Interactive implementations of thermodynamics-based RNA structure and RNA-RNA interaction prediction approaches for example-driven teaching

Martin Raden[1,2*], Mostafa Mahmoud Mohamed[2], Syed Mohsin Ali[2], Rolf Backofen[2,3,4]

1 Chair of Forest Growth and Dendroecology, University of Freiburg, Freiburg, Germany
2 Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany
3 Center for Biological Signaling Studies (BIOSS), University of Freiburg, Freiburg, Germany
4 Center for Biological Systems Analysis (ZBSA), University of Freiburg, Freiburg, Germany

* mmann@informatik.uni-freiburg.de

Abstract

The investigation of RNA-based regulation of cellular processes is becoming an increasingly important part of biological or medical research. For the analysis of this type of data, RNA-related prediction tools integrated into of many pipelines and workflows. In order to correctly apply and tune these programs, the user has to have a precise understanding of their limitations and concepts. Within this manuscript, we provide the mathematical foundations and extract the algorithmic ideas that are core to state-of-the-art RNA structure and RNA-RNA interaction prediction algorithms. To allow the reader to change and adapt the algorithms or to play with different inputs, we provide an open-source web interface to JavaScript implementations and visualizations of each algorithm.

The conceptual, teaching-focused presentation enables a high-level survey of the approaches while providing sufficient details for understanding important concepts. This is boosted by the simple generation and study of examples using the web interface available under http://rna.informatik.uni-freiburg.de/Teaching/. In combination, we provide a valuable resource for teaching, learning and understanding the discussed prediction tools and thus enable a more informed analysis of RNA-related effects.

Author summary

RNA molecules are central players in many cellular processes. Thus, the analysis of RNA-based regulation has provided valuable insights and is often pivotal to biological and medical research. In order to correctly select appropriate algorithms and apply available RNA structure and RNA-RNA interaction prediction software, it is crucial to have a good understanding of their limitations and concepts. Such an overview is hard to achieve by end users, since most state-of-the-art tools are introduced on expert level and are not discussed in text books. Within this manuscript, we provide the mathematical means and extract the algorithmic concepts that are core to state-of-the-art RNA structure and RNA-RNA interaction prediction algorithms. The conceptual, teaching-focused presentation enables a detailed understanding of the

approaches using a simplified model for didactic purposes. We support this process by providing clear examples using the web interface of our algorithm implementation. In summary, we have compiled material and web applications for teaching - and the selfstudy of - several state-of-the-art algorithms commonly used to investigate the role of RNA in regulatory processes.

Background

Bioinformatics analyses has become indispensable to biological research. While platforms like Galaxy enable the setup of tool pipelines without expert knowledge [1, 2], one requires a general understanding of underlying concepts and algorithms to be able to successfully apply and adapt these pipelines to biological data [3, 4]. Thus, bioinformatics is thought in both computer science and biology studies.

It has been established that when teaching mathematics a combination of reflected example study and problem solving by hand fosters learning. This learning effect is heightened when done iteratively with increasing difficulty [5]. Thus, diverse examples covering different aspects of the topic have to be provided to guide the learning process. This is even more important in an e-learning or self-study context, where the study of examples that show different aspects of a problem might compensate for the missing interaction with a teacher [6, 7].

Here, we focus on RNA-related bioinformatics and especially on approaches for RNA structure and RNA-RNA interaction prediction. Both are essential when investigating the vast amount of regulatory RNA that is common to all kingdoms of life [8, 9]. The function of many RNA species is guided by their structure that is defined by the formation of intramolecular base pairs. For instance, prokaryotic small RNAs show evolutionary-conserved unstructured regions that regulate the expression of their target mRNAs via intermolecular base pairing [10, 11]. Thus, the prediction of both functional intramolecular structures of RNAs as well as their intermolecular (RNA-RNA) interaction potentials are central bioinformatics tasks.

Most computational methods for RNA structure or RNA-RNA interaction prediction are based on thermodynamic models and provide an efficient computation since Richard Bellman's principle of optimality [12] can be applied. This means that optimal solutions of a problem can be composed of optimal solutions of (independent) subproblems. This is used by dynamic programming approaches that decompose a problem into smaller problems and tabularize partial solutions. Robert Giegerich and colleagues developed a rigorous framework, namely Algebraic Dynamic Programming (ADP) [13, 14], to systematically study and develop dynamic programming approaches in a computer science context. In addition, they provided an online platform to study ADP programs for various problems also covering RNA related topics [15]. The central idea of ADP is to separate the strategy how a problem is decomposed into subproblems from the evaluation strategy, i.e. the objective of the optimization. We use the counting of structure alternatives for a given RNA to illustrate how dynamic programming can be applied to predictions problems. In particular, we introduce the decomposition strategy for (nested) RNA structure models.

The teaching of dynamic programming approaches is typically split into a theoretical introduction by the lecturer showing individual examples and a subsequent manual application by students where the methods are implemented or applied to solve small-scale problems for exercise. This leads often to a very small set of examples discussed due to the high amount of work needed for manual application and the limited gain of knowledge by iterated usage of once understood solution strategies. To increase the number of examples, e.g., to focus on different aspects of an individual method or to compare different approaches, either partial solutions have to be provided

or implementations made available. Beside single instances like the Nussinov algorithm, most state-of-the-art methods and their underlying algorithmic ideas are not covered by textbooks, e.g. [16–18]. Resorting to the original literature for teaching these algorithm, however, is complicated, as most approaches are introduced for very sophisticated energy models. While these advanced energy models are required for a successful application of these tools in real world scenarios, they often mask the basic and transferable algorithmic ideas for the non-expert reader since they require a high level of background knowledge.

We approach the aforementioned problems in two ways. First, we have stripped the model-specific energy details from the state-of-the-art methods for RNA structure prediction and RNA-RNA interaction prediction and present their underlying (or basic) algorithmic ideas. For that purpose, we use the most simple energy model available. State-of-the-art energy models take the structural context of base pairs into account. To this end, RNA structures are decomposed into loops (i.e., a region that is enclosed by one or more base-pairs) to calculate their overall energy. However, the algorithmic principles are essentially the same when using an energy model that considers bases-pairs without their structural context as basic units. Since all methods are presented using the same mathematical nomenclature, relationships and differences are easy to understand. Second, we provide a web-interface that provides interactive implementations of all algorithms discussed with extensive visualizations. This interface (i) helps to understand and follow the algorithms, (ii) eases the generation of interesting examples for different aspects to teach, and (iii) provides master solutions for comparison with your own calculations or implementations. Each section closes with a list of advanced questions that exemplify what can be studied and answered using the provided web interfaces available at:

http://rna.informatik.uni-freiburg.de/Teaching/.

RNA structure prediction topics covered within this manuscript are the formalization of RNA secondary structures and simplified energy models, computation of the number of structures with regards to the given model [19, 20], identification of the minimum free energy structure [21, 22], computation of partition functions [23], probability calculation for single base pairs and unpaired regions [23, 24], and identification of the maximum expected accuracy structure [25, 26].

RNA-RNA interaction prediction approaches are grouped according to their algorithmic idea as in [27] into hybrid-only interaction prediction [28–30], concatenation-based/co-folding interaction prediction [31, 32], and accessibility-based interaction prediction [24, 33, 34].

Results and Discussion

In the following, we will briefly introduce the available algorithms and their respective application to life science. Most algorithms are dynamic programming approaches. Thus, we also provide the corresponding recursions for the simplified RNA structure model, which we introduce first.

RNA

Ribonucleic acid (RNA) is a linear molecule built from nucleotides. The ribose sugars of the nucleotides are bound via interlinking phosphate groups. Furthermore, each sugar is connected to a nitrogenous base, typically one of adenine (`A`), guanine (`G`), cytosine (`C`) or uracil (`U`). The bases can form hydrogen bonds between two (non-consecutive) nucleotides, which is then called a base pair. Although other forms are possible, the

typically considered base pairs are $\mathtt{G}-\mathtt{C}$, $\mathtt{A}-\mathtt{U}$, and $\mathtt{G}-\mathtt{U}$ in both orientations. Pairing between nucleotides of the same molecule (intramolecular) defines its three-dimensional structure. In order to fulfill a certain regulatory function, typically a stable structure is needed. Thermodynamic analyses have identified base (pair) stacking as the major stabilizing force within RNA structures [35] and according energy estimates have been identified experimentally, e.g. refer to [36]. The functional structure of an RNA can regulate e.g. other RNA molecules by direct (intermolecular) base pairing, i.e. forming base pairs between two RNAs, called RNA-RNA interactions. While the probability of an initial contact is dependent on many factors such as concentration or location, the subsequent formation of a stable RNA-RNA interaction is assumed to follow the same thermodynamic principles as single structure formation. Thus, most ideas and parameters from RNA structure prediction are transfered to RNA-RNA interaction prediction approaches. It is important to note that thermodynamics-based approaches are again models that do not consider all factors that influence structure/interaction formation, as e.g., already bound molecules, specific solution conditions, kinetics of structure formation. Nevertheless, they typically allow for accurate predictions for the majority of RNA molecules [37].

RNA secondary structures

In the following, we provide the mathematical framework needed to define and solve RNA related problems. The *primary structure* of an RNA molecule can be described by its sequence of bases. That is, an RNA molecule of length n is defined by its sequence $S \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{U}\}^n$ of respective IUPAC single letter codes [38].

The *secondary structure* P of an RNA S is defined as a set of (ordered) base pairs, i.e. $P \subset [1,n] \times [1,n]$ with $(i,j) \in P \rightarrow i < j$. Typically it is assumed that each nucleotide can pair with at most one other nucleotide, i.e. $\forall (i,j) \neq (p,q) \in P : \{i,j\} \cap \{p,q\} = \emptyset$, and that only the introduced Watson-Crick or $\mathtt{G}-\mathtt{U}$ base pairs are allowed, i.e. $\forall (i,j) \in P : \{S_i, S_j\} \in \{\{\mathtt{A},\mathtt{U}\},\{\mathtt{C},\mathtt{G}\},\{\mathtt{G},\mathtt{U}\}\}$ extraneous to order. Such base pairs are said to be *complementary*. Furthermore, to restrict computational complexity of prediction algorithms, structures are constrained to be *non-crossing (nested)*, i.e. $\nexists (i,j),(p,q) \in P : i < p < j < q$. Using non-crossing structures generally allow a good estimate of the overall structure stability. However, it is important to note that crossing base pairs do exist, albeit not as abundant as non-crossing base pairs, and contribute to the final stability of the three dimensional shape. It is typically assumed that first non-crossing structural elements are formed that subsequently are linked via few crossing base pairs [39]. Thus, the majority of the structure can be modeled/predicted via nested base pairing, which strongly reduces the computational complexity. Finally, it is commonly enforced that pairing bases have a minimal sequence distance of l, also called *minimal loop length*, to incorporate steric constraints of structure formation. In the following, we will denote with $\mathcal{P}$ the set of all possible structures (also referred to as structural ensemble or structure space) that can be formed by a given sequence S. It has been shown that the size of the structure space $\mathcal{P}$ grows exponentially with sequence length n. For a minimal loop length l of 3, the growth is about 2.3^n [40].

Nested secondary structures can be visualized as outerplanar graphs where nucleotides are represented by nodes and edges represent base pairs or sequential backbone connections. Furthermore, dot-bracket strings can be used that encode for each position i whether it is unpaired '`.`', it is the smaller index (opening) of a base pair '`(`', or the larger (closing) index '`)`'.

As motivated by Ruth Nussinov and co-workers [21], we relate the stability of an RNA structure directly with its number of base pairs. Since some algorithms require explicit energy contributions of individual base pairs (e.g. McCaskill's algorithm to

compute base pair probabilities), we set the energy of any base pair E_{bp} to -1 for simplification purposes. Thus, the energy of a structure is given by $E(P) = |P| \cdot E_{bp}$. Note, this is in stark contrast to state-of-the-art RNA structure prediction approaches (e.g. using Zuker's algorithm [22]), which typically apply a Nearest Neighbor energy model [41, 42] and experimentally derived energy contributions [36]. Furthermore, all algorithms for RNA-RNA interaction prediction ignore concentration-dependence and other factors influencing the duplex formation, which is typically modeled within the Nearest Neighbor model by an 'initiation' energy term [24, 33, 34]. Nevertheless, the use of the simplified base-pair-focused model enables a much clearer presentation of the algorithms, which is better suited (and sufficient) to understanding their ideas and mechanisms. The transfer from the simple base pair maximization to the advanced energy models, as done by Michael Zuker and Patrick Stiegler [22], is generic and can be applied to all problems discussed within this manuscript. References to extended versions and implementations are provided for each approach.

Counting structures via Dynamic Programming

A first task that introduces the general structure of dynamic programming approaches used for RNA structure prediction is to compute the number of structures a sequence S can form, i.e. $|\mathcal{P}|$. Since the structure space $\mathcal{P}$ grows exponentially, explicit enumeration is inefficient. In order to apply dynamic programming, we first have to have a strategy of how to decompose such a problem into independent subproblems. Let us consider the subsequence $S_i..S_j$. We can easily split the problem into two independent problems by introducing a case distinction for its last position S_j: case (1) S_j is *not* involved in any base pairing and case (2) S_j is paired with some position S_k ($i \leq k < j$). Both cases are depicted in Fig. 1. The first case can be easily reduced to a smaller problem, namely to $S_i..S_{j-1}$, since the unpaired position S_j does not allow any structural alternatives. Thus, the reduced problem directly provides a count for case 1. On the contrary, each possible base pairing of S_j in the second case decomposes the problem into two smaller independent problems (one to the left of and one enclosed by the base pair (k, j)), since no base pair is allowed to cross (k, j) (nestedness condition, see above). Since any structural alternative of the left subproblem can be combined with any of the enclosed one, we have to multiply the numbers from these smaller subproblems to get the overall count for case 2.

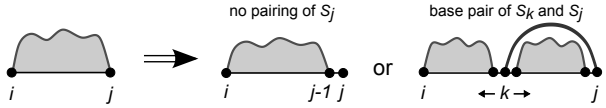

Fig 1. Secondary structure decomposition by Waterman & Smith (1978). The figure illustrates for a given subsequence $S_{i..j}$ a unique nested secondary structure decomposition based on the distinction of all possible pairing states of the last nucleotide S_j. Note, this scheme applies to all RNA structure related algorithms presented here.

Michael S. Waterman and Temple S. Smith applied this idea to solve the counting problem using a table C [19, 20]. An entry $C_{i,j}$ provides the number of structures for a subsequence $S_i..S_j$. Thus, we initialize $C_{i,i} = 1$ for all positions i, since any subsequence of length one is confined to the unpaired structure. The recursion for longer subsequences is given by

$$C_{i,j} \quad = \quad C_{i,j-1} + \sum_{\substack{i \leq k < (j-l) \\ S_k, S_j \text{ compl.}}} C_{i,k-1} \cdot C_{k+1,j-1} \qquad (1)$$

which combines the two discussed cases to consider all possible 'states' of nucleotide S_j

in valid structures. The first ($C_{i,j-1}$) covers all cases where S_j is unpaired, and the second counts all cases where S_j is paired with an S_k within the subsequence (second case). Note, the base pair (k, j) has to respect the minimal loop length l. The overall number of structures is accessed by $|\mathcal{P}| = C_{1,n}$. Given l and an RNA sequence, our user interface computes and depicts the filled matrix C.

Example Questions:

- The decomposition and counting of RNA structures was introduced for a case distinction on S_j. Rewrite Eq. 1 using a case distinction on S_i.
- Compute the numbers of nested structures that can be formed by random RNA sequences of different lengths. Compare the exponential growth of the structure space with the approximation 2.3^n mentioned earlier.

Optimal structure prediction

Ruth Nussinov and co-workers introduced in 1978 [21] a first algorithm that efficiently predicts a nested structure with the maximal number of base pairs for a given RNA sequence S, i.e. $\arg\max_{P\in\mathcal{P}}(|P|)$. The corresponding recursion

$$N_{i,j} = \max \begin{cases} N_{i,j-1} & S_j \text{ unpaired} \\ \max\limits_{\substack{i \le k < (j-l) \\ S_k, S_j \text{ compl.}}} (N_{i,k-1} + N_{k+1,j-1} + 1) & S_k, S_j \text{ pair} \end{cases} \quad (2)$$

is strongly related to the counting approach from Eq. 1. Here, an entry $N_{i,j}$ stores the maximal number of base pairs that can be formed by the subsequence $S_i..S_j$. Thus, summation in Eq. 1 is replaced by maximization and multiplication with summation while the second case considers the formed base pair with '+1'. N is initialized with 0 and can be filled in $O(n^3)$ time while using $O(n^2)$ memory. A depiction of the recursion is given in Fig. 2.

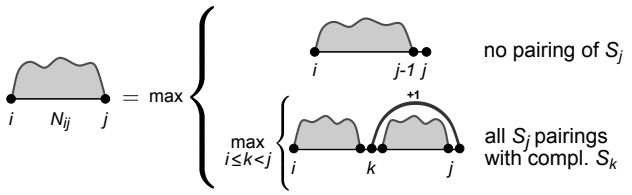

Fig 2. Recursion by Nussinov and coworkers (1978). The figure illustrates the recursion to compute the maximal number of base pairs that can be formed by a given sequence by distinction of all possible pairing states of the last nucleotide S_j.

The maximal number of base pairs formed by any structure can be found in $N_{1,n}$ and a respective optimal structure P can be identified via traceback starting in $N_{1,n}$. Thus, for a given cell $N_{i,j}$, the traceback discovers how the value of $N_{i,j}$ was obtained. To this end, the case distinctions of the (filling) forward recursion (e.g. from Eq. 2) are considered. If it holds $N_{i,j} = N_{i,j-1}$ (first case), position j is found to be unpaired and the traceback proceeds with cell $N_{i,j-1}$. Otherwise, position j has to form a base pair with some position $i \le k < j$, which is identified in accordance to the second case of Eq. 2. The base pair (k, j) is stored as part of the final structure P and the traceback proceeds for *both* subintervals represented by $N_{i,k-1}$ and $N_{k+1,j-1}$.

For the identification of functional structures or the study of structural alternatives, the enumeration of suboptimal structures is of interest. A generic approach was introduced by Stefan Wuchty and coworkers [43] that enables the enumeration of all

structures that are in a certain range of the minimal energy. An implementation is also available in our web interface.

Our interactive user interface enables the computation of both optimal and suboptimal structures. For a user defined sequence as well as recursion and traceback parameters, the dynamic programming table is provided along with a list of (sub)optimal structures. On selection, the according traceback is highlighted within the matrix. This is complemented with a graphical representation of the structure using FORNA [44].

Different recursions can be chosen to examine the effects of ambiguous recursions versus the original one. In the following, such an ambiguous variant from [17] is presented.

$$N_{i,j} = \max \begin{cases} N_{i+1,j} & S_i \text{ unpaired} \\ N_{i,j-1} & S_j \text{ unpaired} \\ N_{i+1,j-1} + 1 & \text{if } S_i, S_j \text{ compl. and } i + l < j \\ \max\limits_{i<k<(j-1)} N_{i,k} + N_{k+1,j} & \text{decomposition} \end{cases} \quad (3)$$

While this recursion also computes the same entries of N and thus maximal number of possible base pairs ($N_{1,n}$), it is not using a unique decomposition of the structure, i.e. the same structural variant is considered by different recursion cases.

This causes duplicated enumeration of (sub)optimal structures when using Wuchty's traceback algorithm, which can be studied in our web server for different recursions. Furthermore, it is not possible to use variants of ambiguous recursions like Eq. 3 to count structures (consider relation of Eq. 2 and 1) or to compute the partition function of the structural ensemble (as discussed next), since both requires a unique consideration of each structure.

In 1981, Michael Zuker and Patrick Stiegler introduced a dynamic programming approach that efficiently computes minimum free energy structures using a Nearest Neighbor energy model [22]. Using further restriction, the same time and space complexity compared to Nussinov's algorithm is kept. The approach with according decomposition depictions and how it relates to Nussinov's algorithm is introduced in detailed e.g. in [45]. Implementations like UNAFOLD [46] (former MFOLD [47]) or RNAFOLD [31, 37] are the current state-of-the-art tools for RNA secondary structure prediction.

Example Questions:

- Find RNA sequences that fold uniquely into i) a single hairpin, ii) two hairpins, and iii) three hairpins. What guided your design?
- Find an RNA sequence that shows the ambiguity of Eq. 3. What are the differences to Eq. 2 that cause this ambiguity?
- Define formally what is represented by the entry $N_{1,n}$ when using an energy minimizing variant of Eq. 2 that uses E_{bp} instead of '+1'. Provide a recursion to compute this value.

Partition function and probabilities

To estimate the probability of a given structure P within the structural ensemble $\mathcal{P}$, statistical mechanics typically dictates a Boltzmann distribution when using minimal

assumptions [48]. Thus, the probability of a structure P is directly related to its energy $E(P)$ by

$$\Pr(P) = \frac{\exp(-E(P)/k_B T)}{\sum_{P' \in \mathcal{P}} \exp(-E(P')/k_B T)} \tag{4}$$

given the Boltzmann factor k_B and the system's temperature T. Note, when using an energy model with units 'per mole', which is typically the case when using a Nearest Neighbor model with measured energy contributions, one has to replace k_B with the gas constant R. Note further, the structure with minimal free energy, e.g. predicted with algorithms discussed above, will always have maximal probability according to Eq. 4. Thus, the most stable structure is automatically the most likely structure.

The nominator of Eq. 4 is called Boltzmann weight (of structure P). The denominator is called canonical *partition function* Z, which is the sum of the Boltzmann weights of all structures in $\mathcal{P}$. Since $\mathcal{P}$ grows exponentially, its exhaustive enumeration to compute Z is impracticable.

Nevertheless, it is possible to compute Z efficiently using a variant of the counting algorithm. This approach was first introduced for the Nearest Neighbor energy model by John S. McCaskill (1990) [23] and we rephrase a variant for the simplified base pair model. First, we have to note that the Boltzmann weight of a structure P can be computed based on the energy of its base pairs E_{bp} as follows

$$\exp(-E(P)/k_B T) \;=\; \exp\left(-\sum_{(i,j)\in P} E_{bp}/k_B T\right) \;=\; \prod_{(i,j)\in P} \exp(-E_{bp}/k_B T). \tag{5}$$

That is the structure's weight is computed by the product of individual base pair weights. To simplify notation in the following, we refer with $q^{bp} = \exp(-E_{bp}/k_B T)$ to the Boltzmann weight of a single base pair. Given this, we can alter the counting recursion from Eq. 1 to

$$Q_{i,j} \;=\; Q_{i,j-1} + \sum_{\substack{i \le k < (j-l) \\ S_k, S_j \text{ pair}}} Q_{i,k-1} \cdot Q_{k+1,j-1} \cdot q^{bp}. \tag{6}$$

This directly provides the partition function $Z = Q_{1,n}$ in $O(n^3)$ time.

For some approaches and research questions, *probabilities of individual base pairs* $\Pr^{\text{bp}}(i,j)$ are of interest. This is the probability that a base pair (i,j) is formed by some structure, which can be calculated by summing up the probabilities of all structures containing (i,j), i.e.,

$$\Pr^{\text{bp}}(i,j) = \frac{\sum_{\substack{P \in \mathcal{P} \\ (i,j) \in P}} \exp(-E(P)/k_B T)}{Z}. \tag{7}$$

As for counting, the base pair (i,j) decomposes all structures into the enclosed and outer subsequence that are independent concerning base pairing. Thus, the partition functions of the according subsequences can be used to compute $\Pr^{\text{bp}}(i,j)$ efficiently. To do so, we need an auxiliary matrix Q^{bp}. Each entry $Q^{bp}_{i,j}$ holds the partition function for the subsequence $S_i..S_j$ with the side constraint that i and j form the base pair (i,j). If this is not possible due to non-complementarity or the minimal loop constraint, the entry is 0. Given this, we can rewrite Eq. 6 as follows

$$Q_{i,j} \;=\; Q_{i,j-1} + \sum_{i \le k < (j-l)} Q_{i,k-1} \cdot Q^{bp}_{k,j} \tag{8}$$

$$Q^{bp}_{i,j} \;=\; \begin{cases} Q_{i+1,j-1} \cdot q^{bp} & \text{if } S_i, S_j \text{ complementary} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

and compute the base pair probability using

$$\begin{aligned} \Pr^{\mathrm{bp}}(i,j) &= \frac{Q_{1,i-1}\cdot Q^{bp}_{i,j}\cdot Q_{j+1,n}}{Q_{1,n}} \\ &+ \sum_{p<i,j<q} \Pr^{\mathrm{bp}}(p,q)\cdot \frac{q^{bp}\cdot Q_{p+1,i-1}\cdot Q^{bp}_{i,j}\cdot Q_{j+1,q-1}}{Q^{bp}_{p,q}}. \end{aligned} \tag{10}$$

The first term in Eq. 10 covers structures where (i,j) is an external base pair, i.e. not enclosed by any other base pair. The second term considers all structures where (i,j) is directly enclosed by a base pair (p,q) and corrects the respective base pair probability $\Pr^{\mathrm{bp}}(p,q)$ by the probability of the structure subensemble that contains both base pairs and no 'in-between spanning' base pair (k,l) with $p<k<i<j<l<q$. The latter probability is defined by the fraction within the second term.. Note (again) that by using a simple energy model, we omit all the complex case distinctions, which allows one to concentrate on the main cases of algorithmic importance. In the full model, the first case would have been the same, whereas the second one would have been split to consider specifically each structural context a base pair can have.

In analogy to base pair probabilities, it is also possible to define and compute the *unpaired probability* $\Pr^{\mathrm{ss}}(i,j)$ of a subsequence $S_i..S_j$ (Eq. 11), i.e. the probability of all structures that show no base pairing in the single stranded subsequence.

$$\Pr^{\mathrm{ss}}(i,j) = \frac{\sum_{P\in\mathcal{P}^{\mathrm{ss}}_{i..j}} \exp(-E(P)/k_BT)}{Z} \tag{11}$$

$$\text{with } \mathcal{P}^{\mathrm{ss}}_{i..j} = \left\{ P \mid \nexists_{(k,l)\in P} : k\in[i,j] \vee l\in[i,j] \right\} \subseteq \mathcal{P} \tag{12}$$

The unpaired probability is also sometimes termed 'accessibility' as an unpaired region in an RNA is accessible for pairing to another RNA. For the computation of $\Pr^{\mathrm{ss}}(i,j)$, we only have to replace $Q^{bp}_{i,j}$ with 1 in Eq. 10, since only the unpaired structure with energy zero has to be considered for $S_i..S_j$, which has a Boltzmann weight of 1.

Stephan H. Bernhart and coworkers provide in [49] details for the extension of the introduced recursions to the Nearest-Neighbor model, which is also nicely detailed in [45]. Implementations are for instance available in the Vienna RNA package [37]. The authors also show how to reduce the time complexity of the probability computation from $O(n^4)$ to $O(n^3)$. To this end, they introduce another auxiliary matrix $\hat{Q}^{bp}$ that provides the 'outer' partition function, which reflects only base pairs not enclosed by respective subsequences.

Our web implementation enables the computation of both base-pair probabilities as well as unpaired probabilities. To provide insights into how the temperature and energy model influence structure and base-pair probabilities, the user can alter the used temperature as well as E_{bp}. Beside a visualization of the partition function tables Q and Q^{bp}, the user is provided with a visualization of the base pair and unpaired probabilities using the established *dot plot* format (e.g. used also by UNAFOLD/MFOLD [46,47] or RNAFOLD [37,50]). Within this matrix-like illustration, each base-pair probability is represented by a dot of proportional size; i.e. the higher the probability, the larger the dot and small probabilities are not visible. With a bit of visual practise, dot plots enable an easy identification of highly probable substructures and the study of structural alternatives.

Example Questions:

- Find an RNA sequence that folds uniquely into a single hairpin but shows an alternative hairpin with high base pair probabilities. What are the difficulties for such a design?

- What changes are observed for the partition functions when increasing the system's temperature? What is expected for $\lim T \to \infty$?
- Where are subsequences with high unpaired probability typically located?

Maximum expected accuracy

So far, individual structures were evaluated based on their number of base pairs or energy. This focus on single structures might hide that some substructures (base pairs or unpaired positions) are very common among highly-probable structures but not found e.g. in the most-probable structure and thus are lost from the prediction. To face this problem, the expected accuracy can be used for structure evaluation [25, 26, 51]. Here, we follow Chuong B. Do and coworkers [25] and define the expected accuracy of a structure P by

$$\mathrm{acc}(P) = \sum_{(i,j)\in P} \gamma \cdot 2 \cdot \Pr^{bp}(i,j) \; + \sum_{k\,:\,(i,k),(k,j)\notin P} \Pr^{u}(k). \tag{13}$$

It is basically the weighted sum of all base pair probabilities of the respective structure together with unpaired probability estimates for all its positions k not involved in any base pair, i.e. features of the whole structural ensemble are mapped to individual structures. The position-wise unpaired probability is computed by

$$\Pr^{u}(k) = 1 - \sum_{i<k} \Pr^{bp}(i,k) - \sum_{k<j} \Pr^{bp}(k,j) \tag{14}$$

from base-pair probabilities, which is equivalent to $\Pr^{\mathrm{ss}}(k,k)$ from Eq. 11. Base pair probabilities in Eq. 13 are weighted by a factor of two to reflect that two sequence positions are covered. Furthermore, a weighting factor γ is introduced, which scales the importance of unpaired vs. base pair probabilities.

Given this measure, we can compute the maximum expected accuracy (MEA) structure, i.e. a structure formed by the most accurate/likely base pairs rather than simply maximizing their number (or minimizing the overall energy). To calculate the MEA and an according structure, a variant of the Nussinov algorithm (Eq. 2) can be applied, i.e.

$$M_{i,j} = \max \begin{cases} M_{i,j-1} + P_j^u & S_j \text{ unpaired} \\ \max\limits_{\substack{i\le k<(j-l) \\ S_k, S_j \text{ compl.}}} \Big(M_{i,k-1} + M_{k+1,j-1} + 2\gamma \Pr^{bp}(k,j) \Big) & S_k, S_j \text{ pair,} \end{cases} \tag{15}$$

where unpaired positions are weighted by $\Pr^u$ (case 1) and base pairs with $2\gamma \Pr^{bp}_{i,j}$ (case 2). M is initialized with 0. The MEA is found in $M_{1,n}$ while a corresponding structure can be identified via traceback. A recursion variant adapting Eq. 3 can be found in [25].

Our MEA web interface computes base pair and unpaired probabilities using the recursions introduced above for the simplified energy model. Thus, the effects of temperature or base-pair energy E_{bp} on MEA computations can be directly studied. As for the Nussinov algorithm, structure and traceback visualization is enabled as well as suboptimal MEA enumeration using our generic implementation of Wuchty's algorithm [43]. An alteration of the γ weighting factor for base pair probabilities provides insights into its importance for accurate structure prediction.

Example Questions:

- Compare the prediction results for MEA and base pair maximization (energy minimization). What do you observe and how could you explain your observations?
- What happens when altering the base-pair probability weight γ?

Hybridization-only interaction prediction

The fastest class of RNA-RNA interaction prediction approaches focuses only on the identification of the interaction site, i.e. only on the intermolecular base pairs, without considering the intramolecular structures of the interacting RNAs. To this end, the prefix-based decomposition scheme of global sequence alignment [52] can be adapted.

Given two RNA sequences S^1 and S^2 of lengths n and m, resp., we denote with $\overleftarrow{S}^2_j$ the reversely indexed S^2 to simplify the index notation, since RNA molecules interact in antiparallel orientation. The latter applies to both intra- and intermolecular base pairing. When considering S^1 and $\overleftarrow{S}^2_j$, we can design a dynamic programming approach for the simplified energy model using a two-dimensional matrix H. An entry $H_{i,j}$ will provide the maximal number of intermolecular base pairs for the prefixes $S^1_{1..i}$ and $\overleftarrow{S}^2_{1..j}$. The decomposition scheme for the recursion of Eq. 16 to compute $H_{i,j}$ is visualized in Fig. 3.

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + 1 & \text{if } S^1_i, \overleftarrow{S}^2_j \text{ are complementary} \\ H_{i-1,j} \\ H_{i,j-1} \end{cases} . \tag{16}$$

As already mentioned, Eq. 16 is a variant of the global sequence alignment approach introduced by Saul B. Needleman and Christian D. Wunsch [52] using an adapted scoring scheme (base pair instead of match/mismatch scoring for $S_i, \overleftarrow{S}_j$ and no gap cost). Thus, initializing all $H_{i,0}/H_{0,j}$ with 0, the entry $H_{n,m}$ provides the maximal number of intermolecular base pairs that can be formed and a traceback starting at $H_{n,m}$ yields the respective interaction details. This approach enables very low runtimes ($O(nm)$), as observed by Brian Tjaden and coworkers who presented in [30] a variant of Eq. 16. When computing hybridization-only interactions via minimizing a more sophisticated energy model, the strategy has to be altered to follow a scheme similar to local sequence alignment as defined by Temple Smith and Michael S. Waterman [53], which is detailed in [30].

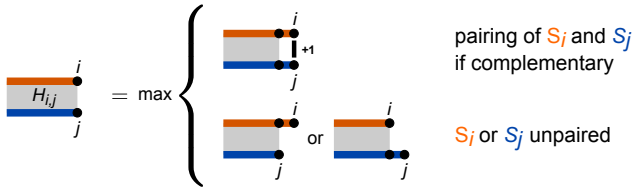

Fig 3. Recursion scheme to maximize intermolecular base pairs between two RNAs S^1 and S^2 represented in orange/blue, respectively. The optimal number for the interaction of $S^1_{1..i}$ and $S^2_{j..n}$ is identified based on a distinction whether or not the right ends S^1_i and S^2_j might form a base pair or not.

The web interface of our implementation identifies and reports all optimal interaction sites. For each, an ASCII visualization of the intermolecular base pairs is provided. Note, to reduce code redundancy, we do not use an implementation of Eq. 16 but a base-pair-maximization variant of Eq. 19, which is discussed in the next section.

Adaptations of this approach to the Nearest Neighbor model have been discussed in [28] and e.g. implemented in the tools TargetRNA [30], RNAhybrid [29] or RNAplex [54]. While such methods have been successfully applied for target site identification of very short RNAs, they often overestimate the length of target sites since intramolecular base pairing is ignored [33, 54]. These problems are tackled by concatenation- and accessibility-based approaches discussed next.

Example Questions:

- Provide a variant of Eq. 16 that uses the original sequence S^2 and according indexing, i.e. entry $H_{i,j}$ provides the maximal number of intermolecular base pairs for $S^1_{1..i}$ and $S^1_{1..j}$. Think about the computation order of entries for this matrix.
- Develop a dynamic-programming recursion for hybridization-only RNA-RNA interaction prediction (base-pair maximization) that restricts the lengths of unpaired subsequences enclosed by interacting base pairs. What is the runtime complexity of your recursion?

Concatenation-based RNA-RNA interaction prediction

Among the first approaches to predict the interacting base pairs for two RNA molecules are *concatenation-based* or *co-folding* approaches [31, 32]. Here, two or more RNA sequences are concatenated into a single sequence with special interspacing linker sequences. The resulting hybrid sequence is used within an adaptation of a standard structure prediction that takes special care of the linker sequences. The linked sequences are forbidden to form base pairs and the structural elements containing linker sequences are treated energetically as external as e.g., discussed by Ivo L. Hofacker and colleagues [31].

The extension of standard structure-prediction approaches to RNA-RNA interaction prediction directly yields the possibility to compute according probabilities of interaction sites or intermolecular base pairs [55]. A first implementation of concatenation-based prediction using the Nearest-Neighbor energy model was reported for mfold [47] and later implemented in e.g. the tools MultiRNAFold [56] and RNAcofold [55].

Our implementation extends the Nussinov recursion from Eq. 2 with a special handling for linker-sequence characters '`X`'. Base pairs (case 2) are not allowed to involve a linker position. No special energy treatment is necessary for the simplified energy model since we treat intra- and intermolecular base pairs equally and without considering their context. The input is restricted to two RNA sequences that are concatenated by a linker of length $l + 1$ (where l is the minimal loop size), to ensure the presence of a linker and that the concatenated sequence ends can form a base pair.

Our interactive co-folding web interface lists (sub)optimal hybridization structures using our generic suboptimal traceback implementation. Within the reported dot-bracket strings, intramolecular base pairs are encoded using parentheses '`()`', intermolecular base pairs (spanning the linker) are represented by brackets '`[]`', and the linker itself is depicted by linker characters '`X`'. For each hybridization structure, a traceback is visualized on selection along with a FORNA 2D structure graph visualization. Furthermore, an ASCII visualization of only the intermolecular base pairs is provided.

Concatentation-based approaches do incorporate the competition of intra- and intermolecular base pairing, which is a central weakness of hybridization-only prediction algorithms. Still, not all important interaction patterns can be predicted using co-folding approaches since the hybrid structure has to be nested. For instance, common kissing stem-loop or kissing hairpin interactions cannot be predicted because

they form a crossing structure in the concatenated model (see Fig. 4). To predict such patterns, accessibility-based approaches, discussed next, can be applied.

Fig 4. RNA-RNA interaction examples. (a) an interaction pattern that can be predicted by co-folding algorithms but not using standard accessibility-based methods, and a (b) kissing stem-loop or (c) kissing hairpin interaction pattern, both cannot be predicted by co-folding but using accessibility-based approaches. The RNA molecules are depicted in orange and blue while the linker is indicated in dotted green. Base pairs are illustrated in black.

Example Questions:

- Find RNA sequence pairs that show (i) only or (ii) no intermolecular base pairs within optimal structures. Study the suboptimals of the latter. Is it possible to find sequence pairs that do not prefer (among optimals) but still enable intermolecular base pairs (within suboptimals) using this model?
- Find example sequences for the interaction patterns from Fig. 4. For Figure 4b, find a sequence that can theoretically form all base pairs of the given pattern but no suboptimal prediction contains all pairs at the same time. Think of other patterns that cannot be predicted by concatenation-based approaches and try to find corresponding sequences.
- Find an RNA sequence pair that shows more intermolecular base pairs within optimal hybrid structures using a hybrid-only approach compared to a concatenation-based prediction. What is key to finding such sequences?

Accessibility-based interaction prediction

The previously introduced concatenation-based approaches directly reflect the competition of intra- and intermolecular base pairing by optimizing both at the same time. Nevertheless, they are neglecting that the intramolecular structure is established *before* an intermolecular interaction is formed. That is, intramolecular base pairs (might) have to be opened/broken such that intermolecular base pairs can form a stable interaction. To be favorable, the interaction energy must outweigh the energy needed to make the subsequences accessible. This two-step process is modeled by accessibility-based interaction prediction approaches.

The following formula, depicted in Fig. 5, is used to compute the final interaction energy values $I_{j,l}^{i,k}$ that incorporate both the hybridization/duplex energy D as well as the penalties $\Delta E^1, \Delta E^2$ for inaccessible sites of the RNAs S^1, S^2, respectively.

$$I_{j,l}^{i,k} = D_{j,l}^{i,k} + \Delta E_{i..k}^{1} + \Delta E_{j..l}^{2}. \tag{17}$$

Note, $\Delta E_{j..l}^{2}$ is computed for the reversely indexed sequence $\overleftarrow{S}^2$ to ease the notation. This reversal has to be taken into account for hybridization energy computations, since e.g. Nearest-Neighbor models have to incorporate the chemical 5'- to 3'-end orientation of RNAs. The entry of I with minimal energy is used to traceback the interaction

details of the optimal interaction. Only entries in I with an energy lower than zero mark favorable interactions, since here the duplex energy D outweighs the ΔE penalties to make the respective subsequences accessible.

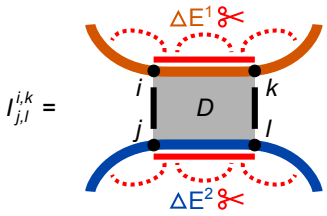

Fig 5. Depiction how accessibility-based approaches score an interaction of two RNAs S^1 and S^2 in orange and blue, respectively. The final interaction energy $I^{i,k}_{j,l}$ is only defined for subsequence combinations enclosed by two intermolecular base pairs $(i,j),(k,l)$ marked in black. It is composed of the duplex contribution $D^{i,k}_{j,l}$ (via intermolecular base pairs) shown in grey and the energy needed to break the intramolecular base pairing of each subsequence, i.e. $\Delta E^1 + \Delta E^2$, depicted in red.

The energy penalties $\Delta E_{i..j}$ resemble the free energy needed to make the interaction site $S_i..S_j$ accessible, i.e. to unfold the site's intramolecular base pairs [24, 33]. To reflect the structural flexibility of RNAs, the terms are based on the structure ensembles that can be formed rather than individual structures. The penalties can be computed from the energy difference of the structure ensemble with accessible site that is single stranded, $E^{\text{ss}}_{i..j}$, versus the whole structure ensemble, E^{ens}. Both energies can be computed from the respective partition functions $Z^{\text{ss}}_{i..j}$ (for $\mathcal{P}^{\text{ss}}_{i..j}$ from Eq. 12) and Z using the inverse Boltzmann weight. In the following, we show the relation of ΔE and the unpaired probability $\Pr^{\text{ss}}$.

$$\begin{aligned}\Delta E_{i..j} &= E^{\text{ss}}_{i..j} - E^{\text{ens}} \\ &= -\left(RT\cdot\log(Z^{\text{ss}}_{i..j}) - RT\cdot\log(Z)\right) \quad = \quad -RT\cdot\log(Z^{\text{ss}}_{i..j}/Z) \\ &= -RT\cdot\log(\Pr^{\text{ss}}(i,j)). \end{aligned} \tag{18}$$

Note, since $\Pr^{\text{ss}}(i,j)$ is ≤ 1, all $\Delta E_{i..j}$ penalties are ≥ 0.

To add such site-specific terms to duplex energies, we cannot simply use the prefix-based recursion from Eq. 16, since $H_{i,j}$ only provides the optimal value for all interaction sites with right ends S^1_i and $\overleftarrow{S}^2_j$ and not for individual sites. Thus, for exact results, we have to relate to a subsequence-based computation that explicitly stores values for all subsequence combinations. To further simplify the recursions, we use dedicated calculations (and matrices) for the duplex energy (matrix D, Eq. 19) and the overall interaction energy including inaccessibility penalties (matrix I, Eq. 17). Both matrices are four-dimensional, where an entry $D^{i,k}_{j,l}$ provides the duplex energy of the interacting sites $S^1_{i..k}$ and $\overleftarrow{S}^2_{j..l}$ under the assumption that the boundaries form the intermolecular base pairs (i,j) and (k,l); otherwise the entry is set to ∞.

$$D^{i,k}_{j,l} = \min\begin{cases} E_{bp} & S^1_i, \overleftarrow{S}^2_j \text{ compl.}, i=k, j=l \\ \min\limits_{\substack{i<p\leq k \\ j<q\leq l}} \left(E_{bp} + D^{p,k}_{q,l}\right) & S^1_i, \overleftarrow{S}^2_j \text{ compl.}, i<k, j<l \\ +\infty & \text{otherwise} \end{cases}. \tag{19}$$

The first case represents the initiation of a new interaction that covers only the intermolecular base pair (i,j) with according energy E_{bp}. The second case extends an already computed interaction of $S^1_{p..k}$, $\overleftarrow{S}^2_{q..l}$ with a new base pair (i,j), while the third

case is applied if the base pair (i, j) can not be formed or the indices violate order constraints. Note, the given recursion has an $O(n^6)$ time complexity due to arbitrarily large gaps in the second case. Given the typically applied thermodynamic model and statistics from known interactions, the sequential distance between neighbored intermolecular base pairs is normally restricted to a small constant < 30 [24], which reduces the time complexity to $O(n^4)$. The space complexity can be reduced to $O(n^2)$, as shown in [33], by interactively computing parts of D for a fixed right boundary base pair (k, l).

Our implementation provides the list of all optimal interactions and visualizes the selected interaction details using an ASCII chart. Due to the four-dimensionality of the matrices D and I, only the value $I_{j,l}^{i,k}$ for the current selection as well as the penalty tables $\Delta E^1 + \Delta E^2$ used for computation are shown.

The interactive web interface enables a straightforward comparison of the effects and restrictions of the three different interaction-prediction approaches introduced. For instance, using the simple example sequences $S^1 =$ `CCC` and $S^2 =$ `CCCGGGGGG`, the hybridization-only optimization reports (as expected) any interaction patterns of S^1 with `G` nucleotides of S^2. In contrast, intermolecular base pairs predicted by the co-folding approach are restricted to the 3'-end of S^2 since the central `G` nucleotides are blocked by an intramolecular hairpin structure (similar to Fig. 4a). Both approaches neglect that RNA S^2 will first (most probably) fold into a hairpin structure (with unpaired/accessible nucleotides in the center) before both interact. Thus it is most likely this central unpaired region of S^2 where interaction formation with S^1 will start. The growing interaction would have to break the already formed intramolecular base pairs for larger interaction patterns, which is not necessarily favorable. This scenario is modeled by accessibility-based approaches, which predict interactions to be restricted to the loop region only. The resulting interaction resembles a kissing stem-loop pattern (see Fig. 4b). Note, while accessibility-based approaches are well suited to predict interaction patterns like stem-loop or kissing-hairpin interactions, they are still not able to model arbitrary interaction patterns. For instance, double kissing-hairpin interactions can not be modeled correctly [57].

The first accessibility-based approach RNAup for the Nearest-Neighbor model was introduced by Ulrike Mückstein an colleagues [24]. While it is still among the state-of-the-art prediction tools [27], its vast runtime requirements of $O(n^4)$ render it inapplicable for large scale data analyses as e.g. genome wide target screens. This problem was tackled by Anke Busch and coworkers with IntaRNA [33, 34], which implements a heuristic version of an accessibility-based approach that extends fast hybridization-only recursions with ΔE penalties. IntaRNA results in a much lower $O(n^2)$ time complexity [33] when using precomputed or approximate ΔE terms as introduced in [58]. A detailed introduction is also given in [45]. A similar heuristic extension was recently reported for TargetRNA2 [59]. Current versions of the initially hybridization-only approach RNAplex [54] and its webserver RNApredator [60] incorporate an approximate, position-specific accessibility model to increase prediction quality [61].

Example Questions:

- Rewrite Eq. 19 to directly compute the final interaction energy values from Eq. 17.
- Why can interaction patterns enclosing intramolecular base pairs (see Fig. 4) not be predicted by the introduced basic accessibility-based approaches?

Implementation

All discussed algorithms and visualizations have been implemented in JavaScript. This enables client-side computation (no backend server hardware needed) as well as local download and application (from github repository) for offline usage. Since all algorithms are dynamic-programming approaches, a generic inheritance hierarchy was implemented to reduce code redundancy and to simplify maintenance and extensibility. We use `knockoutjs` as the controller to bind input/output elements from within the HTML pages with the JavaScript data structures and computations.

Conclusion

The understanding of RNA structure and RNA-RNA interaction prediction approaches is central to ensure correct result interpretation and an awareness of their limitations, both essential to avoid wrong conclusions. Furthermore, it ensures proper embedding in RNA-related analysis pipelines or their extension to new fields of applications.

To gain this level of understanding the original literature is often of limited didactic value since scientific articles are typically not meant for educational use. Thus, approaches are either represented on a very detailed expert level or sketched briefly since the manuscript focuses on the biological results rather than algorithmic details.

Here, we provide a compact summary of the relevant theoretical background for the most common algorithmic approaches and their state-of-the-art instances currently used. Algorithms are stripped from complicating energy-model details to enable an easy understanding of the underlying concepts and the resulting limitations. Furthermore, we provide web-based implementations and visualizations of all presented approaches for their ad hoc use. The latter is of importance, since example-driven (self-)study is known to significantly foster learning and understanding. To further support such self-learning efforts based on our manuscript and web-service, we provide small exemplary tasks for each algorithm group that can be tackled using our web-implementations.

The web-service [62] is being continually extended with the implementation and visualization of additional methods. Planned implementations cover pseudoknotted (crossing) structure prediction approaches as well as comparative approaches for RNA structure and RNA-RNA interaction prediction, e.g. discussed in [57].

Eventually, we provide both a comprehensive review of current RNA thermodynamic-focused prediction approaches to spark ideas for new approaches and interactive teaching material, which will help that available tools are correctly applied and interpreted.

Acknowledgments

We thank Florian Eggenhofer for helpful comments and Sita J. Saunders for her thorough language corrections.

References

1. Afgan E, Baker D, vandenBeek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research. 2016;44(W1):W3. doi:10.1093/nar/gkw343.

2. Grüning BA, Fallmann J, Yusuf D, Will S, Erxleben A, Eggenhofer F, et al. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. Nucleic Acids Research. 2017;45(W1):W560–W566. doi:10.1093/nar/gkx409.

3. Pevzner P, Shamir R. Computing Has Changed Biology - Biology Education Must Catch Up. Science. 2009;325(5940):541–542. doi:10.1126/science.1173876.

4. Qin H. Teaching Computational Thinking Through Bioinformatics to Biology Students. In: Proceedings of the 40th ACM Technical Symposium on Computer Science Education. SIGCSE '09. New York, NY, USA: ACM; 2009. p. 188–191.

5. Renkl A, Atkinson RK, Maier UH, Staley R. From Example Study to Problem Solving: Smooth Transitions Help Learning. The Journal of Experimental Education. 2002;70(4):293–315. doi:10.1080/00220970209599510.

6. Song L, Singleton ES, Hill JR, Koh MH. Improving online learning: Student perceptions of useful and challenging characteristics. The Internet and Higher Education. 2004;7(1):59 – 70. doi:10.1016/j.iheduc.2003.11.003.

7. Oliver J, Pisano ME, Alonso T, Roca P. The Web as an educational tool for/in learning/teaching bioinformatics statistics. Medical Informatics and the Internet in Medicine. 2005;30(4):255–266. doi:10.1080/14639230500367456.

8. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nature Reviews Genetics. 2014;15:469–479. doi:10.1038/nrg3681.

9. Backofen R, Engelhardt J, Erxleben A, Fallmann J, Grüning B, Ohler U, et al. RNA-bioinformatics: Tools, Services and Databases for the Analysis of RNA-based Regulation. Journal of Biotechnology. 2017;261:76–84. doi:10.1016/j.jbiotec.2017.05.019.

10. Wright PR, Richter AS, Papenfort K, Mann M, Vogel J, Hess WR, et al. Comparative genomics boosts target prediction for bacterial small RNAs. 2013;110(37):E3487–96. doi:10.1073/pnas.1303248110.

11. Lott SC, Schäfer RA, Mann M, Backofen R, Hess WR, Voss B, et al. GLASSgo – Automated and Reliable Detection of sRNA Homologs From a Single Input Sequence. Frontiers in Genetics. 2018;9:124. doi:10.3389/fgene.2018.00124.

12. Bellman RE. Dynamic Programming. 1st ed. Princeton, NJ, USA: Princeton University Press; 1957.

13. Giegerich R. A Declarative Approach to the Development of Dynamic Programming Algorithms, Applied to RNA Folding. Bielefeld University; 1998.

14. Giegerich R, Meyer C, Steffen P. Towards A Discipline of Dynamic Programming. In: Informatik bewegt. vol. P-19 of GI-Edition - Lecture Notes in Informatics; 2002. p. 3–44.

15. Steffen P, Giegerich R, ADP-team;. `https://bibiserv.cebitec.uni-bielefeld.de/adp/adpapp.html`.

16. Clair CS, Visick JE. Exploring Bioinformatics: A Project-based Approach. Burlington, MA, USA: Jones & Bartlett Learning; 2013.

17. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press; 1998.

18. Clote P, Backofen R. Computational Molecular Biology: An Introduction. Mathematical and Computational Biology. Chichester: Jon Wiley & Sons; 2000.

19. Waterman MS, Smith TF. RNA secondary structure: a complete mathematical analysis. Mathematical Biosciences. 1978;42(3):257 – 266. doi:10.1016/0025-5564(78)90099-8.

20. Waterman M. Secondary Structure of Single-Stranded Nucleic Acids. In: Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y., 1:167 – 212; 1978. p. 167–212.

21. Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. Algorithms for Loop Matchings. SIAM J Appl Math. 1978;35(1):68–82. doi:10.1137/0135006.

22. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981;9(1):133–48. doi:10.1093/nar/9.1.133.

23. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990;29(6-7):1105–19. doi:10.1002/bip.360290621.

24. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA–RNA binding. Bioinformatics. 2006;22(10):1177. doi:10.1093/bioinformatics/btl024.

25. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006;22(14):e90. doi:10.1093/bioinformatics/btl246.

26. Amman F, Bernhart SH, Doose G, Hofacker IL, Qin J, Stadler PF, et al. In: Setubal JC, Almeida NF, editors. The Trouble with Long-Range Base Pairs in RNA Folding. Springer International Publishing; 2013. p. 1–11.

27. Umu SU, Gardner PP. A comprehensive benchmark of RNA–RNA interaction prediction tools for all domains of life. Bioinformatics. 2017;33(7):988. doi:10.1093/bioinformatics/btw728.

28. Dimitrov RA, Zuker M. Prediction of Hybridization and Melting for Double-Stranded Nucleic Acids. Biophysical Journal. 2004;87(1):215 – 226. doi:10.1529/biophysj.103.020743.

29. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA. 2004;10(10):1507–1517. doi:10.1261/rna.5248604.

30. Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, et al. Target prediction for small, noncoding RNAs in bacteria. Nucleic Acids Research. 2006;34(9):2791. doi:10.1093/nar/gkl356.

31. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie / Chemical Monthly. 1994;125(2):167–188. doi:10.1007/BF00818163.

32. Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH. Predicting oligonucleotide affinity to nucleic acid targets. RNA. 1999;5(11):1458–1469. doi:10.1017/S1355838299991148.

33. 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. doi:10.1093/bioinformatics/btn544.

34. Mann M, Wright PR, Backofen R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. Nucleic Acid Res. 2017;45(W1):W435–W439. doi:10.1093/nar/gkx279.

35. DeVoe H, Tinoco I. The stability of helical polynucleotides: Base contributions. Journal of Molecular Biology. 1962;4(6):500 – 517. doi:10.1016/S0022-2836(62)80105-3.

36. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res. 2010;38(Database issue):D280–2. doi:10.1093/nar/gkp892.

37. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms for Molecular Biology. 2011;6(1):26. doi:10.1186/1748-7188-6-26.

38. IUPAC-IUB Comm on Biochem Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents. Biochemistry. 1970;9(20):4022–4027. doi:10.1021/bi00822a023.

39. Thirumalai D. Native secondary structure formation in RNA may be a slave to tertiary folding. Proceedings of the National Academy of Sciences. 1998;95(20):11506–11508. doi:10.1073/pnas.95.20.11506.

40. Hofacker IL, Schuster P, Stadler PF. Combinatorics of RNA secondary structures. Discrete Applied Mathematics. 1998;88(1):207 – 237. doi:10.1016/S0166-218X(98)00073-0.

41. Tinoco Jr I, Borer P, Dengler B, Levin M, Uhlenbeck O, Crothers D, et al. Improved estimation of secondary structure in ribonucleic acids. Nature New Biology. 1973;246(150):40–41. doi:10.1038/newbio246040a0.

42. Borer PN, Dengler B, Tinoco I, Uhlenbeck OC. Stability of ribonucleic acid double-stranded helices. Journal of Molecular Biology. 1974;86(4):843 – 853. doi:10.1016/0022-2836(74)90357-X.

43. Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. 1999;49(2):145–65. doi:10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G.

44. Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics. 2015;31(20):3377. doi:10.1093/bioinformatics/btv372.

45. Gorodkin J, Ruzzo WL, editors. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. vol. 1097 of Methods in Molecular Biology. Totowa, NJ: Humana Press; 2014.

46. Markham NR, Zuker M. In: Keith JM, editor. UNAFold: software for nucleic acid folding and hybridization. Totowa, NJ: Humana Press; 2008. p. 3–31.

47. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research. 2003;31(13):3406. doi:10.1093/nar/gkg595.

48. Jaynes ET. Information Theory and Statistical Mechanics. Phys Rev. 1957;106:620–630. doi:10.1103/PhysRev.106.620.

49. Bernhart SH, Mückstein U, Hofacker IL. RNA Accessibility in cubic time. Algorithms for Molecular Biology. 2011;6(1):3. doi:10.1186/1748-7188-6-3.

50. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Websuite. Nucleic Acids Research. 2008;36(Webserver issue):W70. doi:10.1093/nar/gkn188.

51. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research. 2003;31(13):3423. doi:10.1093/nar/gkg614.

52. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970;48(3):443 – 453. doi:10.1016/0022-2836(70)90057-4.

53. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981;147(1):195 – 197. doi:10.1016/0022-2836(81)90087-5.

54. Tafer H, Hofacker IL. RNAplex: a fast tool for RNA–RNA interaction search. Bioinformatics. 2008;24(22):2657. doi:10.1093/bioinformatics/btn193.

55. Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, Hofacker IL. Partition function and base pairing probabilities of RNA heterodimers. Algorithms for Molecular Biology. 2006;1(1):3. doi:10.1186/1748-7188-1-3.

56. Andronescu M, Zhang ZC, Condon A. Secondary Structure Prediction of Interacting RNA Molecules. Journal of Molecular Biology. 2005;345(5):987 – 1001. doi:10.1016/j.jmb.2004.10.082.

57. Wright PR, Mann M, Backofen R. Structure and interaction prediction in prokaryotic RNA biology. Microbiol Spectrum. 2018;6(2). doi:10.1128/microbiolspec.RWR-0001-2017.

58. Bernhart SH, Hofacker IL, Stadler PF. Local RNA base pairing probabilities in large sequences. Bioinformatics. 2006;22(5):614–615. doi:10.1093/bioinformatics/btk014.

59. Kery MB, Feldman M, Livny J, Tjaden B. TargetRNA2: identifying targets of small regulatory RNAs in bacteria. Nucleic Acids Research. 2014;42(W1):W124. doi:10.1093/nar/gku317.

60. Eggenhofer F, Tafer H, Stadler PF, Hofacker IL. RNApredator: fast accessibility-based prediction of sRNA targets. Nucleic Acids Research. 2011;39(suppl_2):W149. doi:10.1093/nar/gkr467.

61. Tafer H, Amman F, Eggenhofer F, Stadler PF, Hofacker IL. Fast accessibility-based prediction of RNA–RNA interactions. Bioinformatics. 2011;27(14):1934. doi:10.1093/bioinformatics/btr281.

62. Raden M, Ali SM, Alkhnbashi OS, Busch A, Costa F, Davis JA, et al. Freiburg RNA tools - a central online resource for RNA-focused research and teaching. Nucleic Acids Research. 2018;doi:10.1093/nar/gky329.