Fast and Accurate Structure Probability Estimation for Simultaneous Alignment and Folding of RNAs

Milad Miladi
Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany
miladim@informatik.uni-freiburg.de

Martin Raden
Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany
mmann@informatik.uni-freiburg.de

Sebastian Will
Theoretical Biochemistry Group (TBI), Institute for Theoretical Chemistry,
University of Vienna, Austria
will@tbi.univie.ac.at

Rolf Backofen
Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany
Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Germany
backofen@informatik.uni-freiburg.de

Abstract

Motivation: Simultaneous alignment and folding (SA&F) of RNAs is the indispensable gold standard for inferring the structure of non-coding RNAs and their general analysis. The original algorithm, proposed by Sankoff, solves the theoretical problem exactly with a complexity of $O(n^6)$ in the full energy model. Over the last two decades, several variants and improvements of the Sankoff algorithm have been proposed to reduce its extreme complexity by proposing simplified energy models or imposing restrictions on the predicted alignments.

Results: Here we introduce a novel variant of Sankoff's algorithm that reconciles the simplifications of PMcomp, namely moving from the full energy model to a simpler base pair-based model, with the accuracy of the loop-based full energy model. Instead of estimating pseudo-energies from unconditional base pair probabilities, our model calculates energies from conditional base pair probabilities that allow to accurately capture structure probabilities, which obey a conditional dependency. Supporting modifications with surgical precision, this model gives rise to the fast and highly accurate novel algorithm Pankov (***P**robabilistic **S**an**k**off-like simultaneous alignment and folding of RNAs inspired by Mar**kov** chains*). Pankov benefits from the speed-up of excluding unreliable base-pairing without compromising the loop-based free energy model of the Sankoff's algorithm. We show that Pankov outperforms its predecessors LocARNA and SPARSE in folding quality and is faster than LocARNA. Pankov is developed as a branch of the LocARNA package and available at `https://github.com/mmiladi/Pankov`.

2012 ACM Subject Classification Applied computing → Bioinformatics; Applied computing → Computational biology; Applied computing → Molecular structural biology

Keywords and phrases RNA secondary structure, Structural bioinformatics, Alignment, Algorithms

Digital Object Identifier 10.4230/LIPIcs.WABI.2019.14

Funding MM and MR are supported by the German Research Foundation (DFG *BA2168/14-1* and *BA2168/16-1*). SW is supported by the Austrian Science Fund (FWF *I 2874-N28*). RB is supported by the German Research Foundation (DFG) under Germany's Excellence Strategy (CIBSS – EXC-2189 – Project ID 390939984).

Acknowledgements The authors would like to thank the anonymous reviewers.

1 Introduction

In all forms of life, RNAs play essential roles that go beyond coding as messenger RNAs for the synthesis of proteins. Non-coding RNAs (ncRNAs) directly regulate cellular mechanisms, where some are known to be conserved for billions of years [2]. ncRNAs often have only weak sequence conservation, since their (conserved) structure crucially determines their function. Therefore, inferring the conserved secondary structure of homologs – most often, based on RNA alignments, is central for the discovery and annotation of functional RNAs.

RNA structural alignment algorithms can be classified depending on whether they fold and align simultaneously or in turn [19]. The gold standard for computing reliable alignments (and common structures) of RNAs is still the simultaneous algorithm proposed by Sankoff in 1985 [18]. By simultaneously aligning and folding the RNAs, it resolves the vicious cycle that reliable RNA alignments must consider RNA structures (especially for RNAs of medium to low sequence identity [6]), while computational structure prediction is typically unreliable without comparative information. For a pair of RNA sequences, the algorithm finds the optimal alignment and two compatible secondary structures that minimize the total of sequence-alignment distance score and the free energies of the predicted structures. With a run-time complexity of $O(n^6)$ and $O(n^4)$ for memory, the method requires extreme computational resources, such that its application is largely restricted to small instances and data sets. Efficient alignment algorithms are needed for the multiple alignments of the clusters that can be obtained from large scale clustering of data from high-throughput sequencing experiments [16].

Thus, several approaches have adapted Sankoff's original algorithm to reduce its computational costs. Two main lines of the variants can be distinguished. Methods like Dynalign [13] and FoldAlign [9] reduce the computational complexity by heuristic restrictions, e.g. introducing banding strategies or limiting the maximum size of the comparable sub-structures. Since such methods need to perform expensive energy computations in the nearest neighbor model [14], their applications are still limited without considering heuristic restrictions that in turn could compromise their accuracy.

A highly viable alternative was proposed with the PMcomp algorithm [10], which replaces the nearest neighbor energy model with a still accurate probabilistic model. This model allows to drastically simplify the algorithms, which strongly reduces the computational overhead and supports further algorithmic optimizations in PMcomp's successors. For reducing the overhead, PMcomp-type algorithms evaluate structures based on base pair probabilities, which are precomputed by McCaskill's algorithm [15] instead of calculating nearest neighbor energy terms. Moreover, PMcomp-type algorithms such as LocARNA and similar approaches with a probabilistic energy model [22, 20, 8, 11] further speed up by reducing the search space. To this end, LocARNA considers only base-pairs that pass a defined probability threshold. This sparsification improves over PMcomp's complexity of $O(n^6)$ time and $O(n^4)$ space, each, by a quadratic factor (resulting in the $O(n^4)$ time and $O(n^2)$ space requirements of LocARNA).

With the algorithm SPARSE [21], we introduced a second level of sparsification using loop-closing aware recursions to filter based on the joint probability of base-pairs and associated loop-closing base-pairs. This sparsification reduces the time complexity of the alignment algorithm to $O(n^2)$, starting from the precomputed probabilities. The joint probabilities were computed based on our extension of the McCaskill's algorithm [17].

We emphasize that all these methods, starting with the original Sankoff algorithm, consider only non-crossing structures, even if crossing base pairs occur relatively frequently in physical structures [3]. This is generally justified, since most secondary structures are

dominated by non-crossing base pairs; in turn, the limitation to non-crossing structures allows dynamic programming techniques, which are far more efficient and flexible than comparable techniques that consider (even limited forms of) base pair crossings.

In this work, we utilize the joint probabilities in a novel way – not only for strong sparsification as in SPARSE but as well to evaluate RNA structure more accurately in a PMcomp-type algorithm. We start with showing that joint (or equivalently, conditional) probabilities allow to precisely capture structure probabilities in the nearest neighbor energy model. This corresponds to the exact capture of the nearest neighbor energies themselves. Remarkably, while previous work discussed only the stacking base-pair helices [1], we cope with all loop types. Based on the exact model, we suggest careful simplifications, that allow incorporation of the model into alignment and folding (in the variant of the SPARSE algorithm). Based on the novel precise probabilistic model, we propose the novel algorithm Pankov with $O(n^2)$ time complexity. As fundamental novelty over its predecessors, it applies an accurate full-loop energy model for evaluating the structures.

Performing an established benchmark, we show that Pankov is in practice faster than LocARNA *and* significantly improves structure prediction over both SPARSE and LocARNA. Compared to SPARSE, it even improves the sequence alignment quality.

2 Methods

2.1 Preliminaries

Basic notations

An *RNA sequence* A is a string over the alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$. A *base pair* a *of* A is a pair $(a^{\mathrm{L}}, a^{\mathrm{R}})$ $(1 \leq a^{\mathrm{L}} < a^{\mathrm{R}} \leq |A|)$ such that the respective sequence positions are complementary, i.e. `AU`, `GC` or `GU`. A *non-crossing RNA structure* S *of* A, in the following called *structure*, is a set of base pairs, where each two different base pairs (i, j) and (i', j') of S do not cross, i.e. $i < i' < j < j'$, and do not share any end, i.e. i, j, i', and j' are pairwise different. To treat the external bases pairs of an RNA structure, we introduce a pseudo-base-pair $a_0 := (0, |A| + 1)$, which formally encloses all base pairs of A.

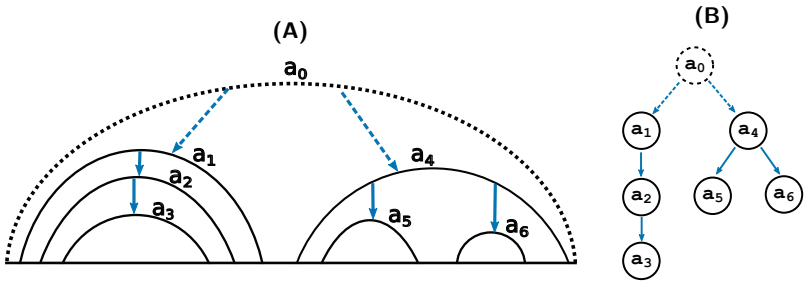
Tree structure

A nested RNA secondary structure S can be represented as a rooted structure tree, exemplified in Figure 1(A,B), where base-pairs are encoded as nodes and the enclosing base-pairs are the *parents* of the directly enclosed base-pairs. The $\mathrm{chi}(a \in S)$ function provides the set of children base pairs that are directly enclosed by a given base pair a. Thus, the cardinality of $\mathrm{chi}(a)$ is zero for hairpin loops (a_3, a_5, a_6 in Fig. 1), at least two for multi-loops (a_4) and one otherwise (a_1, a_2), which represents stackings, bulges or interior loops. Furthermore, the pseudo base pair a_0 recursively encloses all base pairs of any structure that can be formed by A.

Energy and probability of an RNA structure

The *energy of a structure* S can be estimated using the nearest neighbor energy model, which is based on a loop-decomposition of the structure, where a loop is defined as the substructure defined by a base pair a and its enclosed base pairs $\mathrm{chi}(a)$. The energy model provides *contributions* $E^{loop}(a, \mathrm{chi}(a))$ *for such loops*, which are summed up, i.e.

$$E(S) = \sum_{a \in S} E^{loop}(a, \mathrm{chi}(a)). \tag{1}$$



(C)

$$\mathbf{P}^{\text{PMcomp}}(\mathbf{S}) = \mathbf{P}(\mathbf{a_1}) \cdot \mathbf{P}(\mathbf{a_2}) \cdot \mathbf{P}(\mathbf{a_3}) \cdot \mathbf{P}(\mathbf{a_4}) \cdot \mathbf{P}(\mathbf{a_5}) \cdot \mathbf{P}(\mathbf{a_6})$$

$$\mathbf{P}^{\text{Pankov}}(\mathbf{S}) = \mathbf{P}(\mathbf{a_3}\|\mathbf{a_2}) \cdot \mathbf{P}(\mathbf{a_2}\|\mathbf{a_1}) \cdot \mathbf{P}(\mathbf{a_1}\|\mathbf{a_0}) \cdot [\mathbf{P}(\mathbf{a_5}\|\mathbf{a_4}) \cdot \mathbf{P}(\mathbf{a_6}\|\mathbf{a_4})] \cdot \mathbf{P}(\mathbf{a_4}\|\mathbf{a_0})$$

Figure 1 (A, B) Illustrations of an exemplary structure S and the respective structure tree. PMcomp energy model is composed of the probabilities of base-pairs, shown with plain black arcs and nodes. Pankov energy model additionally incorporates in-loop probabilities of paired base-pairs that are illustrated with blue arrows. **(C)** Top: PMcomp structure model assume independence between the base-pair probabilities by multiplying the probabilities. Bottom: Pankov uses a loop-aware scheme to compute the probability of the structure that is efficiently calculated from pre-computed loop-conditional probabilities $P(a\|a')$ based on the parent-child relation of the base-pairs.

By assuming a Boltzmann distribution of the structures based on the principles of statistical mechanics, there is a bijection between energies and probabilities of structures. Thus, the *probability* $P(S)$ *of the structure* S is related to its energy $E(S)$ by

$$P(S) = \exp\left(-E(S)/RT\right)/Z = \frac{1}{Z}\prod_a \exp(-E^{loop}(a, \text{chi}(a))/RT), \tag{2}$$

based on its loops' *Boltzmann weights* and the *partition function* $Z = \sum_S \exp\left(-E(S)/RT\right)$, which is efficiently computed by McCaskill's algorithm [15].

If inverted, this allows transforming structure probabilities back to energies:

$$E(S) = -RT \cdot \log(P(S)) + E^{\text{ens}}, \tag{3}$$

where $E^{\text{ens}} = -RT \cdot \log(Z)$ denotes the ensemble energy. This notation can be further used to derive the common definition of the *probability of a base pair* a $P(a)$ as

$$P(a) = \sum_{S \ni a} P(S), \tag{4}$$

which can be efficiently computed by McCaskill's algorithm.

2.2 PMcomp assumes independence of base-pair probabilities

PMcomp alignment and folding score

The *alignment and folding score* of PMcomp [10], which is assigned to an alignment $\mathcal{A}$ and RNA structures S_A and S_B, can be formulated as

$$\text{alignment-score}^{\text{PMcomp}}(S_A, S_B, \mathcal{A}) = \sum_{a \in S_A} \Psi_a^A + \sum_{b \in S_B} \Psi_b^B + \sum_{(i_A, i_B) \in \mathcal{A}} \sigma(i_A, i_B) + N_{\text{indel}}\gamma.$$

The first two terms define the structural component of the score that is discussed in the next section. The last two terms define the sequence component of the score. σ is the base similarity for two matched sequence positions i_A and i_B from sequence A and B, resp., and γ is the gap penalty. N_{indel} is the number of insertions and deletions in $\mathcal{A}$.

PMcomp structure scoring model

Here, we focus on the structure component of PMcomp's alignment and folding score, since we want to investigate how well the probabilistic model reflects the thermodynamic energy model. The score of a structure S of sequence A normalizes and sums up the log-scores of its base pair probabilities, i.e.

$$\text{score}^{\text{PMcomp}}(S) = \sum_{a \in S} \Psi_a^A = \sum_{a \in S} \log\left(P(a)/P_{\min}\right) = \log(\prod_{a \in S} P(a)) - |S \setminus \{a_0\}| \cdot \log\left(P_{\min}\right). \tag{5}$$

Here, base pair probabilities $P(a)$ are normalized via $P_{\min}$, the minimal probability of a significant base-pair, such that less probable base-pairs are unfavored. The logarithm transfers the probabilistic model back to the energy space similar to Eq. 3.

Putting the normalization term aside, the PMcomp's structure score contains a notion of structure probability (see first log term on the right of Eq 5), which we denote with

$$\text{P}^{\text{PMcomp}}(S) = \prod_{a \in S} P(a). \tag{6}$$

Noticeably, base pairing events are assumed to be independent, which is in violation with the underlying Nearest-Neighbor energy model, as we will show next. Thus, PMcomp's structure score does not relate well with the energy of the respective structure.

2.3 Exact computation of structure probabilities based on conditional loop probabilities

Here we prove that the equilibrium probability of a structure within the ensemble of possible structures can be expressed exactly based on conditional loop probabilities. This provides the theoretical foundation for discussing the Pankov energy model in the subsequent section.

▶ **Theorem 1.** *Let $P(S)$ be the probability of structure S and have $P(loop(a, \text{chi}(a)) \mid a)$ as the conditional probability of the loop in S closed by base-pair a, the following equation holds:*

$$P(S) = \prod_{a} P(loop(a, \text{chi}(a)) \mid a) \tag{7}$$

Proof. The free energy of the secondary structure S is composed of its loop energies E^{loop} in the nearest-neighbor thermodynamic model (Eq. 1), which implies that its probability can be computed from the respective Boltzmann weights of loops (Eq. 2). Decomposing the right term of Eq. 7 by the partition function Z_a inside the base-pairs a, i.e.

$$Z_a = \sum_{S_a \text{ closed by } a} \exp(-E(S_a)/RT), \tag{8}$$

we get:

$$\begin{aligned}\prod_a P(loop(a,\mathrm{chi}(a))|a) &= \prod_a \frac{\left(\prod_{a'\in\mathrm{chi}(a)} Z_{a'}\right)\cdot \exp(-E^{loop}(a,\mathrm{chi}(a))/RT)}{Z_a}\\ &= \frac{\prod_a \prod_{a'\in\mathrm{chi}(a)} Z_{a'}}{\prod_a Z_a}\cdot \prod_a \exp(-E^{loop}(a,\mathrm{chi}(a))/RT)\\ &=_* \frac{\prod_{a'\neq a_0} Z_{a'}}{\prod_a Z_a}\cdot \prod_a \exp(-E^{loop}(a,\mathrm{chi}(a))/RT)\\ &=_{**} \frac{1}{Z}\prod_a \exp(-E^{loop}(a,\mathrm{chi}(a))/RT) =_{(\mathrm{Eq.2})} P(S). \end{aligned} \tag{9}$$

$(=_*)$: Every arc but a_0 occurs exactly once as a child in the numerator product.
$(=_{**})$: a_0 encloses all possible structures, thus $Z_{a_0} = Z$. ◀

Eventually, this work extends and generalizes the approach for canonical helices (only stackings) from [1] to arbitrary secondary structures.

2.4 Pankov structure scoring model

Here, we want to score structures for simultaneous alignment and folding based on the derived Eq. 7, i.e. based on pre-computed conditional loop probabilities, to better reflect the structures' energy within the overall alignment score. Due to the exponential number of possible multi-loop branchings, a polynomial pre-computation and storing of respective multi-loop probabilities $P(loop(a,\mathrm{chi}(a))$ is not feasible. Thus, we propose an approximation of such terms based on pair-in-loop probabilities introduced next.

Approximate loop probabilities using pair-in-loop probabilities

To handle arbitrary multi-loops (closed by a) with any number and composition of children base pairs $\mathrm{chi}(a)$, we restrict computation and storage to all parent-child pairs $a \times a' \in \mathrm{chi}(a)$, i.e. we define the *pair-in-loop probability* $P(a'\|a)$, in the following abbreviated as *in-loop probabilities*, as

$$P(a'\|a) = P(a' \in \mathrm{chi}(a) \mid a) = \frac{P(a, a' \in \mathrm{chi}(a))}{P(a)} = \frac{\sum_{S\supset\{a,a'\}\wedge a'\in\mathrm{chi}(a)} P(S)}{P(a)}, \tag{10}$$

where we calculate the joint probabilities of the form $P(a, a' \in \mathrm{chi}(a))$ by an extension of McCaskill's algorithm introduced in [17], which can be performed in $O(n^3)$ time. To this end, the pair-in-loop information is stored in additional matrices during the partition function computation without increasing the computational complexity. The probability of a base-pair being external, i.e. being enclosed by the pseudo-arc a_0, is also computed and stored. Within the following, we discuss how to approximate loop probabilities from in-loop probabilities.

Pair-in-loop approximation of non-branching-loop probabilities. When using pair-in-loop probabilities to approximate non-branching loop probabilities, i.e. loops with exactly one child base pair, the latter is overestimated since Eq. 10 does not distinguish the loop context of the pair. Therefore, also multi-loops contribute to in-loop probabilities such that it follows

$$P(a'\|a) \geq P(loop(a, \{a'\} = \mathrm{chi}(a)) \mid a). \tag{11}$$

Scoring based on least stable multi-loop branch. An alternative approximation would be to assign the least probable branch to the whole multi-loop. This can be intuited in the energy space as that least stable branching of the multi-loop dominants formation of the multi-loop.

$$P_{\text{ML-min}}(loop(a, \text{chi}(a)) \mid a) = \min_{a' \in \text{chi}(a)} P(a' \| a). \tag{12}$$

Due to the same reasons as for non-branching loops, the following relation holds, i.e.

$$P_{\text{ML-min}}(loop(a, \text{chi}(a)) \mid a) \geq P(loop(a, \text{chi}(a)) \mid a). \tag{13}$$

Assuming multi-loop-branch independence. As a straightforward approach we assume an independence between the multi-loop branches that are conditioned to be closed under the same base-pair, i.e.

$$P_{\text{ML-prod}}(loop(a, \text{chi}(a)) \mid a) = \prod_{a' \in \text{chi}(a)} P(a' \| a). \tag{14}$$

Weighted overall structure scores

For scoring the structure in the implementation of the Pankov alignment algorithm, we assign the ultimate scoring $\text{score}^{\text{Pankov}}$ based on the product of multi-loop contributions following Eq. 14, that we designate as the Pankov's probability of structure.

$$P^{\text{Pankov}}(S) = \prod_{a \in S} \prod_{a' \in \text{chi}(a)} P(a' \| a) \tag{15}$$

Similar to PMcomp, the probabilities are incorporated on the logarithmic scale via the $\Phi(a', a)$ function, which also includes a normalization via the bonus term β. The latter term subsequently balances structure and sequence contributions within the alignment scoring.

$$\Phi(a', a) = \log(P(a' \| a)) + \beta, \tag{16}$$

$$\begin{aligned} \text{score}^{\text{Pankov}}(S) &= \sum_{a \in S} \sum_{a' \in \text{chi}(a)} \Phi(a', a) \\ &= log(P^{\text{Pankov}}(S)) + |S \setminus \{a_0\}| \cdot \beta \end{aligned} \tag{17}$$

2.5 Pankov alignment approach

The Pankov algorithm keeps track of closing-loop base-pairs in an efficient manner during dynamic programming computation of the score matrices. The matrices are defined similar to the dynamic programming recursions of the SPARSE algorithm [21], which achieve a quadratic time complexity for the alignment by exploiting the sparsity of the in-loop probabilities. Thus, Pankov uses the following matrices:

- $D(a, b)$ for the scores of matching the two base pairs $a = (a^{\text{L}}, a^{\text{R}})$ and $b = (b^{\text{L}}, b^{\text{R}})$ and aligning the two enclosed subsequences;
- $M^{ab}(i, k)$ for storing the maximum score of all possible alignments and foldings of the subsequences $A[a^{\text{L}} + 1..i]$ and $B[b^{\text{L}} + 1..k]$ that are under the loops enclosed by a and b; and
- I_{A} and I_{B} for supporting variability in the helix size via deletion and insertion of base-pairs under the loops closed by a and b respectively.

The matrix entries can be calcluted recursively:

$$D(a,b) = \max \begin{cases} M^{ab}(a^{\mathrm{R}}-1, b^{\mathrm{R}}-1) \\ I_{\mathrm{A}}^{ab}(a^{\mathrm{R}}-1) \\ I_{\mathrm{B}}^{ab}(b^{\mathrm{R}}-1) \end{cases}$$

$$M^{ab}(i,k) = \max \begin{cases} M^{ab}(i-1,k-1) + \sigma(i,k) \\ M^{ab}(i,k-1) + \gamma \\ M^{ab}(i-1,k) + \gamma \\ \max\limits_{\substack{P(a_1)\geq\theta, P(b_1)\geq\theta \\ a_1^{\mathrm{R}}=i, b_1^{\mathrm{R}}=k \\ P(a_1,a)\geq\theta' P(b_1,b)\geq\theta'}} \left(\begin{array}{l} M^{ab}(a_1^{\mathrm{L}}-1, b_1^{\mathrm{L}}-1) \\ + D(a_1,b_1) + \sigma(a^{\mathrm{L}},b^{\mathrm{L}}) + \sigma(a^{\mathrm{R}},b^{\mathrm{R}}) \\ +\Phi(a,a_1) + \Phi(b,b_1) \end{array} \right) \end{cases}$$

$$I_{\mathrm{A}}^{ab}(i) = \max \begin{cases} I_{\mathrm{A}}^{ab}(i-1) + \gamma \\ \max\limits_{P(a_1)\geq\theta, P(a_1,a)\geq\theta', a_1^{\mathrm{R}}=i} (a_1^{\mathrm{L}} - a^{\mathrm{L}} + 1)\cdot\gamma + D(a_1,b) + \Phi(a,a_1) \end{cases}$$

$$I_{\mathrm{B}}^{ab}(k) = \max \begin{cases} I_{\mathrm{B}}^{ab}(k-1) + \gamma \\ \max\limits_{P(b_1)\geq\theta, P(b_1,b)\geq\theta', b_1^{\mathrm{R}}=k} (b_1^{\mathrm{L}} - b^{\mathrm{L}} + 1)\cdot\gamma + D(a,b_1) + \Phi(b,b_1) \end{cases}$$

Here θ is the minimum base-pair probability threshold and θ' is the corresponding threshold for the joint in-loop probabilities. Following PMcomp's energy model, SPARSE calculates $\Phi(a,a_1)$ as Ψa_1 (independent of the enclosing base pair a) – in Pankov we make use of the full flexibility.

The asymptotic time complexity of the Pankov alignment algorithm is $O(n^2)$. Similar to SPARSE, the base pairing and joint in-loop probabilities need to be computed only once for each sequence. This is important for computing multiple alignments, using a pairwise aligner like Pankov in a progressive scheme (e.g. via `mlocarna` tool from the LocARNA software package). The Pankov implementation of the recursions keeps track of *both* ends of the base-pairs, while in SPARSE the tracking is relaxed and M matrices with common left-ends are combined for further speed-up, although the complexity is not affected. So in practice, compared to SPARSE, Pankov's run-time is slightly increased.

3 Results and Discussion

3.1 Evaluation of the probabilistic energy models

The evaluation procedure

We evaluated the agreement between the reference and the probabilistic free energy models. Having the Turner [14] nearest-neighbor full energy model as the reference, we compared the performance of PMcomp's base-pair independence model and Pankov's loop-based conditional probability model. The Sankoff-like algorithms maximize (or minimize) the sum of structure and sequence alignment scores over the space of possible formations of alignments and structure. Hence, a higher correlation between the calculated energies of a model with the reference free energy values indicates better modeling of the structure score, that is expected to perform better for the task of RNA simultaneous alignment and folding.

We developed an evaluation procedure to measure the level of agreement between the probabilistic models and the reference energy model with correlation coefficients. The procedure performs these steps for an input RNA sequence: (`i`) suboptimal secondary

structures are generated using RNAsubopt method [24], for the range of the minimum free energy structure up to 5kcal/mol (-e=5 -s) and 500 suboptimals. (`ii`) for each suboptimal structure, the probability or free energy is calculated according to the described energy models and structure scores (`iii`) Over the set of suboptimal structures, the Spearman's rank correlation coefficient between the reference RNAsubopt's free energies and the free energies/scores of the models are computed.

The five approximation variants of the probabilistic models are evaluated which compute the probability or score of a structure. More precisely, these variants are computed according to the equations $\mathrm{P}^{\mathrm{PMcomp}}$(Eq. 6), $\mathrm{score}^{\mathrm{PMcomp}}$(Eq. 5), P^{Pankov} using $P_{\text{ML-min}}$(Eq. 12), P^{Pankov} using $P_{\text{ML-prod}}$ (Eq. 14) and $\mathrm{score}^{\mathrm{Pankov}}$ using $P_{\text{ML-prod}}$ and β terms (Eq. 17). The structure probabilities are transferred to the energy dimension according to Eq. 3. For consistency in the representation, the scores were also scaled with a similar scheme. As a monotonic function, the energy scale transformation does not affect the absolute rank correlations in neither of the cases.

The minimal significance probability P_{min} in Eq. 5 was set to $1/(2 \cdot \text{sequence-length})$, this is used for the PMcomp score scheme of LocARNA. The bonus balance term β from Eq. 17 was set to 1.5, a bonus of zero makes the rank correlation of $\mathrm{score}^{\mathrm{Pankov}}$ same as the rank correlations of P^{Pankov} using $P_{\text{ML-prod}}$. However, a non-zero value is needed to balance the total alignment score by appropriately shifting the energy. The base-pair and conditional in-loop probabilities were computed using the extended implementation of McCaskill's algorithm (see methods). The run-time complexity for calculating these values for a structure, using the precomputed matrices, are linear to the number of base-pairs in structure (O(|S|)) as well as the length of the sequence.

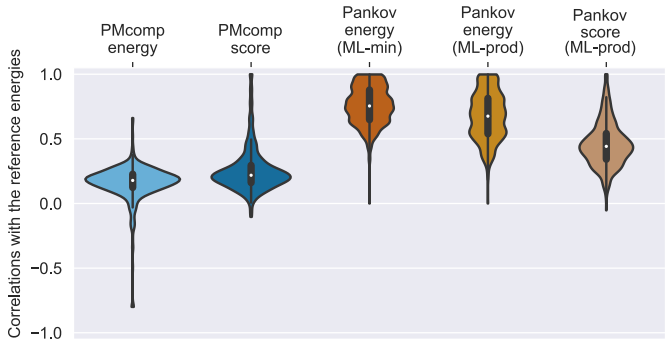


Figure 2 The distribution of rank correlation for evaluating the energy models over around 500 RNA sequences from RNAstrand database. Spearman's rank correlations are computed for each sequence, between the reference and the models' approximations of the suboptimal energies.

Evaluation of real ncRNAs

The described evaluation procedure was repeated for the set of RNA sequences obtained from the RNAstrand database (sequence length [30-200] nucleotides, entries without ambiguous or spurious characters). Figure 2 shows the distribution of the rank correlation evaluation of the sequences.

To inspect the model agreements in details, we visualized the output on a sample tRNA transcript of RNAstrand (ID: SPR_00633) in Figure 3. An evaluation for the top 500 suboptimal structures of lowest free energies is shown. As can be seen in Figure 3,

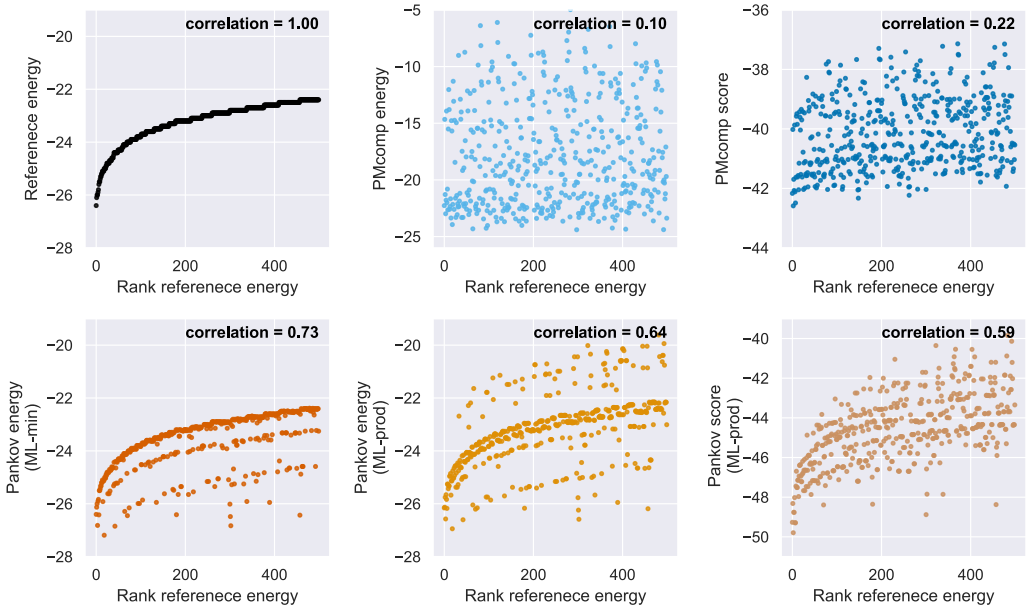


Figure 3 Evaluation of the agreement between the structure probabilistic and scoring models with the reference energies for a tRNA transcript. Each dot corresponds to a suboptimal structure of the tRNA. The structures are sorted according to reference free energy in the x-axis, computed by Vienna package RNAsubopt tool. The probabilities and scores are scaled to the energy dimension (Eq. 3). Annotated correlations of each panel are the Spearman's correlation coefficients between the x and y axes. Each correlation corresponds to an individual entry within the corresponding distributions that are shown as violin plots in Figure 2.

the Pankov's in-loop-probability-based models (bottom row of Figure 3) perform best in preserving the reference free energy ranks. It is also notable that the energy scaled values are precisely scaling back to the range of reference free energies.

The scatter plot for Pankov energy (ML-min) in Figure 3 confirms the relation for the probability of multiloops that was presented in the methods section (Eq. 13), P^{Pankov} with a ML-min approximation is bounded by the exact probability of the structure such that the approximated energies are always less than or equal to the reference energies.

3.2 Alignment performance evaluation

We evaluated our implementation of Pankov alignment algorithm on the pairwise alignment benchmark set, Bralibase 2.1 [23]. For the evaluated methods, the sparsification probability thresholds were set similar. Namely, LocARNA, SPARSE and Pankov with the minimum base-pair probability θ to 0.001 (option -p). For SPARSE and Pankov the in-loop probability threshold θ' was set to 0.0001 (option –prob-basepair-in-loop). The sequence-structure balance term β of Pankov's score (Eq. 17) was set to 1.5, the chosen among the values 1, 1.5, 2 and 2.5 posing a fair balance for the average of sequence and structure scores. The Matthews Correlation Coefficient (MCC) performance was stable for β of 1.5 and larger values.

Figure 4(A,B) show the performance comparison in term of sequence alignment quality sum-of-pairs-score (SPS) and structure prediction quality by Matthews Correlation Coefficient (MCC)[7]. To mediate the Bralibase curve "dent" effect [12], the visualization was done for sequence pairs of sequence identity (SI) between 30 to 80% to avoid a curve dent around SI-80

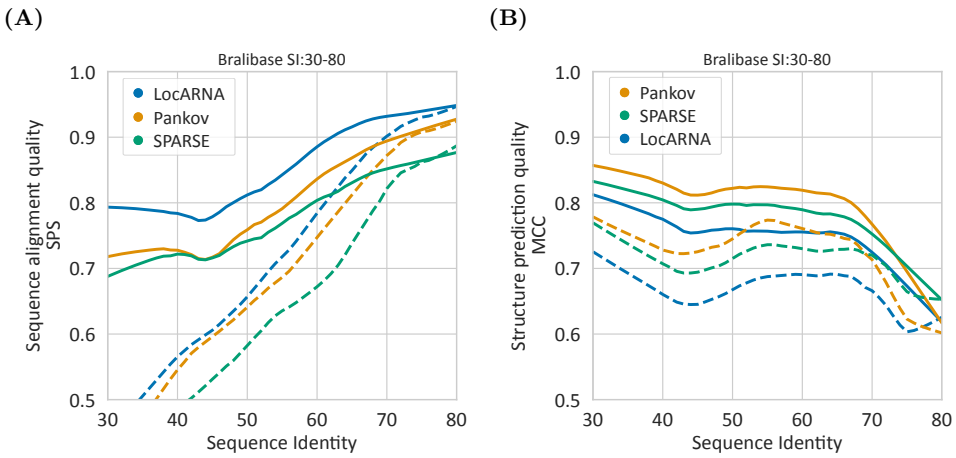


Figure 4 Comparison of the alignment performance on the Bralibase 2.1 pairwise benchmark set K2. **(A)** The sequence-level quality is measured as sum-of-pairs-score (SPS) by comparing the alignment edges of the reference and predicted alignments. **(B)** The structure prediction quality is measured as Matthews Correlation Coefficient (MCC) by comparing the base-pairs of the reference structure with the predicted structures. The solid lines depict all families and the dashed lines depict the subset without tRNA and the two rRNA families.

that seems to be mainly caused by enforcing a continuous curve over a quasi-heterogeneous distribution. Entries with a higher SI are not of particular interest, as they mostly perform fine also using the structure-unaware alignment algorithms. Furthermore, the dashed curves in Figure 4 correspond to the subset of the benchmark by excluding the three ribosomal and tRNAs families. These families are shown to be moderately overrepresented in the Braliabase and could overweight in the overall performance, especially on lower sequence identity range [12].

In the aspect of execution time, Pankov is overall faster than LocARNA and slower than SPARSE since Pankov implements the exact loop-closing track of the alignment recursions (see methods). Our implementation of Pankov had an average run time of 1.8 seconds on Bralibase K2 instances, when running on a AMD Opteron 2.1 GHz processor; this compares to respective average run times of 3.9 and 0.7 seconds of LocARNA and SPARSE. As can be seen in Figure 4, Pankov considerably improves structure prediction over both SPARSE and LocARNA. Compared to its predecessor SPARSE, it even improves the sequence alignment quality.

4 Conclusion

Sankoff's algorithm is the reference standard for simultaneous alignment and folding (SA&F) of RNAs. While the theoretical work integrates the full loop-based nearest-neighbor energy model, the derived algorithms mostly implement a simplified or limited structure energy models or restrict on the alignment and structure formation possibilities, to reduce the high computational complexity. PMcomp proposed a probabilistic light-weight energy model. This empowers the PMcomp-like methods to strongly reduce the computational overhead of the exact thermodynamic folding details and allows further algorithmic optimizations and sparsification based on the equilibrium probability of the base-pairs.

Here we showed that PMcomp's energy model assumes a level of independence between the base-pairing events, which violates the underlying nearest neighbor energy model. To solve this issue, we demonstrated an exact way to compute the probability of an RNA secondary structure from the decomposed loop-probabilities. To circumvent the computational complexity of multi-loop cases, we introduced an energy model to accurately approximate this loop decomposition in an efficient way using the precomputed in-loop probabilities. Our proposed energy model takes care of the nearest-neighbor thermodynamic rule. It was further empirically validated that the novel model has a much closer agreement with the full-loop energy model, based on the dataset of real non-coding RNAs. Using this energy model, we proposed the Pankov algorithm for pairwise simultaneous alignment and folding of RNAs. Benchmark results show that the implementation of Pankov outperforms its predecessors on predicting the secondary structure from the pairs of homologous RNAs.

The concept of conditional and joint in-loop probabilities has some parallels to the production rules of Stochastic Context-Free Grammars (SCFG) that can encode base-pair relations differently [5], they have also been used to solve the SA&F problem [4]. The overhead of treating various nucleotides separately during the alignment procedure is similar to the invocation of the full loop-based energy model, which restrains the implementation towards using simplified grammars that may not benefit from the power of thermodynamic rules. In our proposed model, the probabilistic terms are obtained from the thermodynamic partition function, so the probabilistic transition rules are straightforward and do not need to deal with individual types and combinations of nucleotides separately.

Pankov, to the best of authors' knowledge, is the first SA&F method that dissociates the loop computation details from the alignment and prediction step to efficiently solve the target problem without substantially compromising the power of underlying thermodynamic rules.

References

1 Athanasius F Bompfünewerer, Rolf Backofen, Stephan H Bernhart, Jana Hertel, Ivo L Hofacker, Peter F Stadler, and Sebastian Will. Variations on RNA folding and alignment: lessons from Benasque. *Journal of mathematical biology*, 56(1-2):129–144, 2008.

2 Thomas R Cech and Joan A Steitz. The noncoding RNA revolution—trashing old rules to forge new ones. *Cell*, 157(1):77–94, 2014.

3 Padideh Danaee, Mason Rouches, Michelle Wiley, Dezhong Deng, Liang Huang, and David Hendrix. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic acids research*, 46(11):5381–5394, 2018.

4 R. D. Dowell and S. R. Eddy. Efficient Pairwise RNA Structure Prediction and Alignment Using Sequence Alignment Constraints. *BMC Bioinformatics*, 7:400, 2006.

5 Robin D Dowell and Sean R Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC bioinformatics*, 5(1):71, 2004.

6 Paul P. Gardner, Andreas Wilm, and Stefan Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, 33(8):2433–9, 2005.

7 Jan Gorodkin, Shawn L Stricklin, and Gary D Stormo. Discovering common stem–loop motifs in unaligned RNA sequences. *Nucleic Acids Research*, 29(10):2135–2144, 2001.

8 Arif Ozgun Harmanci, Gaurav Sharma, and David H. Mathews. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics*, 8:130, 2007.

9 Jakob Hull Havgaard, Rune B Lyngsø, Gary D Stormo, and Jan Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 21(9):1815–1824, 2005.

10 I. L. Hofacker, S. H. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–7, 2004.

11 Hisanori Kiryu, Yasuo Tabei, Taishin Kin, and Kiyoshi Asai. Murlet: a practical multiple alignment tool for structural rna sequences. *Bioinformatics*, 23(13):1588–1598, 2007.

12 Benedikt Löwes, Cedric Chauve, Yann Ponty, and Robert Giegerich. The bralibase dent-a tale of benchmark design and interpretation. *Briefings in bioinformatics*, 18(2):306–311, 2016.

13 David H Mathews and Douglas H Turner. Dynalign: an algorithm for finding the secondary structure common to two rna sequences. *Journal of molecular biology*, 317(2):191–203, 2002.

14 DH Mathews, J Sabina, M Zuker, and DH Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–40, 1999.

15 J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990.

16 Milad Miladi, Alexander Junge, Fabrizio Costa, Stefan E Seemann, Jakob Hull Havgaard, Jan Gorodkin, and Rolf Backofen. RNAscClust: clustering rna sequences using structure conservation and graph based motifs. *Bioinformatics*, 33(14):2089–2096, 2017.

17 Christina Otto, Mathias Mohl, Steffen Heyne, Mika Amit, Gad M. Landau, Rolf Backofen, and Sebastian Will. ExpaRNA-P: simultaneous exact pattern matching and folding of RNAs. *BMC Bioinformatics*, 15(1):6602, 2014.

18 David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, 45(5):810–825, 1985.

19 Matthew G Seetin and David H Mathews. Rna structure prediction: an overview of methods. In *Bacterial Regulatory RNA*, pages 99–122. Springer, 2012.

20 Elfar Torarinsson, Jakob H. Havgaard, and Jan Gorodkin. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, 23(8):926–32, 2007.

21 Sebastian Will, Christina Otto, Milad Miladi, Mathias Möhl, and Rolf Backofen. SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, 31(15):2489–2496, 2015.

22 Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4):e65, 2007.

23 Andreas Wilm, Indra Mainz, and Gerhard Steger. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol*, 1:19, 2006.

24 Stefan Wuchty, Walter Fontana, Ivo L Hofacker, and Peter Schuster. Complete suboptimal folding of rna and the stability of secondary structures. *Biopolymers: Original Research on Biomolecules*, 49(2):145–165, 1999.