}^{\text{evo}}(\sigma)\\
&\quad= \sum_{\sigma'} [|\sigma \cap \sigma'| + \alpha |\text{sg}(\sigma) \cap \text{sg}(\sigma')|] \times \Pr(r(\sigma'), A)\\
&\quad= \sum_{(i,j) \in \sigma} \sum_{\sigma'} \delta((i,j) \in \sigma') \times \Pr(r(\sigma'), A)\\
&\qquad+ \alpha \sum_{i \in \text{sg}(\sigma)} \sum_{\sigma'} \delta(i \in \text{sg}(\sigma')) \times \Pr(r(\sigma'), A),
\end{aligned}$$

where the term $\delta(\psi)$ for a proposition $\psi \equiv ((i,j) \in \sigma')$ or $\psi \equiv (i \in \text{sg}(\sigma'))$ is 1 if the $\psi$ is true and 0 otherwise. $\alpha$ is a free parameter that weights single stranded against base pair positions. Since a base pair occupies two positions, $\alpha$ should be smaller than 1\2 (where $\alpha = 1\backslash 2$ means that prediction errors on base pairs and single stranded positions are weighted equally).

We now define the reliability scores for a base pair $(i,j)$ and for a single-stranded base $i$ as in Pfold by

$$\begin{aligned}
\mathcal{R}_{A,T,M}(i,j) &= \sum_{\sigma} \delta((i,j) \in \sigma) \times \Pr(r(\sigma), A),\\
\mathcal{R}^{\text{sg}}{}_{A,T,M}(i) &= \sum_{\sigma} \delta(i \in \text{sg}(\sigma)) \times \Pr(r(\sigma), A).
\end{aligned} \qquad \mathbf{3}$$

Both can be calculated by an inside/outside algorithm. Using the terms in Equation (3), we can redefine the expected overlap of a consensus structure as

$$\text{ex-over}^{\text{evo}}(\sigma) = \sum_{(i,j) \in \sigma} \mathcal{R}_{A,T,M}(i,j) + \alpha \sum_{i \in \text{sg}(\sigma)} \mathcal{R}^{\text{sg}}{}_{A,T,M}(i),$$

which allows us to calculate the consensus structure with the maximal expected overlap using a Nussinov style algorithm.

### Extension by folding energies

Before we can extend the model, we have to more precisely define what it means that two sequences can adopt the same consensus structure under a given alignment matrix $A = (a_{u,l})$, where $u$ denotes the number of sequences and $l$ the number of alignment columns. This is necessary because a consensus structure is defined as a set of paired alignment columns. Hence, let $n$ be the number of sequences and let $f^u_A(i) = l$ be the alignment column corresponding to position $i$ in sequence $s_u$. The mapping $f^u_A$ can be extended to structures:

$$f^u_A(\sigma) = \{(f^u_A(i), f^u_A(j)) \mid (i,j) \in \sigma\}.$$

In the previous section, we searched for a consensus structure that had the maximal expected overlap with other possible consensus structures defined by the probabilistic evolutionary model, thus minimizing the expected number of evolutionary prediction errors. Now, we also want to evaluate the expected overlap for each sequence $s$ with its ensemble of structures as given by the energy model. This implies that for each sequence $s$, we consider the distribution of structures as introduced by McCaskill (28). For this purpose, let $p^s_{k,l} = \sum_{(k,l) \in \sigma} \Pr[\sigma|s]$ be the base pair probabilities for a sequence $s$ as calculated by, e.g. RNAfold –p (1) and $q^s_k = 1 - \sum_{l \neq k} p^s_{k,l}$ the probability for position $k$ being single stranded in sequence $s$. The combined expected overlap now consists of two parts, generally weighted with 1 for the conservation part and $\beta$ for the thermodynamic overlap:

$$\text{ex-over}(\sigma) = \text{ex-over}^{\text{evo}}(\sigma) + \frac{\beta}{n} \times \text{ex-over}^{\text{str}}(\sigma) \qquad \mathbf{4}$$

where ex-over$^{\text{str}}(\sigma)$ is the expected overlap of $\sigma$ with all structures from all sequences. Formally, this is defined by

$$\begin{aligned}
&\text{ex-over}^{\text{str}}(\sigma)\\
&\quad= \sum_{u,\sigma'} [|\sigma \cap f^u_A(\sigma')| + \alpha |\text{sg}(\sigma) \cap \text{sg}(f^u_A(\sigma'))|] \times \Pr[\sigma'|s_u]\\
&\quad= \sum_{(i,j) \in \sigma} \sum_{u} \sum_{\sigma'} \delta((i,j) \in f^u_A(\sigma')) \times \Pr[\sigma'|s_u]\\
&\qquad+ \alpha \sum_{i \in \text{sg}(\sigma)} \sum_{u} \sum_{\sigma'} \delta(i \in \text{sg}(f^u_A(\sigma'))) \times \Pr[\sigma'|s_u]
\end{aligned}$$

$$= \sum_{(i,j)\in\sigma} \sum_{u} p^{u}_{f_A^{-1}(i,j)} + \alpha \sum_{i\in \mathrm{sg}(\sigma)} \sum_{u} q^{u}_{f_A^{-1}(i)},$$

Here, $p^{u}_{f_A^{-1}(i,j)}$ denotes the base pair probability for sequence $s_u$, written by alignment columns:

$$p^{u}_{f_A^{-1}(i,j)} = \begin{cases} p^{s_u}_{k,l} & \text{if columns } i,j \text{ are gap-free in } s_u \\ & \text{and } (k,l) = (f^u_A(i), f^u_A(j)) \\ 0 & \text{else} \end{cases}$$

and analogously for $q^{u}_{f_A^{-1}(i)}$.
Overall, we obtain

$$\text{ex-over}(\sigma) = \sum_{(i,j)\in\sigma} \left( \mathcal{R}_{A,T,M}(i,j) + \frac{\beta}{n} \sum_{u} q^{u}_{f_A^{-1}(i)} \right) + \sum_{i\in \mathrm{sg}(\sigma)} \alpha \left( \mathcal{R}^{\mathrm{sg}}{}_{A,T,M}(i) + \frac{\beta}{n} \sum_{u} p_{f_A^{-1}(i)} \right)$$

The consensus structure maximizing this expectation can again be calculated by a Nussinov-style algorithm.

### Reliably conserved substructure

Highly conserved substructures are under strong evolutionary pressure, possibly caused by interactions with bound molecules or other functional constraints. As mentioned in the Introduction section, these substructures cannot be accounted for in thermodynamic folding algorithms. We refer to them as a *reliably conserved substructure* $\sigma^{rel}$, which is defined as follows. A substructure is a structure that fixes only some of the positions in a sequence, hence being a partial structure. A *partial structure* $\sigma^{p}$ is a tuple $(\mathcal{B}, \mathcal{S})$ consisting of a set of base pairs $\mathcal{B}$ and a set of single-stranded positions $\mathcal{S}$ such that no single-stranded position in $\mathcal{S}$ is part of a base pair in $\mathcal{B}$. Then, a partial structure $\sigma^{p}$ is a *substructure* of a structure $\sigma$ if $\mathcal{B} \subseteq \sigma$ and $\mathcal{S} \subseteq \mathrm{sg}(\sigma)$.

The reliably conserved substructure $\sigma^{rel}$ is a substructure of $\sigma$ that is determined by the evolutionarily highly reliable positions. They are selected by using thresholds on the reliability scores of single-stranded ($p^{\text{threshold}}_{ss}$) and base-paired positions ($p^{\text{threshold}}_{bp}$). We get $\sigma^{rel}$ by using a Nussinov-style approach with the additional condition that

$$\forall (i,j) \in \sigma^{rel} : \mathcal{R}_{A,T,M}(i,j) > p^{\text{threshold}}_{bp}$$
$$\forall i \in \sigma^{rel} : \mathcal{R}_{A,T,M}(i) > p^{\text{threshold}}_{ss}$$

There are two possible approaches to determine these thresholds. First, they can be estimated through parameter tuning using a data set of known RNA families to predict RNA structures most similar to their structure annotations (see Result section). Second, a statistical approach can be applied by determining positions whose reliability is significantly better than the average (see Supplementary Material). Our observation was that in practice the first approach worked better than the second.

For specific cases where there are many additional constraints imposed on the structure, it is a good strategy to identify the reliably conserved substructure $\sigma^{rel}$ *first*, and then search for the consensus structure $\sigma$ that maximizes ex-over($\sigma$) *and* contains $\sigma^{rel}$ as a substructure. Reliably conserved substructures are easily integrated into the earlier described Nussinov-style algorithm by setting the weighting factor of thermodynamic overlap $\beta$ to zero in Equation (2) (see Supplementary Material for details).

### Gaps

Gaps are treated in a two-step procedure. First, all columns are removed from the alignment where ≤75% of the sequences have nucleotides. These columns are integrated as gaps in the consensus secondary structure at the end, a strategy that is adapted from Pfold. Second, sequence-dependent probabilities $p^{s_u}_{k,l}$ and $q^{s_u}_{k}$ are calculated without gaps and probabilities of gaps are estimated as the average in the appropriate column of the alignment.

### Data

Rfam (40) is the most up-to-date comprehensive and freely accessible RNA structure database, so we used 46 Rfam (version 8.0) seed alignments for parameter tuning and for PETfold benchmarking. Our data set consists of 17 RNA families recommended by refs. (21) and (11) and 29 additional families possessing a high-quality alignment documented by the SARSE project (41). The alignments of the second set of data are characterized by a score for inconsistent base pairs ≤4 and a score for novel base pairs ≤4. These criteria are satisfied by 43 families, of which two families overlap the first set and 12 were rejected because of an annotated structure with <40% of bases being involved in base pairs. Large alignments were reduced by sequence similarity as in (41). An overview of all RNA families in the data set is given in Supplementary Table 1.

### Performance evaluation

The RNA structure predictions of PETfold, Pfold and RNAalifold are evaluated by looking for their correlation to the related Rfam structure annotation, ignoring non-canonical base pairs. Therefore, we used the Matthews correlation coefficient (42) defined as

$$\mathrm{MCC} = \frac{P_t N_t - P_f N_f}{\sqrt{(P_t + N_f)(P_t + P_f)(N_t + P_f)(N_t + N_f)}}, \qquad \mathbf{5}$$

where $P_t$ is the number of mutual base pairs in the two assignments (true positives), $N_t$ the number of mutual pairs of bases not base pairing (true negatives), $P_f$ the number of predicted base pairs not in the Rfam assignment (false positives) and $N_f$ the number of base pairs in the Rfam assignment not predicted to pair (false negatives). We also calculated sensitivity, $\mathrm{SEN} = P_t/(P_t + N_f)$ and positive predictive value, $\mathrm{PPV} = P_t/(P_t + P_f)$. Their geometric mean is a good approximation to the MCC for base pair prediction (43). The parameter tuning was done by optimizing the MCC to get a best possible tradeoff between sensitivity and PPV. In addition, we used the more stringent structure evaluation scheme $R_5$ introduced by Gorodkin *et al.* (44) (see Supplementary Material for full details). This was necessary because MCC does not evaluate whether or not two positions coincide in their base pair partners when the two positions do base pair.

## RESULTS

### Pfold similarity

PETfold without thermodynamic probabilities (set parameter $\beta = 0$) calculates exactly the same most reliable structure as Pfold. We interpret highly conserved parts of this RNA structure as evolutionary substructure $\sigma^{rel}$.

### Parameter tuning

The default parameters of PETfold are set by optimizing the MCC of predicted RNA secondary structure to the related Rfam structure annotation over all RNA families in our data set. Hence, we ran PETfold with different values for the single-stranded probability weighting factor ($\alpha \leq 0.5$) and different reliability thresholds for evolutionary highly conserved single-stranded ($p_{ss}^{\text{threshold}}$) and base-paired positions ($p_{bp}^{\text{threshold}}$). The latter two values determine the reliably conserved substructure $\sigma^{rel}$. The conservation part and thermodynamic overlap are equally weighted ($\beta = 1$) as PETfold should not be overfitted to RNA families with strong conservation and lower thermodynamic stability or vice versa.

As shown in Figure 1, PETfold with $\alpha = 0.2$ and $p_{ss}^{\text{threshold}} = 1$ predicts the best RNA structure as compared to the Rfam annotated structure. These settings optimize the structure of 30 RNA families. The total rejection of the reliably conserved substructure $\sigma^{rel}$ ($p_{ss}^{\text{threshold}} = p_{bp}^{\text{threshold}} = 1$) or $p_{bp}^{\text{threshold}} < 0.9$ slightly decreases the performance of PETfold. Therefore, we are pragmatic and use $p_{bp}^{\text{threshold}} = 0.9$, which predicts the best RNA structure for 38 families. In summary, we choose $\alpha = 0.2$, $\beta = 1$, $p_{ss}^{\text{threshold}} = 1$ and $p_{bp}^{\text{threshold}} = 0.9$ as default parameters for PETfold having optimized the structures of 25 RNA families that comprise 54% of the data set.

The parameter tuning of PETfold showed that in many cases the quality improves by adding thermodynamic probabilities even when there are highly conserved substructures. PETfold produces the highest MCCs when single-stranded positions are not considered as part of the reliably conserved substructure $\sigma^{rel}$. Furthermore, defining $\sigma^{rel}$ which consists of evolutionarily conserved base pairs with high reliability resulted only in a minor improvement. This emphasizes that phylogenetic structure information is mostly supported by folding energy. Another unexpected result was the poor weighting of single stranded against base pair positions ($\alpha = 0.2$), i.e. base pair probabilities have a larger impact on RNA structure prediction.

### Performance

Next, we compared the RNA structure predictions of PETfold with Pfold and RNAalifold, using the default parameters for each program. On average over the entire data set, PETfold performed better than the other methods for a wide range of parameter settings. PETfold with default parameters predicts base pairs with 0.85 PPV, 0.88 SEN and 0.86 accuracy. Its mean MCC to the Rfam annotations of 0.85 is significantly higher than the 0.71 MCC obtained for Pfold and the 0.79 obtained for RNAalifold (Table 1).

The MCCs for PETfold, Pfold and RNAalifold are listed for all families in the data set in Supplementary Table 1. PETfold predicts a structure that is better than the ones predicted by Pfold and RNAalifold for 18 RNA families. In contrast, Pfold achieves this for only 7 families and RNAalifold for 16 families. Considering a confidence interval of 0.01, we observed 27 of the most accurate predictions by PETfold, 15 by Pfold and 18 by RNAalifold. These include cases where two or three methods

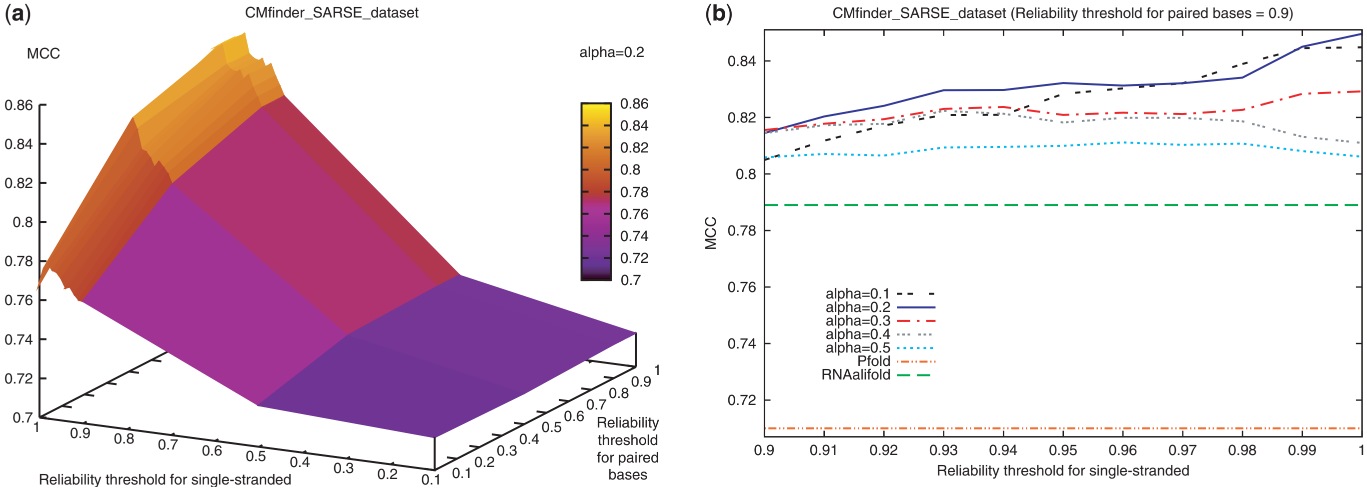

**Figure 1.** Parameter tuning to optimize mean MCC. The mean MCCs to Rfam structures of all RNA families in the data set are shown for structure predictions by PETfold with numerous parameter settings, which include the weighting factor for single-stranded positions $\alpha$, reliability thresholds for evolutionary constrained single-stranded $p_{ss}^{\text{threshold}}$ as well as paired bases $p_{bp}^{\text{threshold}}$. (**a**) $0.1 \leq p_{ss}^{\text{threshold}} \leq 1$ and $0.1 \leq p_{bp}^{\text{threshold}} \leq 1$, which are plotted on the $x$-axis as well as $y$-axis, and $\alpha = 0.2$; (**b**) $0.9 \leq p_{ss}^{\text{threshold}} \leq 1$ and $p_{bp}^{\text{threshold}} = 0.9$ and several $\alpha$. (b) also shows the mean MCC of Pfold and RNAalifold.

**Table 1.** Performance on data set

| | PPV[a] | SEN[b] | ACC[c] | MCC[d] | $R_5$[e] | Time (s) |
|---|---|---|---|---|---|---|
| PETfold | 0.852 | 0.876 | 0.864 | 0.850 | 0.722 | 153.0 |
| Pfold | 0.662 | 0.843 | 0.747 | 0.710 | 0.575 | 65.6 |
| RNAalifold | 0.758 | 0.842 | 0.799 | 0.789 | 0.652 | 2.3 |

[a] Positive predictive value.
[b] Sensitivity.
[c] Accuracy.
[d] Matthews corelation coefficient.
[e] $R_5$ correlation coefficient.

predicted the same structure. More specifically, seven structures were predicted by PETfold and Pfold, three by PETfold and RNAalifold and two structures [*Histone3* (45) and *Rhino_CRE* (46)] were predicted by all three methods.

Concerning the structure predictions that completely agreed with the Rfam annotations, the three methods correctly predict the *Histone3* structure. In addition, Pfold predicts the exact structure of *mir-194* (47) microRNA precursor and RNAalifold of the viral 3′-UTR stem-loop *s2m* (48). In contrast, PETfold provides accurate structure predictions for four viral RNA families [*HepC_CRE* (49), *IBV_D-RNA* (50), *TCV_H5* (51) and *TCV_Pr* (51)]. Their alignments contain between 3 and 64 sequences and have mean pairwise identities (MPI) ranging from 77% to 95%, showing that PETfold is not overfitted to certain alignment features. Furthermore, a scan of all 574 seed alignments in Rfam version 8.0 results in the best performance by PETfold (MCC = 0.63), followed by Pfold (MCC = 0.62) and RNAalifold (MCC = 0.58). However, it should be noted that a large number of annotated RNA structures in Rfam are themselves predictions, many by Pfold.

The performance of PETfold increases greatly in the case of several H/ACA snoRNAs, the already mentioned example of ncRNAs that often do not fold into the structure of lowest free energy, e.g. *HACA_sno_snake* (52) (MCC with PETfold: 0.79, Pfold: 0.25, RNAalifold: 0.41) as a member of our data set and *SNORA51* (53) (MCC with PETfold: 0.91, Pfold: 0.13, RNAalifold: 0.57). The functional RNA element of HIV-1 mRNA, Rev response element (RRE), has a verified RNA secondary structure of 337 nt in length (54). The accuracy of the predicted structure for this element increases by 30% when using PETfold (MCC = 0.90).

PETfold takes $O(L^2)$ space and $O(L^3)$ time when the number of sequences $N$ is much smaller than the sequence length $L$. The most time-consuming parts of the algorithm are the calculation of evolutionarily reliabilities using Pfold ($O(L^3) + O(L^2)$), the calculation of energy-based probabilities of $N$ sequences using RNAfold ($N \times O(L^3)$) and the Nussinov-style algorithm ($O(L^3)$). In practice, the running time of PETfold is approximately twice that of Pfold and much longer than that of RNAalifold (see Table 1 for details). Major reasons for the longer runtimes are the implementation of PETfold in Perl and the external calls to Pfold and RNAfold.

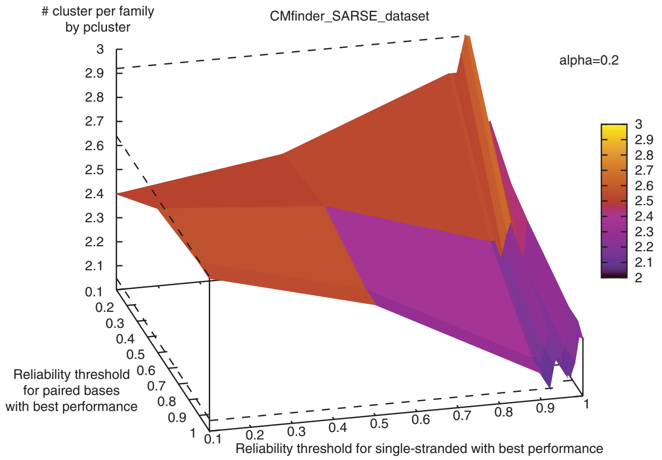

**Figure 2.** Correlation between alignment diversity and evolutionary constraint substructures. The correlation between the diversity (number of structural clusters per family calculated by Pcluster) of alignments and the reliability thresholds used by PETfold to predict the best RNA structure, averaged over all families in the data set. We can see that less diverse alignments perform best almost without evolutionary constraints. *Intron_gpII* (57) is excluded because of its exceedingly high number of 10 structural clusters.

### Family specific parameter settings

The suggested PETfold parameters are based on the average performance over all RNA families in the data set. We observed a common distribution of PETfold performance for different parameter settings in which low-reliability thresholds for evolutionary substructures $\sigma^{rel}$ decrease the prediction accuracy (Figure 1). Thus, the general application of evolutionary constraints has negative influences on the structure prediction.

Nevertheless, several RNA families show a different performance distribution. For instance, PETfold with default parameters performs worse than RNAalifold for the Rotavirus *cis*-acting replication element (*Rota_CRE*) (55). However, if we lower the threshold for conserved base pairs, then PETfold predicts a structure with MCC = 0.84 compared to the MCC = 0.76 achieved by RNAalifold.

We could not find a general correlation of the best performing reliability thresholds with previously used alignment features such as the structural conservation index (SCI) (23) or mean pairwise sequence identity (MPI). However, structurally diverse alignments tend to benefit from lower reliability thresholds due to a higher evolutionary information content. The alignment diversity can be measured by Pcluster (41), which divides the sequences of an RNA alignment into subgroups with different consensus structures. On average, PETfold performs best with high-reliability thresholds for RNA families with low number of structural clusters. This correlation is shown in Figure 2. For instance, the best performing reliability thresholds for *Rota_CRE*, whose alignment is divided in three clusters by Pcluster, are well described by Figure 2.

In summary, PETfold's default parameters should be regarded as a reasonable approximation for the RNA family alignments with different levels of diversity.

## DISCUSSION

We presented a new method, named PETfold, which unifies probabilistic evolutionary and thermodynamic models to predict global RNA structures from multiple alignments. Highly conserved RNA structures are probably naturally selected due to their functional relevance, while less conserved sequences tend to fold in a state of minimum free energy. Combining both types of information increases the predictive power: we rely on phylogenetic information if the structure elements are highly conserved and extend the model by including folding energies to build the surrounding structure.

The 46 Rfam seed alignments used for benchmarking are selected for their consistency. They might include manually optimized structure information, that are not achieved by sequence alignment algorithms. However, we are interested in the best possible structure prediction, which is often only found using conserved structure information in the alignment. The analyses of PETfold revealed that the selection of a reliably conserved substructure $\sigma^{rel}$ is less important than expected. This has been confirmed by estimating the statistical significance of the reliability scores, which shows that only high-reliability scores are significant. On the other hand, the combination of an evolutionary model with a thermodynamic model outperforms the widely used RNA structure predictors Pfold and RNAalifold.

Nevertheless, there are cases where the evolutionary conservation is more important. Especially for RNA families with an active site like H/ACA snoRNAs, the usage of reliably conserved substructures improves the predictions. A first guess, namely that the impact of the evolutionary substructure might be negatively correlated with the SCI, could not be confirmed. However, we observed a correlation to the number of structural clusters in an alignment as measured by Pcluster. This should be investigated in greater detail since it provides a new possibility to classify multiple RNA alignments. Currently, the measurements as used, e.g. in BRAliBase (26) are MPI and SCI.

There are two possible improvements. One could be the application of stacking base pair probabilities (56). These probabilities can be easily integrated in the MEA approach if the rule $F \rightarrow \mathrm{d}F d$ of the SCFG is applied. Nevertheless, a first trial of this extension using the stacking base pair probabilities calculated by RNAfold (1) did not improve the performance of PETfold and may be due to the fact that these are not independent of the simple base pair probabilities.

A second possible improvement considers the fact that structural stability is at least partially accounted twice since Pfold already favors stable base pairs such as G—C. This does not pose a major problem since the thermodynamic part adds a lot of new information like sequence-dependent stacking or properties of the complete structure ensemble (encoded in base pair probabilities). Nevertheless, it would be nice to explicitly separate these different information sources in a future version.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. The source code of PETfold is available under the GNU Public License at http://genome.ku.dk/resources/petfold.

## ACKNOWLEDGEMENTS

This work was supported by the Danish Research Council for Technology and Production (FTP) and the Danish Center for Scientific Computation. S.E.S was supported by the Lundbeck foundation. The authors thank the anonymous referees for their detailed comments.

## FUNDING

Funding for open access charge: Albert-Ludwigs-Universität Freiburg.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chemie*, **125**, 167–188.
2. Zuker,M. (1994) Prediction of RNA secondary structure by energy minimization. *Methods Mol. Biol.*, **25**, 267–294.
3. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
4. Schattner,P., Barberan-Soler,S. and Lowe,T.M. (2006) A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA*, **12**, 15–25.
5. Hertel,J., Hofacker,I.L. and Stadler,P.F. (2007) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, **24**, 158–164.
6. Sato,K., Morita,K. and Sakakibara,Y. (2008) PSSMTS: position specific scoring matrices on tree structures. *J. Math. Biol.*, **56**, 201–214.
7. Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
8. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
9. Gorodkin,J., Heyer,L. and Stormo,G. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.*, **25**, 3724–3732.
10. Havgaard,J.H., Lyngso,R.B. and Gorodkin,J. (2005) The FOLDALIGN web server for pairwise structural RNA alignment and mutual motif search. *Nucleic Acids Res.*, **33** (Web Server Issue), W650–W653.
11. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.
12. Havgaard,J.H., Torarinsson,E. and Gorodkin,J. (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, **3**, 1896–1908.
13. Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.
14. Harmanci,A.O., Sharma,G. and Mathews,D.H. (2007) Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*, **8**, 130.
15. Hofacker,I.L., Bernhart,S.H. and Stadler,P.F. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.

16. Backofen,R. and Will,S. (2004) Local Sequence-Structure Motifs in RNA. *J. Bioinform. Comput. Biol. (JBCB)*, **2**, 681–698.
17. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring Non-Coding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Comput. Biol.* **3**, e65.
18. Harmanci,A.O., Sharma,G. and Mathews,D.H. (2008) PARTS: probabilistic alignment for RNA joinT secondary structure prediction. *Nucleic Acids Res.*, **36**, 2406–2417.
19. Holmes,I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics* **6**, 73.
20. Dowell,R.D. and Eddy,S.R. (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, **7**, 400.
21. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder – a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
22. Meyer,I.M. and Miklos,I. (2007) SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.*, **3**, e149.
23. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
24. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
25. Torarinsson,E., Sawera,M., Havgaard,J.H., Fredholm,M. and Gorodkin,J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.*, **16**, 885–889.
26. Gardner,P.P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
27. Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **31**, 3423–3428.
28. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
29. Carvalho,L.E. and Lawrence,C.E. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl. Acad. Sci. USA*, **105**, 3209–3214.
30. Miyazawa,S. (1995) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng.*, **8**, 999–1009.
31. Do,C.B., Mahabhashyam,M.S.P., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
32. Hofacker,I.L. and Stadler,P.F. (2004) *The Partition Function Variant of Sankoff's Algorithm. Lecture Notes in Computer Science LNCS* 3039. Springer, Heidelberg.
33. Knudsen,B. and Hein,J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15(6)**, 446–454.
34. Kiryu,H., Kin,T. and Asai,K. (2007) Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics*, **23**, 434–441.
35. Anwar,M., Nguyen,T. and Turcotte,M. (2006) Identification of consensus RNA secondary structures using suffix arrays. *BMC Bioinformatics*, **7**, 244.
36. Bafna,V., Tang,H. and Zhang,S (2006) Consensus folding of unaligned RNA sequences revisited. *J. Comput. Biol.* **13**, 283–295.
37. Hamada,M., Tsuda,K., Kudo,T., Kin, T. and Asai, K. (2006) Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, **22**, 2480–2487.
38. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
39. Dowell,R.D. and Eddy,S.R. (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 71.
40. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, **33**, (Database Issue) D121–D124.
41. Andersen,E.S., Lind-Thomsen,A., Knudsen,B., Kristensen,S.E., Havgaard,J.H., Torarinsson,E., Larsen,N., Zwieb,C., Sestoft,P., Kjems,J. *et al.* (2007) Semiautomated improvement of RNA alignments. *RNA*, **13**, 1850–1859.
42. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
43. Gorodkin,J., Stricklin,S.L. and Stormo,G.D. (2001) Discovering Common Stem-Loop Motifs in Unaligned RNA Sequences. *Nucleic Acids Res.*, **29**, 2135–2144.
44. Gorodkin,J. (2004) Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.*, **28**, 367–374.
45. Williams,A.S. and Marzluff,W.F. (1995) The sequence of the stem and flanking sequences at the 3′ end of histone mRNA are critical determinants for the binding of the stem-loop binding protein. *Nucleic Acids Res.*, **23**, 654–662.
46. McKnight,K.L. and Lemon,S.M. (1998) The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA*, **4**, 1569–1584.
47. Lagos-Quintana,M., Rauhut,R., Meyer,J., Borkhardt,A. and Tuschl,T. (2003) New microRNAs from mouse and human. *RNA*, **9**, 175–179.
48. Jonassen,C.M., Jonassen,T.O. and Grinde,B. (1998) A common RNA motif in the 3′ end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. *J. Gen. Virol.*, **79(Pt 4)**, 715–718.
49. You,S., Stump,D.D., Branch,A.D. and Rice,C.M. (2004) A cis-acting replication element in the sequence encoding the nS5B RNA-dependent RNA polymerase is required for hepatitis c virus RNA replication. *J. Virol.*, **78**, 1352–1366.
50. Dalton,K., Casais,R., Shaw,K., Stirrups,K., Evans,S., Britton,P., Brown,T.D. and Cavanagh,D. (2001) Cis-acting sequences required for coronavirus infectious bronchitis virus defective-RNA replication and packaging. *J. Virol.*, **75**, 125–133.
51. McCormack,J.C. and Simon,A.E. (2004) Biased hypermutagenesis associated with mutations in an untranslated hairpin of an RNA virus. *J. Virol.*, **78(14)**, 7813–7817.
52. Chang,L.S., Lin,S.K. and Wu, P.F. (1998) Differentially expressed snoRNAs in Bungarus multicinctus (Taiwan banded krait). *Biochem. Biophys. Res. Commun.*, **245**, 397–402.
53. Kiss,A.M., Jdy,B.E., Bertrand,E. and Kiss,T. (2004) Human box h/aca pseudouridylation guide RNA machinery. *Mol. Cell. Biol.*, **24**, 5797–5807.
54. Le,S.Y., Zhang,K. and Maizel,J.V.J. (2002) RNA molecules with structure dependent functions are uniquely folded. *Nucleic Acids Res.*, **30**, 3574–2582.
55. Chen,D., Barros,M., Spencer,E. and Patton,J.T. (2001) Features of the 3′-consensus sequence of rotavirus mRNAs critical to minus strand synthesis. *Virology*, **282**, 221–229.
56. Walter,A., Turner,D., Kim,J., Lyttle,M., Muller,P., Mathews,D. and Zuker,M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.
57. Bonen, L. and Vogel, J. (2001) The ins and outs of group II introns. *Trends Genet.*, **17**, 322–331.