AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions

Stefan E Seemann[†1], Andreas S Richter[†2], Jan Gorodkin[1] and Rolf Backofen*[2]

Abstract

Background: Many regulatory non-coding RNAs (ncRNAs) function through complementary binding with mRNAs or other ncRNAs, *e.g.*, microRNAs, snoRNAs and bacterial sRNAs. Predicting these RNA interactions is essential for functional studies of putative ncRNAs or for the design of artificial RNAs. Many ncRNAs show clear signs of undergoing compensating base changes over evolutionary time. Here, we postulate that a non-negligible part of the existing RNA-RNA interactions contain preserved but covarying patterns of interactions.

Methods: We present a novel method that takes compensating base changes across the binding sites into account. The algorithm works in two steps on two pre-generated multiple alignments. In the first step, individual base pairs with high reliability are found using the `PETfold` algorithm, which includes evolutionary and thermodynamic properties. In step two (where high reliability base pairs from step one are constrained as unpaired), the principle of cofolding is combined with hierarchical folding. The final prediction of *intra*- and *inter*-molecular base pairs consists of the reliabilities computed from the constrained expected accuracy scoring, which is an extended version of that used for individual multiple alignments.

Results: We derived a rather extensive algorithm. One of the advantages of our approach (in contrast to other RNA-RNA interaction prediction methods) is the application of covariance detection and prediction of pseudoknots between *intra*- and *inter*-molecular base pairs. As a proof of concept, we show an example and discuss the strengths and weaknesses of the approach.

Background

Predicting RNA-RNA interactions is a rapidly growing area within RNA bioinformatics and is essential for the process of assigning function to known as well as *de novo* predicted non-coding RNAs (ncRNAs) such as those identified in *in silico* screens for RNA structures [1-7]. This candidate information along with the data generated from deep sequencing analyses emphasise the need to predict RNA-RNA interactions. In part, this is because there currently is no high-throughput method available for the reliable analysis of RNA-RNA interactions; however, computational prediction of RNA-RNA interactions is also essential for the identification of putative targets of known and *de novo* predicted ncRNAs. With the main exception of microRNA target prediction, the current approaches essentially evaluate the stabilities of the common complexes between ncRNAs and target RNAs by computing the overall free energy using two major strategies (see, *e.g.*, [8] for a recent review).

The first strategy, represented through the implementations of `RNAup`[9] and `IntaRNA`[10], uses pre-calculated values for all possible regions of interaction to determine the energy required to make that site accessible (called the ED-value for the energy difference). The ED-value is then used to calculate a combined energy of the energy given by the duplex formed by the two interaction regions and the ED-values of both interaction regions. `RNAup` has a complexity of $O(n^3 + nw^5)$, whereas `IntaRNA` has a complexity of $O(n^2)$, which makes it fast enough to be used in genome-wide screens. Both methods are able to predict complex interactions, like kissing

* Correspondence: backofen@informatik.uni-freiburg.de
[2] Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, Freiburg, 79110, Germany
[†] Contributed equally
Full list of author information is available at the end of the article

hairpins, *as long as* the interaction is restricted to one region. However, there are well-known examples where several interaction sites were found, especially for longer ncRNAs. A prominent example is the interaction between OxyS and *fhlA* shown in [11].

The second strategy for RNA-RNA interaction predictions is usually handled with a class of approaches that simultaneously predict a common structure for both RNAs including their interaction. Some of the first approaches, *e.g.*, `pairfold`[12], `RNAcofold`[13] and the method presented by Dirks *et al*. as part of the `NUpack` package [14], concatenate the two sequences using a special linker character. Then, a modified version of the usual RNA folding methods (like `Mfold`[15] and `RNAfold`[16]) is applied to cope with the linker symbol to predict the correct energies. Otherwise, a loop containing the linker symbol would be treated like a hairpin or internal loop, leading to incorrect energy values.

The main disadvantage of the concatenation approach is that the set of candidate joint structures becomes restricted. For this reason, double kissing hairpin interactions (like in OxyS-*fhlA*) cannot be considered. However, alternative (but also most resource demanding) methods have been introduced and extend the class of allowed joint structures. The `IRIS` tool [17] allowed several kissing hairpins using a maximum number of base pair energy model. Then, Alkan *et al*. [18] presented a more realistic energy model and showed the NP-completeness of an unrestricted model. Both approaches predict structures with minimum free energy.

A more stable approach is to consider the partition function because it allows the calculation of interaction probabilities and melting temperatures. This problem was solved independently by Chitsaz *et al*. [19] and Huang *et al*. [20]. In [21], hybrid probabilities were calculated. These approaches have high time complexities of $O(n^6)$, which makes them infeasible for genome-wide applications. Methods to reduce the complexity range from approximation approaches [22,23] to sparsification of the dynamical programming matrix [24].

Here, we present an algorithm for the prediction of RNA-RNA interactions in existing multiple alignments of RNA sequences. Its rationale is based on the assumption that a non-negligible amount of the RNA-RNA interactions contain compensatory base changes across the binding sites. The algorithm presented herein is an extension of the `PETfold` algorithm [25] and makes further use of the principles from `RNAcofold` [13] and computational strategies for hierarchical folding, *e.g.* [26,27]. The latter approach was chosen due to the high computational costs of pseudoknot searches.

Algorithm

The main idea of the introduced method is to use a hierarchical approach to predict an interaction by predicting reliable base pairs within a ncRNA and a mRNA (or another ncRNA), which is followed by prediction of reliable base pairs in the combined sequence. Via this approach, we are able to predict combined pseudoknotted structures, like kissing hairpins, that would be missed otherwise. In both steps, we apply a combined scoring method that predicts consensus base pairs from an alignment using evolutionary conservation and thermodynamic stability information.

The scoring for the first step is according to the standard `PETfold` approach, where we use thresholds for reliable base pairs that have been identified according to training on more than 30 verified interactions in bacteria, which is described later. For the second step, we define a constrained version of the `PETfold` scoring scheme.

Throughout this paper, we consider the concatenation of the two alignments and subsequently (in the base pair prediction process) the concatenation of the corresponding structures. σ will denote a set of base pairs, where the substructures in each part (*e.g.*, ncRNA, mRNA and the base pairs that participate in the interaction) in respective alignments are concatenated or nested (in the dot bracket notation, these substructures have alignment lengths of the ncRNA and mRNA respectively). We use (i, j) to denote a Watson-Crick or G-U wobble base pair between columns i and j. This base pair could be an *intra*-molecular pair in each of the RNA molecules (ncRNA or mRNA) or an *inter*-molecular pair that is involved in an interaction between molecules.

Depending on the context, σ will either be interpreted as a specific structure that implicitly defines the single-stranded positions or as a partial structure that describes an ensemble of structures. In the first case, we define the set of single-stranded positions of a sequence s as

$$\mathrm{ss}(\sigma) = \left\{ i \,\middle|\, \begin{array}{l} 1 \le i \le |\, s \,| \wedge \\ \forall j = 1, \ldots, |\, s \,| : ((i,j) \notin \sigma \wedge (j,i) \notin \sigma) \end{array} \right\}.$$

In the second case, we use $\mathcal{E}(\sigma) = \{\sigma' | \sigma' ? \sigma\}$ to denote the ensemble of all specific structures σ' extending σ. $\mathcal{S}(s)$ denotes the set of nested secondary structures that are defined for the sequence s. We use the same notation for the consensus structures of a given multiple alignment $\mathcal{A}$ with n sequences $s^1 \ldots s^n$. In this case, a position $1 \le i \le |\mathcal{A}|$ refers to a column in the alignment. Furthermore, we use $s \quad \mathcal{A}$ to indicate a sequence $s^1 \ldots s^n$ from the alignment.

The algorithm, like `PETfold` is a maximum expected scoring approach that combines the evolutionary probabilities $\Pr^{ev}[\sigma|\mathcal{A}]$ of a consensus structure, σ, given an alignment, with the thermodynamic probabilities of the associated structures in each sequence. $\Pr^{ev}[\sigma|\mathcal{A}]$ is generated using the stochastic context-free grammar (SCFG) from the Pfold model [28]. The `Pfold` model allows the computation of the probability $\Pr[\sigma|\mathcal{A}, T, M]$ of a consensus structure σ given an alignment $\mathcal{A}$, a phylogenetic tree T for that alignment, and a general background model M for secondary structures. Because the tree T is calculated from the alignment $\mathcal{A}$, and M is constant, we use $\Pr^{ev}[\sigma|\mathcal{A}]$ as short for $\Pr[\sigma|\mathcal{A}, T, M]$.

The (secondary structure) model itself is based on a SCFG that provides a distribution of secondary structures for a given alignment. The combined probability of an alignment $\mathcal{A}$ and a consensus structure σ is

$$\Pr[\mathcal{A}, \sigma \mid T, M] = \Pr[\mathcal{A} \mid T, \sigma] \Pr[\sigma \mid T, M],$$

where $\Pr[\sigma|T, M]$ is the prior distribution of secondary structures and $\Pr[\mathcal{A}|T, \sigma]$ is the probability of the alignment, given a known consensus structure. This is then transformed into $\Pr[\mathcal{A}, \sigma|T, M]$ by applying the Bayesian rule, and further into the posterior distribution $\Pr[\sigma|\mathcal{A}, T, M]$ of consensus structures σ by dividing by $\Pr[\mathcal{A}|T, M]$, which is the sum of all parse trees for an alignment $\mathcal{A}$ given T and M. Note that the comma sign here is just a shortcut for $\wedge$, *i.e.* $\Pr[A, B] = \Pr[A \wedge B]$. We will still use $\wedge$ where it is appropriate.

The probability distributions themselves are formed as follows. For $\Pr[\mathcal{A}|T, \sigma]$, there is an independent evaluation of all base pairs and single-stranded positions:

$$\Pr[\mathcal{A} \mid T, \sigma] = \prod_{(i,j)\in\sigma} \Pr_{bp}[\vec{\mathcal{A}}^i \vec{\mathcal{A}}^j \mid T] \times \prod_{i\in ss(\sigma)} \Pr_{ss}[\vec{\mathcal{A}}^i \mid T],$$

where $\vec{\mathcal{A}}^i$ is the i*th* column of $\mathcal{A}$, and $\sigma \notin \sigma_1^{\mathrm{P}} \cup \sigma_2^{\mathrm{P}}$ for the constrained folding, where σ_1^{P} (σ_2^{P} resp.) is the constrained structure on the first (second resp.) of the two concatenated alignments. For the prior model, the probability $\Pr[\sigma|T, M]$ provides an overall distribution of the secondary structures, which is estimated from rRNA and tRNA sequences. M is given by the following simple SCFG:

$$S \to LS \mid L \quad F \to dFd \mid LS \quad L \to s \mid dFd\,.$$

The evolutionary model and the prior model for RNA structures used in the `Pfold` model are combined into a single SCFG that provides a distribution over $\Pr[A, \sigma|T, M]$ (see additional file 1 for details).

To model the thermodynamic probabilities, we define $\sigma(s^k, \mathcal{A})$ as the structure for the k-th sequence s^k of an alignment $\mathcal{A}$ associated with the consensus structure σ of $\mathcal{A}$. $\Pr^{th}[\sigma(s^k, \mathcal{A})|s^k]$ is the corresponding thermodynamic probability as defined by McCaskill's partition function approach [29].

Using the maximum expected scoring approach, these probabilities are transformed into reliabilities in a two-step approach. Throughout the paper, $\mathcal{R}_{ss}^{\ell}(i)$ is used to denote the reliability of a single-stranded region at alignment position i and $\mathcal{R}_{bp}^{\ell}(i, j)$ the reliability of a consensus base pair (i, j), where ℓ = 1, 2 refers to Step 1 or Step 2 of the combined approach.

Refresher: `PETfold` scoring

Here, we briefly recall the scoring of `PETfold`, which is a maximum expected accuracy scoring method. For simplicity, we will exclude a description of the scoring of single-stranded positions. However, they are scored the same way as in the original `PETfold` approach; for more details, see [25].

The `PETfold` score is the sum of the evolutionary accuracy values plus the average sum of the thermodynamic accuracy values. For the evolutionary part, we compute the expected accuracy (or overlap) $\mathrm{EA}^{ev}(\sigma)$ of a specific consensus structure σ with all possible consensus structures, which are weighted according to their probabilities:

$$\mathrm{EA}^{ev}(\sigma) = \sum_{\sigma'} |\sigma \cap \sigma'| \times \Pr^{ev}[\sigma' \mid \mathcal{A}]. \tag{1}$$

Recall that $\Pr^{ev}[\sigma'|\mathcal{A}]$ denotes the evolutionary probability of a structure σ' according to the `Pfold` SCFG as described above. $|\sigma \cap \sigma'|$ is the number of base pairs that are common between σ and σ' and thus denotes the overlap between these two structures.

For the thermodynamic part, the expected accuracy $\mathrm{EA}^{th}(\sigma)$ of σ with all structures for all sequences according to the thermodynamic ensembles is defined by

$$\mathrm{EA}^{th}(\sigma) = \sum_{s} \sum_{\sigma'\in\mathcal{S}(s)} |\sigma \cap \sigma'| \times \Pr^{th}[\sigma'(s, \mathcal{A}) \mid s]. \tag{2}$$

The combined expected accuracy consists of both parts, generally weighted with 1 for the conservation portion and β for the thermodynamic accuracy:

$$\mathrm{EA}(\sigma) = \mathrm{EA}^{\mathrm{ev}}(\sigma) + \frac{\beta}{n} \times \mathrm{EA}^{\mathrm{th}}(\sigma), \tag{3}$$

where n is the previously described number of sequences in the alignment. As shown previously [25], this final score can be calculated using the base pair reliabilities, where the combined reliability $\Re_{\mathrm{bp}}(i, j)$ for a base pair (i, j) is given by

$$\begin{aligned}\mathcal{R}_{\mathrm{bp}}(i,j) &= \sum_{\substack{\sigma' \text{ with} \\ (i,j)\in\sigma'}} 1 \times \Pr^{\mathrm{ev}}[\sigma' \mid \mathcal{A}] \\ &\quad + \frac{\beta}{n} \times \sum_{s} \sum_{\substack{\sigma' \text{ with} \\ (i,j)\in\sigma'}} 1 \times \Pr^{\mathrm{th}}[\sigma'(s,\mathcal{A}) \mid s] \\ &= \mathcal{R}^{\mathrm{ev}}_{\mathrm{bp}}(i,j,A) + \frac{\beta}{n} \times \sum_{s} \Pr^{\mathrm{th}}_{\mathrm{bp}}(i,j,s),\end{aligned} \tag{4}$$

where $\Pr^{\mathrm{th}}_{\mathrm{bp}}(i, j, s)$ is the base pair probability of the pair (k, l) associated with columns (i, j) in sequence s. These reliabilities are calculated with an inside/outside algorithm and are central to the hierarchical approach presented in the following sections. The expected accuracy can then be calculated from the base pair reliabilities by

$$\mathrm{EA}(\sigma) = \sum_{(i,j)\in\sigma} \mathcal{R}_{\mathrm{bp}}(i,j). \tag{5}$$

The consensus structure with the maximal reliability is then calculated using a Nussinov-style algorithm [30], where the base pairs are evaluated with reliabilities.

Step 1: Intra-molecular partial structures

We use two alignments $\mathcal{A}_1$ and $\mathcal{A}_2$ of sequences $s_1^1 \ldots s_1^n$ and $s_2^1 \ldots s_2^m$, where s_1^k is a ncRNA and s_2^k is its target sequence. For convenience, we adopt the convention of `RNAcofold` and assume that the positions in s_1^k are numbered $1 \le i \le |\, s_1^k \,|$ and the positions in s_2^k are numbered $|\, s_1^k \,| + 1 \le i \le |\, s_1^k \,| + |\, s_2^k \,|$.

Selection of the initial structure

In the first step of the pipeline, we obtain the base pair reliabilities from Equation (4), which we denote $\mathcal{R}^1_{\mathrm{bp}}(i, j)$. Using these reliabilities, the partial (constrained) structures σ_1^{p} and σ_2^{p} are determined independently for $\mathcal{A}_1$ and $\mathcal{A}_2$. In the following steps, let $\mathcal{A}$ be either $\mathcal{A}_1$ or $\mathcal{A}_2$ and σ^{p} be the partial structure calculated for $\mathcal{A}$. This is done by selecting only base pairs (i, j) with

$$\mathcal{R}^1_{\mathrm{bp}}(i,j) \ge \delta,$$

where δ is a cut-off that must be ≥ 0.5 to avoid crossing structures. This is similar to the method by which consensus structures are predicted for single sequences [31] and has been shown to be more reliable for the prediction of consensus structures from alignments.

Here, however, we also have to estimate the contribution of each of the partial structures to the complete solution. Because the set of base pairs from a predicted consensus structure do not necessarily form a reasonable structure, we account for this by introducing a second threshold γ. High values for this threshold guarantee that each sequence used to create the consensus structure has a high likelihood and that the approximation, which we apply in the second step (as will be described by Equation (14)), is accurate.

To find the optimal value of the reliability threshold δ, its value is increased until the resulting ensemble of structures $\mathcal{E}(\sigma^{\mathrm{p}})$ that are compatible with the partial structure σ^{p} is probable enough in the evolutionary model, in the thermodynamic model, or in both models, which is when

$$\Pr^{\mathrm{ev}}[\mathcal{E}(\sigma^{\mathrm{p}}) \mid \mathcal{A}] \ge \gamma \text{ or } \frac{1}{n} \sum_{s\in\mathcal{A}} \Pr^{\mathrm{th}}[\mathcal{E}(\sigma^{\mathrm{p}}(s,\mathcal{A})) \mid s] \ge \gamma.$$

Here, $\Pr^{\mathrm{ev}}[\mathcal{E}(\sigma^{\mathrm{p}}) \mid \mathcal{A}]$ (= $\Pr^{\mathrm{ev}}[\mathcal{E}(\sigma^{\mathrm{p}}) \mid \mathcal{A}, T, M]$) is the probability of the partial structure σ^{p} given the alignment $\mathcal{A}$ in the evolutionary model M and tree T. This can be calculated from `Pfold` with the SCFG that combines the prior structural model with evolutionary information from the alignment (see additional file 1) as follows:

$$\begin{aligned}&\Pr^{\mathrm{ev}}[\mathcal{E}(\sigma^{\mathrm{p}}) \mid \mathcal{A}, T, M] \\ &= \frac{\Pr[\mathcal{E}(\sigma^{\mathrm{p}}) \mid \mathcal{A}, T, M] \times \Pr[\mathcal{A} \mid T, M]}{1 \times \Pr[\mathcal{A} \mid T, M]} \\ &= \frac{\Pr[\mathcal{E}(\sigma^{\mathrm{p}}), \mathcal{A} \mid T, M]}{\Pr[\mathcal{A} \mid T, M]}\end{aligned} \tag{6}$$

The term $\Pr[\mathcal{A} \mid T, M]$ has already been calculated (personal communication with Bjarne Knudsen) in `Pfold` as

the sum of all possible parse trees for an alignment $\mathcal{A}$, given T, M:

$$\Pr[\mathcal{A} \mid T, M] = \sum_{\sigma} \Pr[\mathcal{A}, \sigma \mid T, M].$$

Here, we add the calculation of

$$\Pr[\mathcal{E}(\sigma^{\mathrm{p}}), \mathcal{A} \mid T, M] = \sum_{\sigma \in \mathcal{E}(\sigma^{\mathrm{p}})} \Pr[\mathcal{A}, \sigma \mid T, M]$$

to `Pfold` by summing over all possible parse trees that are compatible with σ.

$\Pr^{\mathrm{th}}[\mathcal{E}(\sigma^{\mathrm{p}}(s, \mathcal{A}))|s]$ is the probability of the partial structure σ^{p} given a sequence s in the thermodynamic model. This probability can be calculated using constrained partition folding as follows:

$$\Pr^{\mathrm{th}}[\mathcal{E}(\sigma^{\mathrm{p}}(s,\mathcal{A})) \mid s] = \frac{e^{\frac{-E^s_{\mathcal{E}(\sigma^{\mathrm{p}}(s,\mathcal{A}))}}{RT}}}{e^{\frac{-E^s_{all}}{RT}}} = e^{\frac{E^s_{all} - E^s_{\mathcal{E}(\sigma^{\mathrm{p}}(s,\mathcal{A}))}}{RT}}, \quad (7)$$

where E^s_{all} is the free energy of the whole ensemble (as determined by `RNAfold` with parameters `-p -d2`) and $E^s_{\mathcal{E}(\sigma^{\mathrm{p}}(s,\mathcal{A}))}$ is the free energy of the ensemble of structures $?(\sigma^{\mathrm{p}}(s, \mathcal{A}))$ with the base pairs in $\sigma^{\mathrm{p}}(s, \mathcal{A})$ as constraints, which can be calculated by `RNAfold` with parameters `-C -p -d2`.

Extension of constrained stems

Reliable *intra*-molecular base pairs are constrained as single-stranded in Step 2 of the algorithm because we are interested in pseudoknots of the concatenated sequence and the interactions in these induced loop regions. The drawback of this *ansatz* is that *intra*-molecular stems get instable because of intermediate unbased constraints. Thus, we may get incomplete stems. To deal with this problem, we extend the constrained stems. Inner and outer base pairs are added as long as the average reliability of the inner or outer extended stem, respectively, is larger than the threshold δ, and the probability of the partial structure is greater than or equal to γ either in the evolutionary or the thermodynamic model. That is, the average reliability of the total, extended stem has to be larger than a threshold.

Step 1 is summarised as pseudocode in Figure 1.

for Alignment $\mathcal{A}_1$, $\mathcal{A}_2$ **do**
  calculate tree T,
  phylogenetic reliabilities $\mathcal{R}^{1,\mathrm{ev}}$,
  thermodynamic probabilities $\Pr^{1,\mathrm{th}}$
  $\Longrightarrow \mathcal{R}^1_{\mathrm{bp}}(i,j), \mathcal{R}^1_{\mathrm{ss}}(i) \leftarrow$ PETfold model
  repeat
    for all (i,j) **do**
      if $\mathcal{R}^1_{\mathrm{bp}}(i,j) \geq \delta$ **then**
        add base pair (i,j) to σ^{p}
      end if
    end for
    calculate partial structure probabilities
      $\Pr^{\mathrm{ev}}[\mathcal{E}(\sigma^{\mathrm{p}})|\mathcal{A}]$ and $\Pr^{\mathrm{th}}[\mathcal{E}(\sigma^{\mathrm{p}})|s]$
    increase δ
  until
    $\Pr^{\mathrm{ev}}[\mathcal{E}(\sigma^{\mathrm{p}})|\mathcal{A}] \geq \gamma \;||\; \frac{1}{n}\sum_s \Pr^{\mathrm{th}}[\mathcal{E}(\sigma^{\mathrm{p}})|s] \geq \gamma$
  for all stem $\mathcal{S} \subset \sigma^{\mathrm{p}}$ **do**
    for *adjacent* = (inner, outer) **do**
      repeat
        $b = adjacent$ base pair of $\mathcal{S}$
        $S_{old} = \mathcal{S}$; $\mathcal{S} = \mathcal{S} \cup \{b\}$
        $\sigma^{\mathrm{p}}_{old} = \sigma^{\mathrm{p}}$; $\sigma^{\mathrm{p}} = \sigma^{\mathrm{p}} \cup \mathcal{S}$
        calculate $\Pr^{\mathrm{ev}}[\mathcal{E}(\sigma^{\mathrm{p}})|\mathcal{A}]$, $\Pr^{\mathrm{th}}[\mathcal{E}(\sigma^{\mathrm{p}})|s]$
      until average $\mathcal{R}^1_{\mathrm{bp}}$ of $\mathcal{S} < \delta$ $||$
    $(\Pr^{\mathrm{ev}}[\mathcal{E}(\sigma^{\mathrm{p}})|\mathcal{A}] < \gamma$ && $\frac{1}{n}\sum_s \Pr^{\mathrm{th}}[\mathcal{E}(\sigma^{\mathrm{p}})|s] < \gamma)$
      $S = \mathcal{S}_{old}$; $\sigma^{\mathrm{p}} = \sigma^{\mathrm{p}}_{old}$
    end for
  end for
end for

Figure 1 Pseudocode for Step 1.

Step 2: Constrained expected accuracy scoring

In the following, $s_1 \& s_2$ denote the concatenated sequences of the two sequences s_1, s_2 using the additional linker symbol & as done in `RNAcofold`. For Step 2 of the scoring, we calculate the expected accuracy of the ensemble of structures σ of $s_1 \& s_2$, which constitutes an interaction under the constraint that σ contains the partial reliable structures σ^{p}_1 and σ^{p}_2 of s_1 and s_2, respectively. Because we use the numbering convention of `RNAcofold`, the union $\sigma^{\mathrm{p}}_1 \cup \sigma^{\mathrm{p}}_2$ of the two partial structures σ^{p}_1 and σ^{p}_2 is the partial structure of $s_1 \& s_2$.

Now we have two problems to solve. On the one hand, we want to calculate the constrained accuracy given the partial structures σ^{p}_1 and σ^{p}_2, which is defined as

$$\mathrm{EA}_{\sigma^{\mathrm{p}}_1, \sigma^{\mathrm{p}}_2}(\sigma) = \mathrm{EA}^{\mathrm{ev}}_{\sigma^{\mathrm{p}}_1, \sigma^{\mathrm{p}}_2}(\sigma) + \frac{\beta}{n} \times \mathrm{EA}^{\mathrm{th}}_{\sigma^{\mathrm{p}}_1, \sigma^{\mathrm{p}}_2}(\sigma). \quad (8)$$

On the other hand, we have to find a combined score for the partial structures σ_1^{p} and σ_2^{p}, and the interaction σ_{int} to evaluate the quality of an predicted interaction. The score must be maximal according to Equation (8).

We will demonstrate the problem and our solution for the thermodynamic folding. However, the same analysis applies to the evolutionary part, which is described later.

The thermodynamic part

The simplest formal solution to this problem would be to investigate directly the expected accuracy of joint structures σ:

$$\begin{aligned}\mathrm{EA}^{\mathrm{th}}(\sigma) &= \sum_{s_1 \& s_2} \sum_{\sigma'} |\, \sigma(s_1 \,\&\, s_2, \mathcal{A}) \cap \sigma' \,| \times \Pr^{\mathrm{th}}[\sigma' \mid s_1 \,\&\, s_2] \\ &= \sum_{s_1 \& s_2 \in \mathcal{A}} \mathrm{EA}^{\mathrm{th}}_{s_1 \& s_2}(\sigma(s_1 \,\&\, s_2, \mathcal{A})),\end{aligned}$$

where $\mathrm{EA}^{\mathrm{th}}_{s_1 \& s_2}(\sigma)$ is the expected accuracy of a structure in one sequence pair $s_1 \& s_2 \quad \mathcal{A}$.

However, this would require that we compute the distribution $\Pr^{\mathrm{th}}[\sigma|s_1 \& s_2]$, which can be done by a partition function approach for interacting structures. This is NP-complete in the full model [18] and even $O(n^6)$ in a restricted model [19,20], which is why the two-step approach is necessary. In the following, we ignore the index "th" for simplicity.

The relationship between $\mathrm{EA}_{\sigma_1^{\mathrm{p}}, \sigma_2^{\mathrm{p}}}(\sigma)$ and $\mathrm{EA}(\sigma)$ is now quantified. In the following, for a structure σ, we use $\sigma_1 \cup \sigma_2 \cup \sigma_{\mathrm{int}}$ to denote the partition of the base pairs of the first sequence, σ_1, the base pairs of the second sequence, σ_2, and the interacting base pairs, σ_{int}. Furthermore, for the partial structure σ, we use $\varepsilon_1(\sigma)$ to denote the set of structures that extends σ using base pairs within the first sequence, *i.e.*,

$$\begin{aligned}\mathcal{E}_1(\sigma) &= \{\sigma' \supseteq \sigma \mid \forall (i,j) \in \sigma' \setminus \sigma : 1 \le i < j \le |\, s_1 \,|\} \\ \mathcal{E}_2(\sigma) &= \{\sigma' \supseteq \sigma \mid \forall (i,j) \in \sigma' \setminus \sigma : \\ &\qquad |\, s_1 \,| + 1 \le i < j \le |\, s_2 \,|\} \\ \mathcal{E}_{\mathrm{int}}(\sigma) &= \{\sigma' \supseteq \sigma \mid \forall (i,j) \in \sigma' \setminus \sigma : \\ &\qquad 1 \le i \le |\, s_1 \,| \wedge |\, s_1 \,| + 1 \le j \le |\, s_2 \,|\}\end{aligned}$$

The ensembles $\varepsilon_{1,\mathrm{int}}(\sigma)$, $\varepsilon_{2,\mathrm{int}}(\sigma)$ and $\varepsilon_{1,2}(\sigma)$ are defined analogously.

Our approach uses one simplification, namely the assumption that the reliabilities for *intra*-molecular base pairs are dominated by the *intra*-molecular folding. This is equivalent to the assumption that the two structures fold independently. We formulate this as follows:

$$\Pr[\sigma_1 \mid \sigma_2, s_1 \,\&\, s_2] = \Pr[\sigma_1 \mid s_1].$$

Because σ_1 and σ_2 are *partial joint* structures, this can be written using the ensemble function

$$\Pr[\mathcal{E}_{2,\mathrm{int}}(\sigma_1) \mid \mathcal{E}_{1,\mathrm{int}}(\sigma_2), s_1 \,\&\, s_2] = \Pr[\sigma_1 \mid s_1]. \quad (9)$$

The implication of this assumption is that the probabilities of the two structures σ_1 and σ_2 are merged independently into the joint probability $\Pr[\varepsilon_{\mathrm{int}}(\sigma_1 \cup \sigma_2)|s_1 \& s_2]$, see Equation (11) below. First, note that for two partial structures

$$\Pr[\mathcal{E}(\sigma^{\mathrm{p}} \cup \sigma^{\mathrm{p}'}) \mid s] = \Pr[\mathcal{E}(\sigma^{\mathrm{p}}) \wedge \mathcal{E}(\sigma^{\mathrm{p}'}) \mid s],$$

by definition. Hence,

$$\begin{aligned}&\Pr[\mathcal{E}_{\mathrm{int}}(\sigma_1 \cup \sigma_2) \mid s_1 \,\&\, s_2] \\ &\quad = \Pr[\mathcal{E}_{2,\mathrm{int}}(\sigma_1) \wedge \mathcal{E}_{1,\mathrm{int}}(\sigma_2) \mid s_1 \,\&\, s_2] \\ &\quad = \Pr[\mathcal{E}_{2,\mathrm{int}}(\sigma_1) \mid \mathcal{E}_{1,\mathrm{int}}(\sigma_2), s_1 \,\&\, s_2] \\ &\qquad \times \Pr[\mathcal{E}_{1,int}(\sigma_2) \mid s_1 \,\&\, s_2] \\ &\quad \overset{Eq.(9)}{=} \Pr[\sigma_1 \mid s_1] \times \Pr[\mathcal{E}_{1,\mathrm{int}}(\sigma_2) \mid s_1 \,\&\, s_2].\end{aligned}$$

Intuitively, $\Pr[\varepsilon_{1,\mathrm{int}}(\sigma_2)|s_1 \& s_2]$ should be the same as $\Pr[\sigma_2|s_2]$. This can be derived using the total probability formula:

$$\begin{aligned}&\Pr[\mathcal{E}_{1,int}(\sigma_2) \mid s_1 \,\&\, s_2] \\ &\quad = \sum_{\sigma_1} \begin{pmatrix} \Pr[\mathcal{E}_{1,int}(\sigma_2) \mid \mathcal{E}_{2,int}(\sigma_1), s_1 \,\&\, s_2] \\ \times \Pr[\mathcal{E}_{2,int}(\sigma_1) \mid s_1 \,\&\, s_2] \end{pmatrix} \\ &\quad \overset{Eq.(9)}{=} \sum_{\sigma_1} \Pr[\sigma_2 \mid s_2] \times \Pr[\mathcal{E}_{2,int}(\sigma_1) \mid s_1 \,\&\, s_2] \\ &\quad = \Pr[\sigma_2 \mid s_2] \times \sum_{\sigma 1} \Pr[\mathcal{E}_{2,int}(\sigma_1) \mid s_1 \,\&\, s_2] \\ &\quad = \Pr[\sigma_2 \mid s_2] \times 1\end{aligned} \quad (10)$$

Combining these equations we obtain the independence property:

$$\Pr[\mathcal{E}_{\mathrm{int}}(\sigma_1 \cup \sigma_2) \mid s_1 \,\&\, s_2] = \Pr[\sigma_1 \mid s_1] \times \Pr[\sigma_2 \mid s_2 \quad (11)$$

Now we will use this property to relate $\mathrm{EA}_{\sigma_1^{\mathrm{p}},\sigma_2^{\mathrm{p}}}(\sigma)$ to $\mathrm{EA}(\sigma)$. The independence property, as described in Equation (9), and the additivity of the expectation is the implication of the expected accuracy of a joint structure, which is the sum of the expected accuracy of the *intra*-molecular structures and the expected accuracy of the *inter*-molecular portion. To illustrate this, note that for any σ, σ'

$$|\sigma \cap \sigma'| = |\sigma_1 \cap \sigma_1'| + |\sigma_2 \cap \sigma_2'| + |\sigma_{\mathrm{int}} \cap \sigma_{\mathrm{int}}'|$$

by definition. Hence, by the additivity of the expectation we get

$$\begin{aligned}\mathrm{EA}^{\mathrm{th}}_{s_1\&s_2}(\sigma) &= \sum_{\sigma'} |\sigma \cap \sigma'| \times \Pr^{\mathrm{th}}[\sigma' \mid s_1 \,\&\, s_2] \\ &= \sum_{\sigma'} |\sigma_1 \cap \sigma_1'| \times \Pr[\sigma' \mid s_1 \,\&\, s_2] \\ &+ \sum_{\sigma'} |\sigma_2 \cap \sigma_2'| \times \Pr[\sigma' \mid s_1 \,\&\, s_2] \\ &+ \sum_{\sigma'} |\sigma_{\mathrm{int}} \cap \sigma_{\mathrm{int}}'| \times \Pr[\sigma' \mid s_1 \,\&\, s_2].\end{aligned}$$

Now we can rewrite the first term $\sum_{\sigma'} |\sigma_1 \cap \sigma_1'| \times \Pr[\sigma' \mid s_1 \,\&\, s_2]$ using the independence property as follows:

$$\begin{aligned}&\sum_{\sigma'} |\sigma_1 \cap \sigma_1'| \times \Pr[\sigma' \mid s_1 \,\&\, s_2] \\ &= \sum_{\sigma_1'} \sum_{\sigma_2', \sigma_{\mathrm{int}}'} |\sigma_1 \cap \sigma_1'| \times \Pr[\sigma_1' \cup \sigma_2' \cup \sigma_{\mathrm{int}}' \mid s_1 \,\&\, s_2] \\ &= \sum_{\sigma_1'} |\sigma_1 \cap \sigma_1'| \times \Pr[\mathcal{E}_{2,\mathrm{int}}(\sigma_1') \mid s_1 \,\&\, s_2] \\ &\overset{Eq.(10)}{=} \sum_{\sigma_1'} |\sigma_1 \cap \sigma_1'| \times \Pr[\sigma_1' \mid s_1],\end{aligned}$$

which is the expected accuracy of σ_1 in the sequence s_1. Analogously, we can do this for the second term $\sum_{\sigma'} |\sigma_2 \cap \sigma_2'| \times \Pr[\sigma' \mid s_1 \,\&\, s_2]$. Thus, $\mathrm{EA}^{\mathrm{th}}_{s_1\&s_2}(\sigma)$ is the sum of the expected accuracies in the first and the second sequences and the expected accuracy of the interaction:

$$\begin{aligned}\mathrm{EA}^{\mathrm{th}}_{s_1\&s_2}(\sigma) &= \mathrm{EA}^{\mathrm{th}}_{s_1}(\sigma_1) + \mathrm{EA}^{\mathrm{th}}_{s_2}(\sigma_2) \\ &+ \sum_{\sigma'} |\sigma_{\mathrm{int}} \cap \sigma_{\mathrm{int}}'| \times \Pr[\sigma' \mid s_1 \,\&\, s_2].\end{aligned} \tag{12}$$

For the expected accuracy of the interaction

$$\mathrm{EA}^{\mathrm{th,int}}(\sigma) = \sum_{\sigma'} |\sigma_{\mathrm{int}} \cap \sigma_{\mathrm{int}}'| \times \Pr[\sigma' \mid s_1 \,\&\, s_2] \tag{13}$$

we still need to define $\Pr[\sigma | s_1 \& s_2]$. For every $\sigma = \sigma_1 \quad \sigma_2 \quad \sigma_{\mathrm{int}}$,

$$\begin{aligned}&\Pr[\sigma_1 \cup \sigma_2 \cup \sigma_{\mathrm{int}} \mid s_1 \,\&\, s_2] \\ &\quad = \Pr[\mathcal{E}_{\mathrm{int}}(\sigma_1 \cup \sigma_2) \wedge \mathcal{E}_{1,2}(\sigma_{\mathrm{int}}) \mid s_1 \,\&\, s_2] \\ &\quad = \Pr[\mathcal{E}_{1,2}(\sigma_{\mathrm{int}}) \mid \mathcal{E}_{\mathrm{int}}(\sigma_1 \cup \sigma_2), s_1 \,\&\, s_2] \\ &\quad \times \Pr[\mathcal{E}_{\mathrm{int}}(\sigma_1 \cup \sigma_2) \mid s_1 \times s_2] \\ &\quad \overset{Eq.(11)}{=} \Pr[\mathcal{E}_{1,2}(\sigma_{\mathrm{int}}) \mid \mathcal{E}_{\mathrm{int}}(\sigma_1 \cup \sigma_2), s_1 \,\&\, s_2] \\ &\qquad \times \Pr[\sigma_1 \mid s_1] \times \Pr[\sigma_2 \mid s_2]\end{aligned}$$

Thus, in principle, to calculate the expected accuracy $\mathrm{EA}^{\mathrm{th,\,int}}(\sigma)$ for the interaction, we must sum over all structures in σ_1 and σ_2:

$$\begin{aligned}&\mathrm{EA}^{\mathrm{th,int}}(\sigma) \\ &= \sum_{\sigma_{\mathrm{int}}'} \sum_{\sigma_1', \sigma_2'} |\sigma_{\mathrm{int}} \cap \sigma_{\mathrm{int}}'| \times \Pr[\sigma_{\mathrm{int}}' \cup \sigma_1' \cup \sigma_2' \mid s_1 \,\&\, s_2] \\ &= \sum_{\sigma_{\mathrm{int}}'} |\sigma_{\mathrm{int}} \cap \sigma_{\mathrm{int}}'| \times \sum_{\sigma_1', \sigma_2'} \Pr[\sigma_{\mathrm{int}}' \cup \sigma_1' \cup \sigma_2' \mid s_1 \,\&\, s_2]\end{aligned}$$

Because this is not feasible, we restrict ourselves to an ensemble of structures. Thus, instead of summing over all possible σ_1 and σ_2, we use the partial structures σ_1^{p} and σ_2^{p} that were determined in the first step and approximate $\mathrm{EA}^{\mathrm{th,int}}(\sigma)$ by

$$\begin{aligned}&\mathrm{EA}^{\mathrm{th,int'}}(\sigma) \\ &= \sum_{\sigma_{\mathrm{int}}'} |\sigma_{\mathrm{int}} \cap \sigma_{\mathrm{int}}'| \times \sum_{\substack{\sigma_1' \in \mathcal{E}(\sigma_1^{\mathrm{p}}) \\ \sigma_2' \in \mathcal{E}(\sigma_2^{\mathrm{p}})}} \Pr[\sigma_{\mathrm{int}}' \cup \sigma_1' \cup \sigma_2' \mid s_1 \,\&\, s_2]\end{aligned}$$

The second sum can now be simplified as follows:

$$\begin{aligned}
&\sum_{\substack{\sigma'_1 \in \mathcal{E}(\sigma_1^{\mathrm{p}}) \\ \sigma'_2 \in \mathcal{E}(\sigma_2^{\mathrm{p}})}} \Pr[\sigma'_{\mathrm{int}} \cup \sigma'_1 \cup \sigma'_2 \mid s_1 \,\&\, s_2] \\
&= \Pr[\mathcal{E}_{1,2}(\sigma'_{\mathrm{int}} \cup \sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}) \mid s_1 \,\&\, s_2] \\
&= \Pr[\mathcal{E}_{1,2}(\sigma'_{\mathrm{int}}) \wedge \mathcal{E}(\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}) \mid s_1 \,\&\, s_2] \\
&= \Pr[\mathcal{E}_{1,2}(\sigma'_{\mathrm{int}}) \mid \mathcal{E}(\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}), s_1 \,\&\, s_2] \\
&\quad \times \Pr[\mathcal{E}(\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}) \mid s_1 \,\&\, s_2] \\
&\overset{Eq.(11')}{=} \Pr[\mathcal{E}_{1,2}(\sigma'_{\mathrm{int}}) \mid \mathcal{E}(\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}), s_1 \,\&\, s_2] \\
&\qquad \times \Pr[\mathcal{E}_1(\sigma_1^{\mathrm{p}}) \mid s_1] \times \Pr[\mathcal{E}_2(\sigma_2^{\mathrm{p}}) \mid s_2],
\end{aligned}$$

where Equation (11') indicates the variation of the independence assumption of Equation (11) for the structure ensembles (see additional file 1). Thus, we finally have

$$\begin{aligned}
&\mathrm{EA}^{\mathrm{th,int'}}(\sigma) \\
&= \sum_{\sigma'_{\mathrm{int}}} \begin{pmatrix} |\sigma_{\mathrm{int}} \cap \sigma'_{\mathrm{int}}| \times \Pr[\mathcal{E}_1(\sigma_1^{\mathrm{p}}) \mid s_1] \times \Pr[\mathcal{E}_2(\sigma_2^{\mathrm{p}}) \\ \times \Pr[\mathcal{E}_{1,2}(\sigma'_{\mathrm{int}}) \mid \mathcal{E}(\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}), s_1 \,\&\, s_2] \end{pmatrix}
\end{aligned} \tag{14}$$

Now $\Pr[\mathcal{E}_{1,2}(\sigma'_{\mathrm{int}}) \mid \mathcal{E}(\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}), s_1 \,\&\, s_2]$ is the constrained folding, where the positions covered by σ_1^{p} and σ_2^{p} are fixed. However, we have the problem that these structures might contain pseudoknots. Recall that the positions in σ_1^{p} and σ_2^{p} are fixed for folding and that we are considering all structures σ that contain $\sigma_1^{\mathrm{p}} \cup \sigma_1^{\mathrm{p}}$ and are nested on $\sigma_{\mathrm{int}} = \sigma \setminus (\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}})$. Technically, we solve the problem using the fact that the set of structures that is nested on σ_{int} *and* compatible with $\sigma_1^{\mathrm{p}} \cup \sigma_1^{\mathrm{p}}$ is selected by considering all structures where the positions of $\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}$ are constrained as single-stranded. This implies that we use constrained cofolding via `RNAcofold` (parameters `-C -p -d2`), and the constraint ... x_1x_1 ... & ... x_2x_2 ..., where x_1 (resp. x_2) denotes a position from σ_1^{p} (resp. σ_2^{p}) that is constrained as single-stranded. The main difference is that the energy contributions could be slightly different, and therefore, we obtain only an approximation of the real distribution. For example, an extension of a helix in σ_1^{p} would be evaluated as an internal loop or hairpin. Note that this is not a major problem because we are mainly interested in the *inter*-molecular base pairs between s_1 and s_2 in this step. However, the recursion scheme of `RNAcofold` could easily be adapted to use new symbols for base pair constraints and a scoring scheme that is common to hierarchical approaches of pseudoknot structure prediction, which would avoid these problems.

Finally, we can rewrite the thermodynamic accuracy as the sum of probabilities as indicated in Equation (5). As shown in Equation (12), for a base pair $(i, j) \in \sigma_\ell^{\mathrm{p}}$ ($\ell = 1, 2$), we want to use the probability of the associated sequence. To avoid competition with the probabilities for the *intra*-molecular base pairs calculated from `RNAcofold`, we set all of these base pairs to the same probability $\mathsf{Pr}^{\mathsf{th}}[\mathcal{E}_\ell(\sigma_\ell^{\mathrm{p}}(\mathsf{s}_\ell, \mathcal{A})) \mid \mathsf{s}_\ell]$ as described in Equation (7). For the *inter*-molecular base pairs, we use the base pair probabilities as provided by `RNAcofold` with constraints, which model $\Pr[\sigma' \mid \mathcal{E}(\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}) \wedge s_1 \,\&\, s_2]$ from the constrained cofolding. However, these raw base pair probabilities (in the following denoted by $\Pr_{\mathrm{bp,raw}}^{2,\mathrm{th}}(i, j)$) are calculated under the constraint of $\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}}$ and have therefore (to obtain the final base pair probabilities) to be multiplied by $\Pr[\mathcal{E}_1(\sigma_1^{\mathrm{p}}) \mid s_1] \times \Pr[\mathcal{E}_2(\sigma_2^{\mathrm{p}}) \mid s_2]$ as indicated by Equation (14). Thus, we can score each base pair as follows:

$$\begin{aligned}
&\Pr_{\mathrm{bp}}^{2,\mathrm{th}}(i, j, s_1 \,\&\, s_2) \\
&= \begin{cases} \mathbf{1} \times \Pr_{\mathrm{bp}}^{\mathrm{th}}(i, j, s_1) & \text{if } 1 \le i, j \le |s_1| \\ \mathbf{1} \times \Pr_{\mathrm{bp}}^{\mathrm{th}}(i, j, s_2) & \text{if } |s_1| + 1 \le i, j \le |s_2| \\ \Pr_{\mathrm{bp,raw}}^{2,\mathrm{th}}(i, j) \times \prod_{\ell=1,2} \Pr^{\mathrm{th}}[\mathcal{E}_\ell(\sigma_\ell^{\mathrm{p}}(s_\ell, \mathcal{A})) \mid s_\ell] \end{cases}
\end{aligned} \tag{15}$$

where the 1 reflects the fixed reliability. However, we deviate from this scoring to weaken the independence assumption for the *intra*-molecular base pairs, which allows us to determine new *intra*-molecular base pairs from the constrained application of RNA-cofold. Thus, we score only the base pairs from the partial structures σ_1^{p} and σ_2^{p} with the probability in the associated sequence. In addition, to avoid competition with the probabilities for these base pairs calculated from `RNAcofold`, we simply set all of these base pairs to the same probability $\mathsf{Pr}^{\mathsf{th}}[\mathcal{E}_\ell(\sigma_\ell^{\mathrm{p}}(\mathsf{s}_\ell, \mathcal{A})) \mid \mathsf{s}_\ell]$ as described in Equation (7). To summarise, given the partial consensus structures σ_1^{p} and σ_2^{p} for an alignment $\mathcal{A}_1$ & $\mathcal{A}_2$ as cal-

culated in Step 1, the probability for a base pair (i, j) in sequence $s_1 \& s_2 \in \mathcal{A}_1 \& \mathcal{A}_2$ in the second step is:

$$\Pr_{\text{bp}}^{2,\text{th}}(i, j, s_1 \,\&\, s_2) = \begin{cases} 1 \times \Pr^{\text{th}}[\mathcal{E}_1(\sigma_1^{\text{p}}(s_1, \mathcal{A})) \mid s_1] & \text{if } (i,j) \in \sigma_1^{\text{p}} \\ 1 \times \Pr^{\text{th}}[\mathcal{E}_2(\sigma_2^{\text{p}}(s_2, \mathcal{A})) \mid s_2] & \text{if } (i,j) \in \sigma_2^{\text{p}} \\ \Pr_{\text{bp,raw}}^{2,\text{th}}(i,j) \times \prod_{\ell=1,2} \Pr^{\text{th}}[\mathcal{E}_\ell(\sigma_\ell^{\text{p}}(s_\ell, \mathcal{A})) \mid s_\ell] & \end{cases} \quad (16)$$

Single-stranded probabilities

Single-stranded probabilities are integrated in a similar way as the base pair probabilities, but with different weighting. The single-stranded probabilities are as follows:

$$\Pr_{\text{ss}}^{2,\text{th}}(i, s_1 \,\&\, s_2) = \begin{cases} 0 & \text{if } \exists j \text{ with } (i,j) \in \sigma_1^{\text{p}} \text{ or } (j,i) \in \sigma_1^{\text{p}} \\ 0 & \text{if } \exists j \text{ with } (i,j) \in \sigma_2^{\text{p}} \text{ or } (j,i) \in \sigma_2^{\text{p}} \\ \Pr_{\text{ss,raw}}^{2,\text{th}}(i) \times \prod_{\ell=1,2} \Pr^{\text{th}}[\mathcal{E}_\ell(\sigma_\ell^{\text{p}}(s_\ell, \mathcal{A})) \mid s_\ell] & \text{el} \end{cases} \quad (17)$$

Given the structure σ on an alignment $\mathcal{A}$ with m columns, the set of all single-stranded positions in the consensus structure is denoted as $\text{ss}(\sigma) = \{i \notin \sigma | 1 \le i \le m\}$. Taking this into consideration, the complete version of Equation (2) is

$$\text{EA}^{\text{th}}(\sigma) = \sum_{s,\sigma'} [\,|\, \sigma \cap \sigma' \,|\, + \alpha \,|\, \text{ss}(\sigma) \cap \text{ss}(\sigma') \,|\,] \times \Pr^{\text{th}}[\sigma'(s, \mathcal{A}) \mid s]$$

and the evolutionary accuracy is determined similarly. The combined score is the sum of the base pair reliabilities and single-stranded reliabilities (weighted with the parameter α). For details, see [25].

The evolutionary part

The calculation for the presented thermodynamic accuracy is purely based on constrained folding. To obtain the complete constrained folding, we use the same approach for the evolutionary accuracy by applying a version of `Pfold`[28] that incorporates the constraints. For that purpose, the raw structural reliabilities $\mathcal{R}_{\text{bp,raw}}^{2,\text{ev}}(i, j)$ and $\mathcal{R}_{\text{ss,raw}}^{2,\text{ev}}(i)$ are calculated by the constrained folding with `Pfold` using the phylogenetic tree deduced from the concatenated alignment. As a linker, three prior-free columns are inserted between both alignments. The evolutionary reliabilities $\mathcal{R}_{\text{bp}}^{2,\text{ev}}(i, j)$ for a base pair (i, j) and $\mathcal{R}_{\text{ss}}^{2,\text{ev}}(i)$ for a single-stranded position i are calculated in the same manner as $\Pr_{\text{bp}}^{2,\text{th}}(i, j, s_1 \,\&\, s_2)$ in Equation (16):

$$\mathcal{R}_{\text{bp}}^{2,\text{ev}}(i, j, \mathcal{A}_1 \,\&\, \mathcal{A}_2) = \begin{cases} 1 \times \Pr^{\text{ev}}[\mathcal{E}_1(\sigma_1^{\text{p}}) \mid \mathcal{A}_1] & \text{if } (i,j) \in \sigma_1^{\text{p}} \\ 1 \times \Pr^{\text{ev}}[\mathcal{E}_2(\sigma_2^{\text{p}}) \mid \mathcal{A}_2] & \text{if } (i,j) \in \sigma_2^{\text{p}} \\ \mathcal{R}_{\text{bp,raw}}^{2,\text{ev}}(i,j) \times \prod_{\ell=1,2} \Pr^{\text{ev}}[\mathcal{E}_\ell(\sigma_\ell^{\text{p}}) \mid \mathcal{A}_\ell] & \text{e} \end{cases} \quad (18)$$

as well as $\Pr_{\text{ss}}^{2,\text{th}}(i, s_1 \,\&\, s_2)$ in Equation (17):

$$\mathcal{R}_{\text{ss}}^{2,\text{ev}}(i, \mathcal{A}_1 \,\&\, \mathcal{A}_2) = \begin{cases} 0 & \text{if } \exists j \text{ with } (i,j) \in \sigma_1^{\text{p}} \text{ or } (j,i) \in \sigma_1^{\text{p}} \\ 0 & \text{if } \exists j \text{ with } (i,j) \in \sigma_2^{\text{p}} \text{ or } (j,i) \in \sigma_2^{\text{p}} \\ \mathcal{R}_{\text{ss,raw}}^{2,\text{ev}}(i) \times \prod_{\ell=1,2} \Pr^{\text{ev}}[\mathcal{E}_\ell(\sigma_\ell^{\text{p}}) \mid A_\ell] & \text{else} \end{cases} \quad (19)$$

The probabilities of the partial structures $\Pr^{\text{ev}}[\mathcal{E}_1(\sigma_1^{\text{p}}) \mid \mathcal{A}_1]$ and $\Pr^{\text{ev}}[\mathcal{E}_2(\sigma_2^{\text{p}}) \mid \mathcal{A}_2]$ are calculated

Input:
$\mathcal{A}_1 \& \mathcal{A}_2 = \text{concatenate}(\mathcal{A}_1, \mathcal{A}_2)$
$C_{ss} = \ldots x_1 x_1 \ldots \& \ldots x_2 x_2 \ldots$, where x's are single-stranded constraints and $x_1 \in \sigma_1^{\text{p}}$, $x_2 \in \sigma_2^{\text{p}}$
Search σ_{int} constrained by C_{ss}:
calculate tree $T_{\mathcal{A}_1 \& \mathcal{A}_2}$,
phylogenetic reliabilities $\mathcal{R}_{\text{raw}}^{2,\text{ev}}$,
thermodynamic probabilities $\Pr_{\text{raw}}^{2,\text{th}}$
for all (i, j) **do**
  if $(i, j) \in \sigma_\ell^{\text{p}}$ for ℓ=(1,2) **then**
    $\mathcal{R}_{\text{bp,raw}}^{2,\text{ev}}(i, j) \leftarrow \Pr^{\text{ev}}[\mathcal{E}_\ell(\sigma_\ell^{\text{p}}) | \mathcal{A}_\ell]$
    $\mathcal{R}_{\text{ss,raw}}^{2,\text{ev}}(i) \leftarrow 0$
    $\Pr_{\text{bp,raw}}^{2,\text{th}}(i, j) \leftarrow \Pr^{\text{th}}[\mathcal{E}_\ell(\sigma_\ell^{\text{p}}) | s_\ell]$
    $\Pr_{\text{ss,raw}}^{2,\text{th}}(i) \leftarrow 0$
  else
    $\mathcal{R}^{2,\text{ev}} \leftarrow \mathcal{R}_{\text{raw}}^{2,\text{ev}} \times \prod_{\ell=1,2} \Pr^{\text{ev}}[\sigma_\ell^{\text{p}} | \mathcal{A}_\ell]$
    $\mathcal{R}^{2,\text{th}} \leftarrow \Pr_{\text{raw}}^{2,\text{th}} \times \prod_{\ell=1,2} \Pr^{\text{th}}[\sigma_\ell^{\text{p}}(s_\ell, \mathcal{A}) | s_\ell]$
  end if
  $\Longrightarrow \mathcal{R}_{\text{bp}}^2(i, j), \mathcal{R}_{\text{ss}}^2(i) \leftarrow$ PETfold model
end for
$\sigma_{\text{int}} \leftarrow$ MEA-structure constrained by C_{ss}
Output:
$\sigma_1^{\text{p}} \cup \sigma_2^{\text{p}} \cup \sigma_{\text{int}}$

Figure 2 Pseudocode for Step 2.

as described in Equation (6). Step 2 is summarised as pseudocode in 1.

The final scoring

To summarise the reliabilities, a combined structure will be determined using the Nussinov algorithm on the following reliabilities:

$$\mathcal{R}^2_{\mathrm{bp}}(i,j) = \mathcal{R}^{2,\mathrm{ev}}_{\mathrm{bp}}(i,j,\mathcal{A}_1 \,\&\, \mathcal{A}_2) + \frac{\beta}{n}\sum_{s_1 \& s_2} \mathrm{Pr}^{2,\mathrm{th}}_{\mathrm{bp}}(i,j,s_1 \,\&\, s_2)$$

$$\mathcal{R}^2_{\mathrm{ss}}(i) = \mathcal{R}^{2,\mathrm{ev}}_{\mathrm{ss}}(i,\mathcal{A}_1 \,\&\, \mathcal{A}_2) + \frac{\beta}{n}\sum_{s_1 \& s_2} \mathrm{Pr}^{2,\mathrm{th}}_{\mathrm{ss}}(i,s_1 \,\&\, s_2),$$

where $\mathcal{R}^{2,\mathrm{ev}}_{\mathrm{bp}}(i,j,\mathcal{A}_1 \,\&\, \mathcal{A}_2)$ and $\mathcal{R}^{2,\mathrm{ev}}_{\mathrm{ss}}(i,\mathcal{A}_1 \,\&\, \mathcal{A}_2)$ are defined as above, $\mathrm{Pr}^{2,\mathrm{th}}_{\mathrm{bp}}(i,j,s_1 \,\&\, s_2)$ as in Equation (16) and $\mathrm{Pr}^{2,\mathrm{th}}_{\mathrm{ss}}(i,s_1 \,\&\, s_2)$ as in Equation (17).

Note that the base pairs in $\sigma^{\mathrm{p}}_1 \cup \sigma^{\mathrm{p}}_2$ have a weight of 0 during folding of the constrained structure to allow for pseudoknot formation. Finally, we add the base pairs in $\sigma^{\mathrm{p}}_1 \cup \sigma^{\mathrm{p}}_2$ to the constrained structure of Step 2. The flow of the structure reliabilities in the pipeline is summarised in Figure 3.

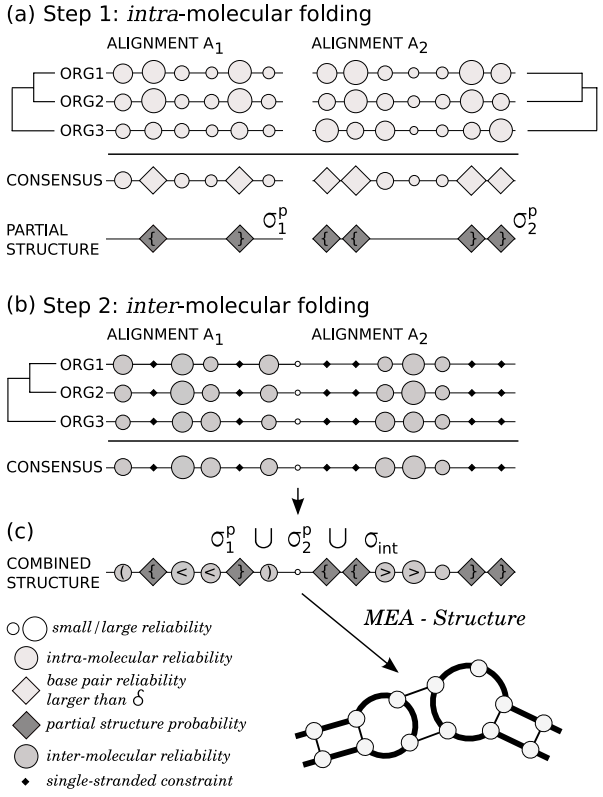

Figure 3 Scoring pipeline. The pipeline illustrates the flow of base pair probabilities during the structure scoring. The PETcofold pipeline consists of two steps: (a) *intra*-molecular folding by `PETfold` of both alignments and selection of a set of highly reliable base pairs that form the partial structures σ^{p}; (b) *inter*-molecular folding by an adaptation of `PETfold` of the concatenated alignments using the constraints from Step 1. In the end, (c) the partial structures and constrained *inter*-molecular structures are combined to generate the joint secondary structure including pseudoknots.

Results and discussion

The algorithm presented herein was implemented in `PETcofold` (Seemann *et al.*, submitted). As a proof of concept, we present an example of a bacterial sRNA-mRNA interaction. The in-depth analysis is described elsewhere (Seemann *et al.*, submitted).

Joint structure prediction of bacterial sRNA OxyS and its target mRNA fhlA

The small RNA OxyS represses the translation of the mRNA *fhlA*, which is mediated through base pairing at the ribosome binding site [11]. However, the OxyS-*fhlA* interaction involves a second binding site within the coding region of *fhlA*. Both interaction sites reside in stem loops such that OxyS and *fhlA* form a double kissing hairpin interaction.

Figure 4 shows the alignment and joint secondary structure prediction of the OxyS-*fhlA* complex, *i.e.*, the secondary structures of OxyS and *fhlA* and the interaction between them, as predicted by our algorithm. The result of the prediction without extending the constrained stems is shown in Figure 4a, and the result with the extension of the constrained stems is shown in Figure 4b.

For OxyS-*fhlA*, our algorithm was able to consistently predict one of the two interaction sites. The second interaction site, which is situated in the *fhlA* coding region, was only predicted when the constrained stems were not extended in Step 1 of our algorithm. Otherwise the stem of *fhlA* that resides the second interaction site was extended both by inner and outer base pairs. Consequently, the unpaired region of the hairpin containing the second interaction site became shorter such that no interaction was predicted at this site.

Algorithmic restrictions and potentials

The algorithm supports pseudoknots between the i*ntra*-molecular and *inter*-molecular base pairs, while the time complexity of $O(N \times I \times L^3)$ is much lower than that of other approaches with the same ability. The time complexity is in the magnitude of `PET-fold` for the added sequence length L of both alignments, and it is linear with respect to the number of sequences N in the alignments

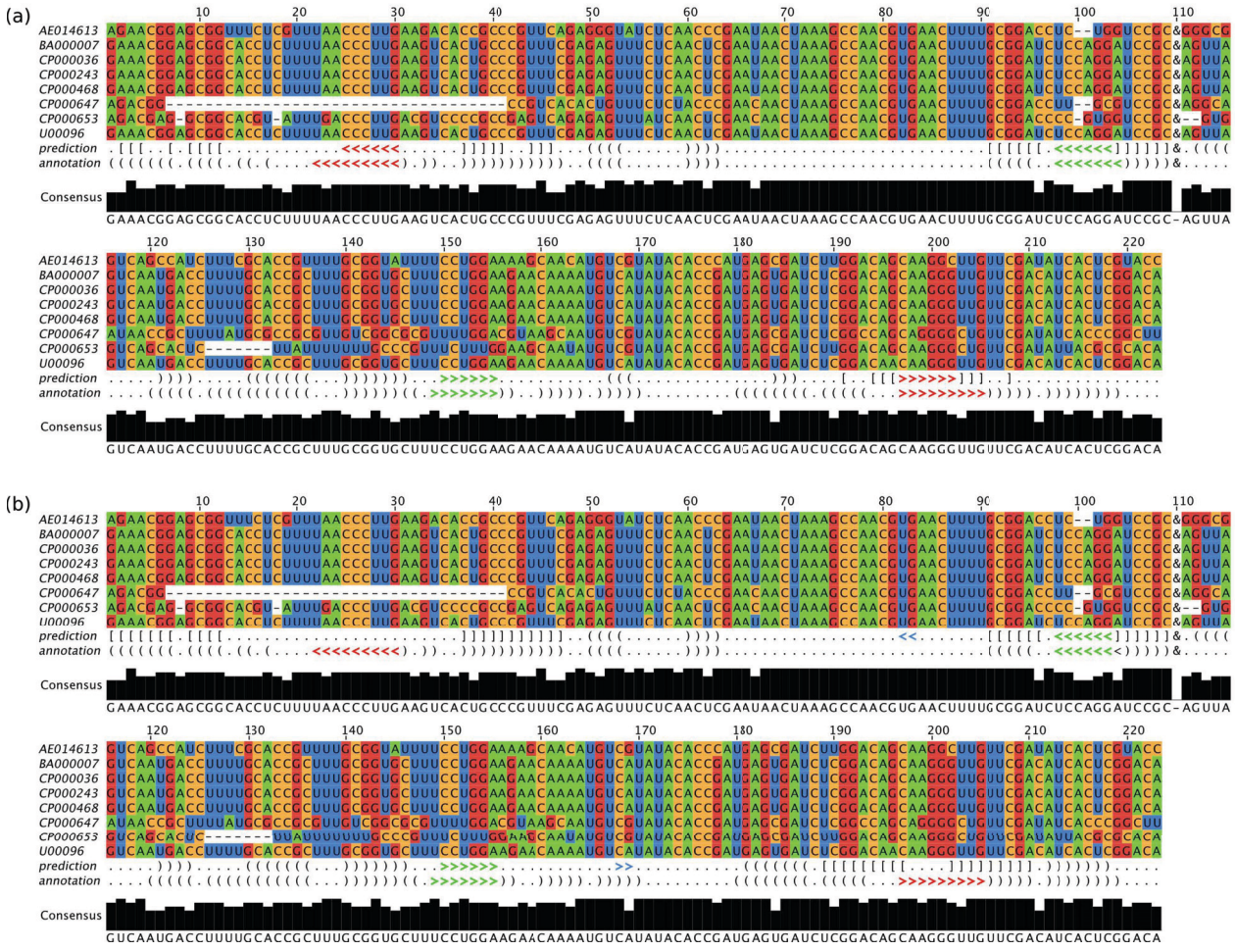

Figure 4 Joint secondary structure prediction of the sRNA OxyS and its target mRNA fhlA. The sequence alignment shows the two input alignments concatenated by the linker symbol "&", the joint secondary structure predicted by our algorithm (with $\delta = 0.9$ and $\gamma = 0.1$ and disallowing base pairs that only occur isolated in the thermodynamic part) and the interaction model of OxyS-*fhlA* [11]. The prediction was performed (a) without and (b) with extension of the constrained stems. Angle brackets indicate *inter*-molecular base pairs between the two RNAs. Round and square brackets indicate *intra*-molecular base pairs. Square brackets indicate base pairs of the partial structure σ^{P}. Shown are all columns with < 75% gaps. Jalview [33] was used for alignment visualisation.

and the number of iterations I in the adaptation of δ to find probable partial structures ($I < M/2$, where M is the sequence length of the longer alignment).

`Pfold` combines a SCFG with evolutionary information from an alignment of RNA sequences through an explicit evolutionary model. It is not clear whether the model learned from tRNA and rRNA secondary structures is appropriate for RNA cofolding. To avoid the bias of a wrong prior probability distribution $\Pr[\sigma|T, M]$ during the evolutionary scoring of the cofolding step, we omitted all SCFG rule probabilities as well as base pair substitution rates. In these cases, all base pair probabilities were calculated independently, and the different substitution rates of base pairs were ignored; thus, we observed that the entire performance decreased. A possible optimisation would be an adapted prior distribution for the cofolding step, which could be generated when sufficient verified RNA-RNA interaction data becomes available. However, the prediction of RNA secondary structure using evolutionary history is robust for different evolutionary speeds and substitution rate variations [32]. Hence, it is reasonable to assume that the deviation is fairly low using the prior probability distribution of *intra*-molecular structures.

The presented method assumes that both alignments have the same evolutionary history during the cofolding step. A more accurate method would consider independent phylogenetic trees for both RNAs, such as in Step 1, and a common tree for the interaction site. However, we do not know the interacting region in advance; thus, we would need an expectationm maximization algorithm, which would increase the running time of the algorithm

to an unreasonable level. Furthermore, the energy contribution of the cofolding step might be slightly biased due to the constraint of the partial structures as single-stranded. We partly solve the resulting *intra*-molecular false predictions by extending the reliable stems in the partial structures, and, as already mentioned above, the `RNAcofold` algorithm and scoring scheme could be adapted to handle base pair constraints as single-stranded.

Furthermore, there is a limitation of the presented method with regard to interaction sites that are located outside conserved RNA structures. These regions are hard to align if they are, in addition, sequentially unconserved. Thus, our method will most likely miss binding sites located in unstructured and otherwise unrelated regions, *e.g.*, miRNA target sites in UTR regions. However, once a correct alignment is found for these regions, then the presented approach still works if the interaction region is conserved or shows enough covariation.

Our algorithm is able to predict pseudoknots between the *intra*-molecular and (inter-)molecular base pairs. In addition, we are interested in more pseudoknots that can be predicted in a similar way using a pipeline of constrained structures. During an iteration of Step 2, additional reliable partial *inter*-molecular structures are constrained as long as new reliable base pairs appear. The final consensus structure is the union of all cofolding base pairs and the partial structures. The main unsolved problem is the weighted combination of the decreasing partial structure probabilities in one scoring scheme when the amount of constraints increases with each iteration.

Conclusions

In summary, we introduced an extension of the `PETfold` algorithm for the identification of interactions between two sets of multiple aligned RNA sequences, which exploits compensating base changes while taking *intra*-molecular partial structures and interaction sites into account. The implementation of the algorithm in `PETcofold` and its application are described in Seemann *et al.* (submitted).

Additional material

Additional file 1 Probability distributions of the `Pfold` model and Implications of the independence property. Combined probability distributions of the Pfold model and theimplications of the independence property (Equation (11)) for partial structures are described.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SES, RB and JG designed the algorithm. SES implemented the algorithm. ASR designed and performed the analysis of the algorithm. RB drafted the manuscript. All authors contributed to the manuscript, read and approved the final manuscript.

Acknowledgements

We thank Bjarne Knudsen for inspiring discussions about extensions of the `Pfold` method.

This work was supported by the Lundbeck Foundation (grant 374/06 to J.G.), the Danish Research Council for technology and production (grant 274-09-0282 to J.G.), the Danish Center for Scientific Computation (J.G.), the German Federal Ministry of Education and Research (BMBF grant 0313921 FRISYS to R.B.) and the German Research Foundation (DFG grant BA 2168/2-1 SPP 1258 to A.S.R. and R.B.).

Author Details

[1]Center for non-coding RNA in Technology and Health, IBHV, University of Copenhagen, Grønnegårdsvej 3, Frederiksberg C, 1870, Denmark and [2]Bioinformatics Group, University of Freiburg, Georges-Köhler-Allee 106, Freiburg, 79110, Germany

Received: 10 April 2010 Accepted: 21 May 2010
Published: 21 May 2010

References

1. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF: **Genome-wide mapping of conserved RNA secondary structure structures predicts thousands of functional non-coding RNAs in human.** *Nature Biotechnology* 2005, **23:**1383-1390.
2. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D: **Identification and classification of conserved RNA secondary structures in the human genome.** *PLoS Comput Biol* 2006, **2:**e33.
3. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J: **Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure.** *Genome Research* 2006, **16:**885-889. [(Erratum in: Genome Res. 2006 16:1439)].
4. Uzilov AV, Keegan JM, Mathews DH: **Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change.** *BMC Bioinformatics* 2006, **7:**173.
5. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigó R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF: **Structured RNAs in the ENCODE selected regions of the human genome.** *Genome Research* 2007, **17:**852-864.
6. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R: **Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering.** *PLoS Computational Biology* 2007, **3:**e65.
7. Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, Gorodkin J: **Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions.** *Genome Research* 2008, **18:**242-251.
8. Backofen R, Hess WR: **Computational prediction of sRNAs and their targets in bacteria.** *RNA Biol* 2010, **7:**.
9. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL: **Thermodynamics of RNA-RNA binding.** *Bioinformatics* 2006, **22(10):**1177-82.
10. Busch A, Richter AS, Backofen R: **IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions.** *Bioinformatics* 2008, **24(24):**2849-56.
11. Argaman L, Altuvia S: ***fhlA* repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex.** *Journal of Molecular Biology* 2000, **300(5):**1101-12.
12. Andronescu M, Zhang ZC, Condon A: **Secondary structure prediction of interacting RNA molecules.** *Journal of Molecular Biology* 2005, **345(5):**987-1001.
13. Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, Hofacker IL: **Partition function and base pairing probabilities of RNA heterodimers.** *Algorithms Mol Biol* 2006, **1:**3.
14. Dirks RM, Bois JS, Schaeffer JM, Winfree E, Pierce NA: **Thermodynamic Analysis of Interacting Nucleic Acid Strands.** *SIAM Review* 2007, **49:**65-88.
15. Zuker M: **Prediction of RNA secondary structure by energy minimization.** *Methods in Molecular Biology* 1994, **25:**267-94.

16. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures.** *Monatshefte Chemie* 1994, **125:**167-188.
17. Pervouchine DD: **IRIS: intermolecular RNA interaction search.** *Genome Inform* 2004, **15(2):**92-101.
18. Alkan C, Karakoc E, Nadeau JH, Sahinalp SC, Zhang K: **RNA-RNA interaction prediction and antisense RNA target search.** *Journal of Computational Biology* 2006, **13(2):**267-82.
19. Chitsaz H, Salari R, Sahinalp SC, Backofen R: **A partition function algorithm for interacting nucleic acid strands.** *Bioinformatics* 2009, **25(12):**i365-73.
20. Huang FWD, Qin J, Reidys CM, Stadler PF: **Partition function and base pairing probabilities for RNA-RNA interaction prediction.** *Bioinformatics* 2009, **25(20):**2646-54.
21. Huang FWD, Qin J, Reidys CM, Stadler PF: **Target prediction and a statistical sampling algorithm for RNA-RNA interaction.** *Bioinformatics* 2010, **26(2):**175-81.
22. Chitsaz H, Backofen R, Sahinalp SC: **biRNA: Fast RNA-RNA Binding Sites Prediction.** *Proc. of the 9th Workshop on Algorithms in Bioinformatics (WABI), Volume 5724 of Lecture Notes in Computer Science* 2009:25-36.
23. Salari R, Backofen R, Sahinalp SC: **Fast prediction of RNA-RNA Interaction.** *Proc. of the 9th Workshop on Algorithms in Bioinformatics (WABI), Volume 5724 of Lecture Notes in Computer Science* 2009:261-272.
24. Salari R, Möhl M, Will S, Sahinalp SC, Backofen R: **Time and space efficient RNA-RNA interaction prediction via sparse folding.** *Proc of RECOMB 2010* 2010 in press.
25. Seemann SE, Gorodkin J, Backofen R: **Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments.** *Nucleic Acids Research* 2008, **36:**6355-6362.
26. Gaspin C, Westhof E: **An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints.** *J Mol Biol* 1995, **254:**163-174.
27. Jabbari H, Condon A, Pop A, Pop C, Zhao Y: **HFold: RNA Pseudoknotted Secondary Structure Prediction Using Hierarchical Folding.** In *Algorithms in Bioinformatics, 7th International Workshop, WABI Philadelphia, PA, USA, September 8-9, 2007, Proceedings* 2007:323-334.
28. Knudsen B, Hein JJ: **RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.** *Bioinformatics* 1999, **15:**446-454.
29. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29(6-7):**1105-19.
30. Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ: **Algorithms for Loop Matchings.** *SIAM Journal on Applied Mathematics* 1978, **35:**68-82.
31. Ding Y, Chan CY, Lawrence CE: **RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble.** *RNA* 2005, **11(8):**1157-66.
32. Knudsen B, Andersen ES, Damgaard C, Kjems J, Gorodkin J: **Evolutionary rate variation and RNA secondary structure prediction.** *Comput Biol Chem* 2004, **28(3):**219-226.
33. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ: **Jalview Version 2-a multiple sequence alignment editor and analysis workbench.** *Bioinformatics* 2009, **25(9):**1189-91.

doi: 10.1186/1748-7188-5-22
Cite this article as: Seemann *et al.*, Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions *Algorithms for Molecular Biology* 2010, **5**:22