Bioinformatics, 33(3), 2017, 382–389
doi: 10.1093/bioinformatics/btw621
Advance Access Publication Date: 30 September 2016
Original Paper

OXFORD

Structural bioinformatics

Sequence–structure relations of biopolymers

Christopher Barrett, Fenix W. Huang and Christian M. Reidys*

Biocomplexity Institute of Virginia Tech, Virginia Tech University, Blacksburg, VA, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 6, 2015; revised on May 16, 2016; accepted on September 26, 2016

Abstract

Motivation: DNA data is transcribed into single-stranded RNA, which folds into specific molecular structures. In this paper we pose the question to what extent sequence- and structure-information correlate. We view this correlation as structural semantics of sequence data that allows for a different interpretation than conventional sequence alignment. Structural semantics could enable us to identify more general embedded 'patterns' in DNA and RNA sequences.

Results: We compute the partition function of sequences with respect to a fixed structure and connect this computation to the mutual information of a sequence–structure pair for RNA secondary structures. We present a Boltzmann sampler and obtain the *a priori* probability of specific sequence patterns. We present a detailed analysis for the three PDB-structures, 2JXV (hairpin), 2N3R (3-branch multi-loop) and 1EHZ (tRNA). We localize specific sequence patterns, contrast the energy spectrum of the Boltzmann sampled sequences versus those sequences that refold into the same structure and derive a criterion to identify native structures. We illustrate that there are multiple sequences in the partition function of a fixed structure, each having nearly the same mutual information, that are nevertheless poorly aligned. This indicates the possibility of the existence of relevant patterns embedded in the sequences that are not discoverable using alignments.

Availability and Implementation: The source code is freely available at http://staff.vbi.vt.edu/fenixh/Sampler.zip

Contact: duckcr@vbi.vt.edu

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

1 Introduction

2015 is the 25th year of the human genome project. A recent signature publication (The 1000 Genomes Project Consortium, 2015) is a comprehensive sequence alignment-based analysis of whole genome nucleotide sequence variation across global human populations. Notwithstanding the importance of this achievement, there is the possibility of information encoded as patterns in the genome that current methods cannot discover.

In this paper we study the information transfer from RNA sequences to RNA structures. This question is central to the processing of DNA data, specifically the role of DNA nucleotide sequences being transcribed into RNA, stabilized by molecular folding. In a plethora of interactions it is this specific configuration and not the particular sequence of nucleotides (aside from, say small docking areas, where specific bindings occur) that determines biological functionality. We find that here are multiple sequences in the partition function of a fixed structure, each having nearly the same mutual information with respect to the latter, that are nevertheless poorly aligned. This indicates the possibility of the existence of relevant patterns embedded in the sequences that are not discoverable using alignments.

RNA, unlike DNA, is almost always single-stranded and all RNA is folded. (There are double-stranded RNA viruses.) Here we only consider single-stranded RNA. An RNA strand has a backbone made of alternating sugar (ribose) and phosphate groups. Attached to each sugar is one of four bases – adenine (**A**), uracil (**U**), cytosine (**C**), or guanine (**G**). There are various types of RNA: messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and many others. Recent transcriptomic and bioinformatic studies suggest the existence of numerous of so called non-coding RNA, ncRNAs, that is RNA that does not translate into protein (Cheng *et al.*, 2005; Eddy, 2001).

RNA realizes folded molecular conformations consistent with the Watson–Crick base as well as the wobble base pairs. In the following we consider RNA secondary structures, presented as diagrams obtained by drawing the sequence in a straight line and placing all Watson–Crick and Wobble base pairs as arcs in the upper half-plane, without any crossing arcs, see Figure 1.

DNA information processing refers to replication, transcription and translation. Additionally, RNA information processing includes replication (Koonin *et al.*, 1989), reverse transcription (from RNA to DNA in e.g. retroviruses; Temin and Mizutani, 1970) and a direct translation from DNA to protein (in cell-free systems, using extracts from *E. coli* that contains ribosomes; McCarthy and Holland, 1965; Uzawa *et al.*, 2002).

In the following we offer an alternative view of DNA–RNA information processing. We focus on the information transfer from DNA/RNA sequences to the folded RNA (after transcription). We speculate that the sequential DNA information may transcribe into single-stranded RNA in order to allow subsequent biological processes to interpret DNA data.

DNA data are viewed as sequences of nucleotides. We currently use sequence alignment tools as a means of arranging the sequences of DNA, RNA, or proteins to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences (Mount, 2004). Here we suggest that the transcription into RNA with the implied self-folding is a way of lifting DNA information to a new and different level: RNA structures provide sequence semantics.

In order to study this idea we consider the folding of RNA sequences into minimum free energy (mfe) secondary structures (Waterman, 1978). Pioneered by Waterman more than three decades ago (Smith and Waterman, 1978) and subsequently studied by Schuster *et al.* (1994) in the context of the RNA toy world (Schuster, 1997) there is detailed information about this folding. In particular we have fairly accurate energy values for computing loop-based mfe (Mathews *et al.*, 1999, 2004; Turner and Mathews, 2010) that are employed by the folding algorithms (Hofacker *et al.*, 1994; Zuker and Stiegler, 1981). More work has been done on loop-energy models in Do *et al.* (2006) and Mathews (2004). We plan on a more detailed analysis of the framework proposed here in the context of the MC-model (Parisien and Major, 2008).

McCaskill (1990) observed that the dynamic programming routines folding mfe structures allow one to compute the partition function of all possible structures for a given sequence. The partition function is tantamount to computing the probability space of structures that a fixed sequence is compatible with. Predictions such as base pairing probabilities are obtained in Hofacker *et al.* (1994) and Hofacker (2003) and are parallelized in Fekete *et al.*(2000). Ding and Lawrence (2003) and Tacker *et al.* (1996) derive a statistically valid sampling of secondary structures in the Boltzmann ensemble and calculate the sampling statistics of structural features.

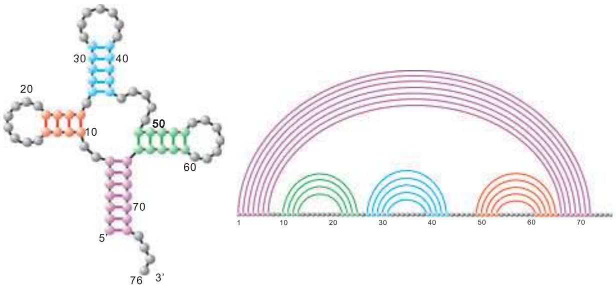


Fig. 1. tRNA: its secondary structure and its diagram presentation

In view of the above we are led to consider the 'dual' of McCaskill's partition function, i.e. the partition function of all sequences that are compatible with a fixed structure. More generally we consider the pairing

$$\varepsilon : \mathcal{Q}_4^n \times \mathcal{S}_n \to \mathbb{R}^+, \tag{1}$$

where $\mathcal{Q}_4^n$ and $\mathcal{S}_n$ denote the space of sequences, σ, and the space of secondary structures, S, respectively and $\varepsilon(\sigma, S) = e^{-\frac{\eta(\sigma,S)}{RT}}$ as well as the energy function $\eta(\sigma, S) \in \mathbb{R}$ are discussed in Section 2.1.

We show in Section 3 how ε allows us to capture the mutual information between sequences and structures, where the mutual information between x and y is given by

$$I(x, y) = \mathbb{P}(x, y) \log \left(\frac{\mathbb{P}(x, y)}{\mathbb{P}(x)\mathbb{P}(y)} \right).$$

Here $\mathbb{P}(x, y)$ denotes the joint probability distribution. In our case, $\mathbb{P}(\sigma,S) = \epsilon(\sigma,S) / \sum_{\sigma \in \mathcal{Q}_4^n, S \in \mathcal{S}_n} \epsilon(\sigma,S), \mathbb{P}(\sigma) = \sum_{S \in \mathcal{S}_n} \mathbb{P}(\sigma,S)$ and $\mathbb{P}(S) = \sum_{\sigma \in \mathcal{Q}_4^n} \mathbb{P}(\sigma,S)$.

In addition, ε allows us to express folding by considering

$$\{S | \varepsilon(\sigma, S) = \max_{S \in \mathcal{S}_n} \varepsilon(\sigma, S)\},$$

and inverse folding as to compute $\{\sigma | \varepsilon(\sigma, S) = \max_{S \in \mathcal{S}_n} \varepsilon(\sigma, S)\}$, for fixed S. Accordingly, the dual to folding is tantamount to computing for fixed S

$$\{\sigma | \varepsilon(\sigma, S) = \max_{\sigma \in \mathcal{Q}_4^n} \varepsilon(\sigma, S)\}.$$

This has direct implications to the 'inverse' folding of structures. Inverse folding is by construction about the sequence constraints induced by a fixed structure while avoiding competing configurations. Point in case: it has been observed in Busch and Backofen (2006), Levin, A., *et al.* (2012) and Reinharz,V., *et al.* (2013) that starting with a sequence that is mfe w.r.t. to a fixed structure, without necessarily folding into it, constitutes a significantly better initialization than starting with a random sequence.

The paper is organized as follows: we first recall in Section 2.1 the decomposition of secondary structures as well as the loop-based thermodynamic model. This in turn facilitates (Sections 2.3 and 2.4) the derivation of the partition function and Boltzmann sampling. In Sections 2.3 and 2.4 we compute $Q(S)$, Boltzmann sampling and the *a priori* probability of sequence patterns.

2 Method

2.1 Secondary structures and loop decomposition

RNA structures can be represented as diagrams where we consider the labels of the sequence to be placed on the x-axis and the Watson–Crick as well as Wobble base pairs drawn as arcs in the upper half plane see Figure 1. That is, we have a vertex-labeled graph whose vertices are drawn on a horizontal line labeled by $[n] = \{1, 2, \ldots, n\}$, presenting the nucleotides of the RNA sequence and the linear order of the vertices from left to right indicates the direction of the backbone from $5'$-end to $3'$-end. Furthermore each vertex can be paired with at most one other vertex by an arc drawn in the upper half-plane. Such an arc, (i, j), represents the base pair between the ith and jth nucleotide (here we assume $j - i > 3$ to meet the minimum size requirement of a hairpin loop.).

Two arcs (i, j) and (r, s) are called crossing if and only if $i < r$ and $i < r < j < s$ holds. An RNA structure is called pseudoknot-free, or secondary structure, if it does not contain any crossing arcs. Furthermore, the arcs of a secondary structure can be endowed with the partial order: $(r, s) \prec (i, j)$ if and only if $i < r < s < j$.

A filtration based on the individual contributions of base pairs of RNA structures was computed via the Nussinov model (Nussinov *et al.*, 1978). Smith and Waterman (1978) were the first bringing energy into the picture, computing the free-energy accurately via loops. A loop in a diagram consists of a sequence of intervals on the backbone $([a_i, b_i])_i$, $1 \leq i \leq k$, where (a_1, b_k), (b_i, a_{i+1}), for all $1 \leq i < k$ are base pairs. By construction, each base pair (i, j) is involved in exactly two loops: one where (i, j) is maximal respect to $\prec$, and one where (i, j) is not. Furthermore, there is a distinguished loop, L_{ex}, called the exterior loop, where $a_1 = 1$, $b_k = n$ and (a_1, b_k) is not a base pair. Depending on the number of base pairs, and unpaired bases inside a loop, a loop is categorized as hairpin-, containing exactly one base pair and one interval, helix, containing two base pairs and two empty intervals, interior-, containing two non-empty intervals and two base pairs, bulge-, containing two base pairs and two intervals, where one of them is empty and the other one is not and multi-loops, see Figures 2 and 3.

Further developments on RNA secondary structure prediction were given by Zuker and Stiegler (1981) and Hofacker *et al.* (1994). In particular, accurate thermodynamic energy parameters can be found in Mathews *et al.* (1999, 2004), Turner and Mathews (2010), Parisien and Major (2008), Deigan *et al.* (2009), Hajdin *et al.* (2013) and Lorenz *et al.* (2016).

In the following, we briefly recall the Turner energy model (Mathews *et al.*, 1999, 2004; Turner and Mathews, 2010) for RNA secondary structures. Let $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ be a sequence, where $\sigma_i \in \{A, U, C, G\}$ for all $1 \leq i \leq n$. To an arbitrary loop, L, we assign the energy $\eta(\sigma, L)$, where $\eta(\sigma, L_{ex}) = 0$ and $\eta(\sigma, L)$ depends on two factors: its type and the underlying backbone. Specifically this is the number of bases pairs, the number of unpaired bases and the particular nucleotides involved. The energy of a structure S over an RNA sequence σ is then given by the sum of the energies of individual loops i.e.

$$\eta(\sigma, S) = \sum_{L \in S} \eta(\sigma, L). \tag{2}$$

A hairpin, L_H is a loop having exactly one base pair with a non-empty interval containing k unpaired bases, where $k \geq 3$ due to flexibility constraints imposed by the backbone of the molecule.

In case of $3 \leq k \leq 4$ we call L a tetra-loop, which has a particular energy that depends on the two nucleotides incident to its unique arc $(\sigma_i, \sigma_{i+k+1})$ as well as the particular nucleotides corresponding to the unpaired bases of its unique non-empty interval $(\sigma_{i+1}, \ldots, \sigma_{i+k})$.

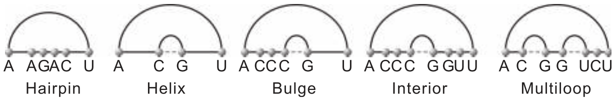


Fig. 2. Hairpin-, helix-, bulge-, interior- and multi-loops in secondary structures

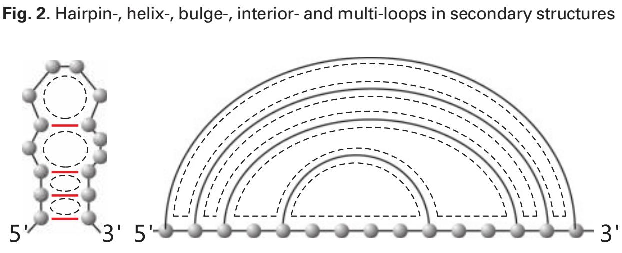


Fig. 3. Loops and their correspondence in a diagram

For any other number of unpaired bases, k, the energy calculation depends only on k and not the particular nucleotide sequence, except of $(\sigma_i, \sigma_{i+k+1})$ and σ_{i+1} and σ_{i+k}. We have

$$\eta(\sigma, L_H) = \begin{cases} \eta_H((\sigma_i, \sigma_{i+k+1}), \sigma_{i+1}, \ldots, \sigma_{i+k}) & \text{if } 3 \leq k \leq 4 \\ \eta_H((\sigma_i, \sigma_{i+k+1}), \sigma_{i+1}, \sigma_{i+k}, k) & \text{otherwise}\,. \end{cases} \tag{3}$$

An interior, bulge or helix loop, L_*, can be represented as two intervals and two base pairs $L_* = \{[i, r], [s, j], (i, j), (r, s)\}$. The energy of L_* is computed as

$$\eta(\sigma, L_*) = \begin{cases} \eta_*((\sigma_i, \sigma_j), (\sigma_r, \sigma_s)) \quad \text{(helix)} \\ \eta_*((\sigma_i, \sigma_j), (\sigma_r, \sigma_s), \sigma_{i+1}, \sigma_{r-1}, \\ \sigma_{s+1}, \sigma_{j-1}), k_1) \quad \text{(bulge)} \\ \eta_*((\sigma_i, \sigma_j), (\sigma_r, \sigma_s), \sigma_{i+1}, \sigma_{r-1}, \\ \sigma_{s+1}, \sigma_{j-1}), k_1, k_2) \quad \text{(interior)} \end{cases} \tag{4}$$

where $k_1 = \max\{r-i-1, j-s-1\}$ and $k_2 = \min\{r-i-1, j-s-1\}$.

A multi-loop L_M contains p base pairs and p intervals, some of which being possibly empty, where $p \geq 3$. $\eta_M(\sigma, L_M)$ is computed by

$$\eta_M(\sigma, L_M) = \alpha + p \cdot \beta + u \cdot \gamma. \tag{5}$$

Here α is the constant multi-loop penalty, β and γ are constants and u is the number of all unpaired bases contained in the respective intervals.

2.2 The partition function

Definition 1 *Let S be a secondary structure over n nucleotides. Then the partition function of S is given by*

$$Q(S) = \sum_{\sigma \in \mathcal{Q}_4^n} e^{-\frac{\eta(\sigma, S)}{RT}}, \tag{6}$$

where $\eta(\sigma, S)$ is the energy of S on σ, R is the universal gas constant and T is the temperature.

In analogy to the partition function of a fixed sequence $Q(\sigma)$ (McCaskill, 1990), $Q(S)$ can be computed recursively. Given the structure S, we consider an arbitrary arc (i, j), where $i < j$. Let $S_{i,j}$ denote the substructure of S over the interval $[i, j]$. Since S contains no crossing arcs all arcs of $S_{i,j}$ are contained in $[i, j]$, whence $S_{i,j}$ is well defined. Let

$$Q(\sigma_i, \sigma_j) = \sum_{\substack{\sigma \in \mathcal{Q}_4^{j-i+1} \\ \sigma|_i = \sigma_i, \sigma|_j = \sigma_j}} e^{-\frac{\eta(\sigma, S_{i,j})}{RT}}.$$

Since S has no crossing arcs, the interval $[i, j]$ is covered by the arc (i, j), i.e. (i, j) induces a loop L for which (i, j) is maximal. Suppose L consist of intervals $[i, p_1], [q_1, p_2], \ldots, [q_k, j]$, where $(p_1, q_1) \ldots, (p_k, q_k)$ are L-arcs different from (i, j). Removal of L renders substructures covered by $(p_1, q_1) \ldots, (p_k, q_k)$. Considering all combinations of the nucleotides in position p_i and q_i, $1 \leq i \leq k$, we derive the following recursion, see Figure 4:

$$Q(\sigma_i, \sigma_j) = \sum_{\sigma_{p_t}, \sigma_{q_t} \in \mathcal{Q}_4} e^{-\frac{\eta(\sigma, L)}{RT}} \prod_t^k Q(\sigma_{p_t}, \sigma_{q_t}). \tag{7}$$

The partition function $Q(S)$ is then obtained as the weighted sum of the terms $Q(\sigma_{a_t}, \sigma_{b_t})$, where (a_t, b_t), $\forall 1 \leq t \leq k$ are base pairs in the exterior loop L_{ex}:

$$Q(S) = \sum_{\sigma_{a_t}, \sigma_{b_t} \in \mathcal{Q}_4} e^{-\frac{\eta(\sigma, L_{ex})}{RT}} \prod_t^k Q(\sigma_{a_t}, \sigma_{b_t}). \tag{8}$$

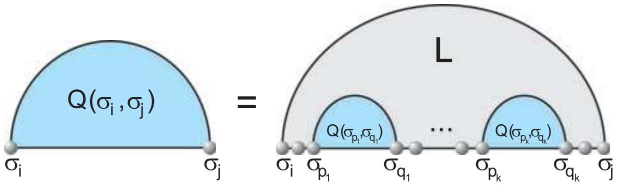


Fig. 4. The recursion for computing the partition function $Q_T(\sigma_i, \sigma_j)$

REMARK. *The routine of computing $Q(S)$ is similar to the one for finding an optimal sequence for a given structure inBusch and Backofen (2006), Levin, A., et al. (2012) and Reinharz,V., et al. (2013). Passing to a topological model for RNA structures (*Bon et al., *2008; Orland and Zee, 2002; Penner, 2004; Reidys et al., 2011), the above recursions can be extended to pseudoknotted RNA structures, i.e. RNA structures containing crossing arcs. The key here is a general bijection between maximal arcs and topological boundary components (loops).*

2.3 Boltzmann sampling and patterns

Having computed the partition function $Q(S)$ as well as the $Q(\sigma_i, \sigma_j)$ terms, puts us in position to Boltzmann sample sequences for fixed secondary structure S. Here the probability of a sequence σ to be sampled is given by

$$\mathbb{P}(\sigma|S) = \frac{e^{-\frac{\eta(\sigma,S)}{RT}}}{Q(S)}.$$

We build σ recursively from top to bottom, starting with the exterior loop, L_{ex}. Suppose (p_t, q_t) are base pairs contained in L_{ex} and let u denote the number of unpaired bases in L_{ex}. Since $\eta(\sigma, L_{ex}) = 0$, the unpaired nucleotides in L_{ex} are sampled uniformly, i.e. with probability 1/4. Then the probability of the event σ_r being the nucleotide in position $r \in L_{ex}$, is given by

$$\mathbb{P}(\sigma_r|S) = \frac{e^{-\frac{\eta(\sigma,L_{ex})}{RT}} \prod_{t=1}^{k} Q(\sigma_{p_t}, \sigma_{q_t})}{Q(S)} = \frac{\left(\frac{1}{4}\right)^u \prod_{t=1}^{k} Q(\sigma_{p_t}, \sigma_{q_t})}{Q(S)},$$

where the dependence on σ_r of the RHS stems from $\sigma|_r = \sigma_r$ or potentially $p_t = r$ or $q_t = r$. We continue the process inductively from top to bottom. Suppose we are given a loop L with the maximal base pair (i, j). Since any two arcs in S are not crossing, any arc (i, j) is contained in exactly two loops (except for the exterior loop) where (i, j) is the maximal arc for one and not for the other. As a result, the nucleotides σ_i, σ_j associated with (i, j) are sampled as part of the preceding loop (in which (i, j) is not maximal). It remains to sample the nucleotides other than σ_i and σ_j in L. Let σ_r be the nucleotides in L and $r \neq i, j$. The probability of the event σ_r being the nucleotide in position $r, r \neq i, j$ is given by

$$\mathbb{P}(\sigma_r|S) = \frac{e^{-\frac{\eta(\sigma,L)}{RT}} \prod_{t=1}^{k} Q(\sigma_{p_t}, \sigma_{q_t})}{Q(\sigma_i, \sigma_j)}.$$

Here (p_t, q_t), for $1 \leq t \leq k, k \geq 0$ are base pairs contained in L, that are different from (i, j). In particular, L is a hairpin loop in case of $k = 0$, an interior-, bulge- or a helix-loop in case of $k = 1$, and a multi-loop for $k \geq 2$.

By construction, for any arc there is a unique loop for which the arc is maximal and a unique loop where the arc is not. As a result, the probability of a sequence σ to be sampled is given by

$$\mathbb{P}(\sigma|S) = \prod_{(i,j)\in S} \frac{e^{-\frac{\eta(\sigma,L(i,j))}{RT}} \prod_{t=1}^{k} Q(\sigma_{p_t}, \sigma_{q_t})}{Q(\sigma_i, \sigma_j)} \cdot \frac{\left(\frac{1}{4}\right)^u \prod_{t=1}^{k} Q(\sigma_{p_t}, \sigma_{q_t})}{Q(S)}$$

In view of Eq. (2) and the fact that the term $Q(\sigma_{p_t,q_t})$ appears exactly once for each arc (p_t, q_t), we arrive at

$$\prod_{(i,j)\in S} (\prod_{t=1}^{k} Q(\sigma_{p_t}, \sigma_{q_t})) = \prod_{(i,j)\in S} Q(\sigma_i, \sigma_j).$$

This in turn implies

$$\mathbb{P}(\sigma|S) = \frac{\left(\prod_{(i,j)\in S} e^{-\frac{\eta(\sigma,L(i,j))}{RT}}\right)\left(\prod_{(i,j)\in S} Q(\sigma_i, \sigma_j)\right)}{Q(S) \prod_{(i,j)\in S} Q(\sigma_i, \sigma_j)} = \frac{e^{-\frac{\eta(\sigma,S)}{RT}}}{Q(S)}.$$

The time complexity for computing the partition function of a structure and Boltzmann sampling depends solely on the complexity of the energy function, $\eta(\sigma, L)$. Clearly, there are $O(n)$ loops in the structure and reviewing Eqs. (3), (4) and (5), at most eight nucleotides are taken into account. From this we can conclude that the time complexity is $O(n)$, as claimed.

Next, we compute the probability of a given sequence pattern, i.e. the subsequence of σ over $[i, j]$ being $p_{i,j}$. We shall refer to a sequence containing $p_{i,j}$ by $\sigma|_{p_{i,j}}$.

The partition function of all sequences σ containing $p_{i,j}$ is given by

$$Q(S|p_{i,j}) = \sum_{\sigma|_{p_{i,j}} \in \mathcal{Q}_4^n} e^{-\frac{\eta(\sigma,S)}{RT}} \tag{9}$$

and the probability of $p_{i,j}$ is $\mathbb{P}(p_{i,j}|S) = \frac{Q(S|p_{i,j})}{Q(S)}$.

We have shown how to compute $Q(S)$ recursively in Section 2.3. It remains to show how to compute $Q(S|p_{i,j})$. To do this we use the same routine as for computing $Q(S)$, but eliminating any subsequences that are not compatible with $p_{i,j}$. By construction, for any pattern, this process has the same time complexity as computing $Q(S)$.

3 Discussion

Let us begin by discussing the mutual information of sequence–structure pairs. Then we ask to what extent does a structure determine particular sequence patterns and finally derive a criterion that differentiates native from random structures.

The mutual information of a sequence–structure pair can be computed by normalizing ϵ

$$\mathbb{P}(\sigma, S) = \frac{e^{-\frac{\eta(\sigma,S)}{RT}}}{\sum_{\sigma\in\mathcal{Q}_4^n} \sum_{S\in\mathcal{S}_n} e^{-\frac{\eta(\sigma,S)}{RT}}},$$

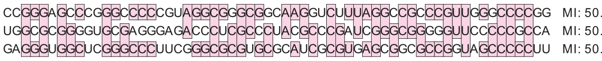
where $U = \sum_{\sigma\in\mathcal{Q}_4^n} \sum_{S\in\mathcal{S}_n} e^{-\frac{\eta(\sigma,S)}{RT}}$ is a constant. Then we have

$$I(\sigma, S) = \left(e^{-\frac{\eta(\sigma,S)}{RT}} \log \frac{(e^{-\frac{\eta(\sigma,S)}{RT}})}{Q(\sigma)Q(S)}\right)/U + \left(e^{-\frac{\eta(\sigma,S)}{RT}} \log U\right)/U.$$

Since U is a large constant, we observe that one term of $\mathbb{P}(\sigma, S)$, namely

$$e^{-\frac{\eta(\sigma,S)}{RT}} \log \frac{e^{-\frac{\eta(\sigma,S)}{RT}}}{Q(S)Q(\sigma)}$$

contributes the most. Accordingly, $Q(S)$ and $Q(\sigma)$ allow us to quantify how a probability space of structures determines a probability space of sequences.



CCGGGAGCCCGGGCCCCCGUAGGCGGGCGGCAAGGUCUUUAGGCCGCCCGUUGGGCCCCGG MI: 50.28
UGGCGCGGGGUGCGAGGGAGACCCUCGCCCUACGCCCGAUCGGGCGGGGGUUCCCCCGCCA MI: 50.01
GAGGGUGGCUCGGGCCCUUCGGGCGCGUGCGCAUCGCGUGAGCGGCGCCGGUAGCCCCCUU MI: 50.86

Fig. 5. Three sequences, having nearly the same mutual information with respect to the PDB structure 2N3R. The sequences differ pairwise by more than 50% of their nucleotides which indicates that there is information that cannot be captured by conventional sequence alignment. Accordingly BLAST outputs no significant homology between the sequences

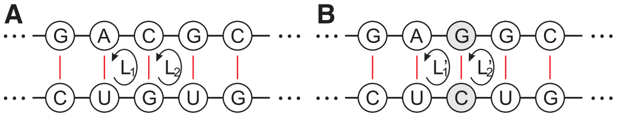


Fig. 6. Isolated replacement of **G- C** by **C- G**: (A) $L_1 = (\mathbf{U}, \mathbf{G}, \mathbf{C}, \mathbf{A})$ and $L_2 = (\mathbf{G}, \mathbf{U}, \mathbf{G}, \mathbf{C})$, (B) replacement induces the new loops: $L'_1 = (\mathbf{U}, \mathbf{C}, \mathbf{G}, \mathbf{A})$ and $L'_2 = (\mathbf{C}, \mathbf{U}, \mathbf{G}, \mathbf{G})$, which changes the free energy

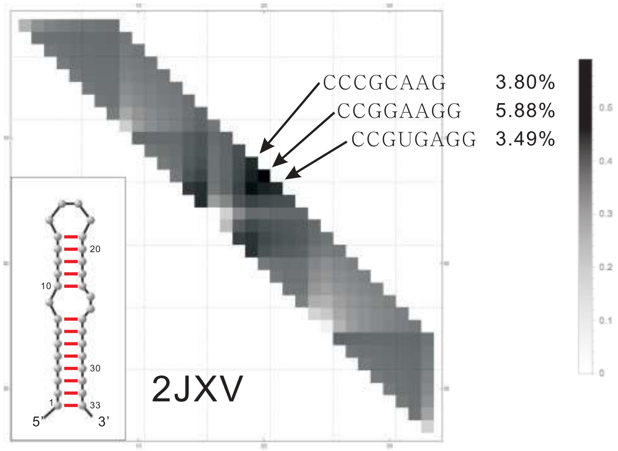


Fig. 7. The secondary structure of 2JXV and its heat-map. We display the most frequent sampled patterns for the largest interval having $R_{i,j} > 0.52$. The sample size is 10^4

In Figure 5 we display three sequences sampled from $Q(S)$ where S is the PDB-structure 2N3R (Bonneau *et al.*, 2015), see Figure 9. All three sequences have similar mutual information and more than 50% of the nucleotides in the sequences are pairwise different.

We point out that replacing a **G- C** base pair in a helix by a **C- G** base pair does change the energy, see Figure 6. This due to the fact that the loops are traversed in a specific orientation. The isolated replacement of **G-C** by **C-G** changes this sequence and hence the energy.

Since the energy model underlying the current analysis does not take non-canonical base pairs into account, we defer a detailed analysis of the mutual information to a later study where we use the MC-model (Parisien and Major, 2008).

Let $p_{i,j}$ be a subsequence on the interval $[i, j]$ with concrete nucleotides, having probability $\mathbb{P}(p_{i,j})$. Its Shannon entropy $E_{i,j}$ is given by

$$E_{i,j} = -\sum_{\forall p_{i,j}} \mathbb{P}(p_{i,j}) \log_4 \mathbb{P}(p_{i,j}).$$

By construction, $0 \leq E_{i,j} \leq (j - i + 1)$, where $E_{i,j} = (j - i + 1)$ when all $p_{i,j}$ have the same probability, i.e. uniformly distributed, and $E_{i,j} = 0$ when $p_{i,j}$ is completely determined, i.e. $\mathbb{P}(p_{i,j}) = 1$. Let $R_{i,j} = 1 - (E_{i,j}/(j - i + 1))$ be the heat of $[i, j]$, i.e. $R_{i,j} = 0$ for random sequences and $R_{i,j} = 1$ if there exists only one pattern $p_{i,j}$. We display the collection of $R_{i,j}$ as a matrix (heat-map), in which we display $R_{i,j} = 0.59$ as black and $R_{i,j} = 0$ as white. For a proof of concept, we restrict ourselves to $R_{i,j}$ for $j - i + 1 \leq 8$.

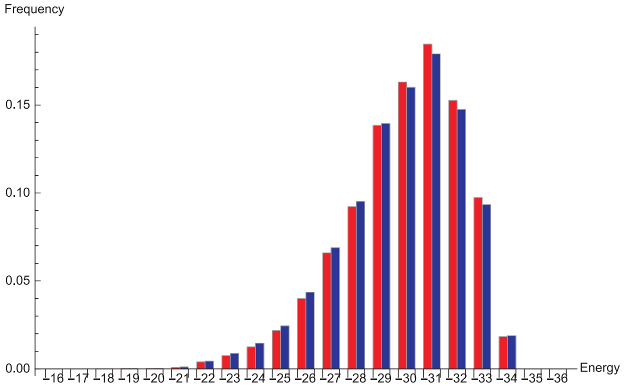


Fig. 8. The energy distribution of the Boltzmann sample for 2JXV. We display the frequency of sequences having a particular energy (right bars) and the frequency of sequences that fold into 2JXV (left bars)

The heat-maps presented here are obtained by Boltzmann sampling an ensemble of 10^4 sequences from $Q(S)$. We present the energy distribution of this ensemble in Figures 8, 10(A) and 12(A) and in addition the energy spectrum of those sequences that actually fold into S via the classic folding algorithm using the same energy functions here. The Inverse folding rate (IFR),

$$\text{IFR} = \frac{\#\,\text{of sequences folding into}\, S}{\#\text{of sampled sequences}}$$

measures the rate of successful re-folding from that ensemble.

Let σ be a sequence from a Boltzmann sample w.r.t. the structure S. Let $\bar{S}$ denote the structure that σ folds to. We consider

$$\Delta\eta(\sigma) = |\eta(\sigma, S) - \eta(\sigma, \bar{S})|$$

and compare the $\Delta G(\sigma, S)$ of several native structures contained in PDB with those of a several random structures (obtained by uniformly sampling RNA secondary structures).

The PDB structure 2JXV (Cevec *et al.*, 2008) represents a segment of an mRNA, having length 33. The structure exhibits a tetraloop, an interior loop and two stacks of length 8 and 5, respectively, see Figure 7. We Boltzmann sample 10^4 sequences for this structure observing an AU ratio of 18.18%, while CG ratio is 81.82%. The IFR reads 95.16%, i.e. almost all sampled sequences refold into 2JXV. The heat-map of 2JXV is given in Figure 7 and we list the most frequent 10 patterns of the largest interval having $R_{i,j} > 0.52$ in Supplementary Table S1 together with their *a priori* pattern probabilities. We observe that the tetra-loop determines specific patterns. This finding is not entirely straightforward as the hairpin-loops are the last to be encountered when Boltzmann sampling. I.e. they are the most correlated loop-types in the sense that structural context influences them the most.

The energy distribution of the Boltzmann sample is presented in Figure 8 and we observe that the inverse folding solution is not simply the one that minimizes the free energy w.r.t. 2JXV with the best energy. $\Delta\eta(\sigma)$-data are not displayed here in view of the high IFR.

The PDB structure 2N3R (Bonneau *et al.*, 2015) consist of 61 nucleotides and has a 3-branch multi-loop, two tetra-loops, interior loops and helixes, see Figure 9. The ratios of AU and CG pairs are 19.67% and 80.33%, respectively, again in a Boltzmann sample of

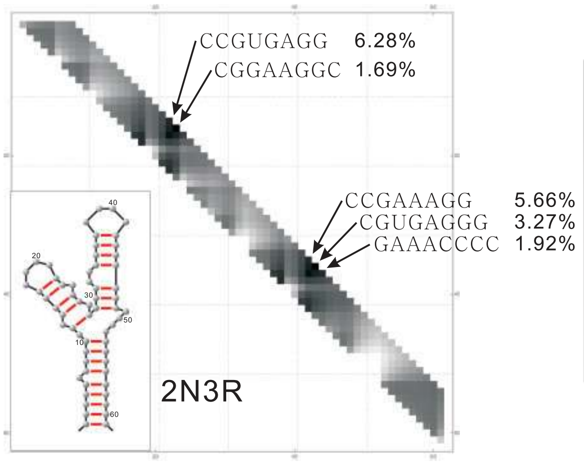


Fig. 9. The secondary structure of 2N3R and its heat-map. We show the most frequent patterns for the largest interval having $R_{i,j} > 0.52$. The sample size is 10^4

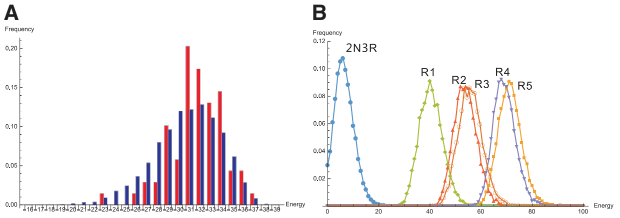


Fig. 10. (**A**) The energy distribution of Boltzmann sampled sequences. The frequency of sequences having a particular energy level (right bars), the frequency of sequences folding into 2N3R (left bars). (**B**) $\Delta\eta(\sigma)$-data of 2N3R versus $\Delta\eta(\sigma)$-data of five random structures

10^4 sequences. The IFR is at 0.69 quite high, despite the fact that 2N3R is much more complex than 2JXV. We illustrate the heat-map of 2N3R in Figure 9 and list the most frequent 10 patterns in the largest interval having $R_{i,j} > 0.52$ in the Supplementary Material, Supplementary Table S2 together with their *a priori* pattern probabilities computed by eq. (9)

Comparing the sequence segments [17, 24] and [37, 44], both of which being tetra-loops with additional two nucleotides. The $R_{i,j}$ values of these segments are similar, approximately 0.59, however, their most frequently sampled patterns appear at different rates. For [17, 24] this pattern is CGGAAGGC and it occurs with a Boltzmann sampled frequency of 1.69% and pattern probability 1.44%, while for [37, 44] it is CGUGAGGG with sampled frequency 3.27% and pattern probability 3.24%. This makes the point that pattern frequency distributions are strongly correlated with structural context.

The energy distribution of the Boltzmann sample is given in Figure 10(A) and we display the $\Delta\eta(\sigma)$-data in Figure 10(B) where we contrast the data with $\Delta\eta(\sigma)$-values obtained from Boltzmann sampling 10^4 sequences of 5 random structures of the same length. We observe that the $\Delta\eta(\sigma)$-values for 2N3R are distinctively lower than those for random structures.

The PDB structure 1EHZ (Shi and Moore, 2000) is a tRNA over 76 nucleotides exhibiting a 4-branch multi-loop. We display the heat-map of 1EHZ in Figure 11 The IFR is 1.3×10^{-3} w.r.t. our Boltzmann sample of size 10^4 and we display the energy distribution of the sampled sequences in Figure 12(A). Interestingly we still find many inverse fold solutions by just Boltzmann sampling $Q(S)$ and these sequences are not concentrated at low free energy values.

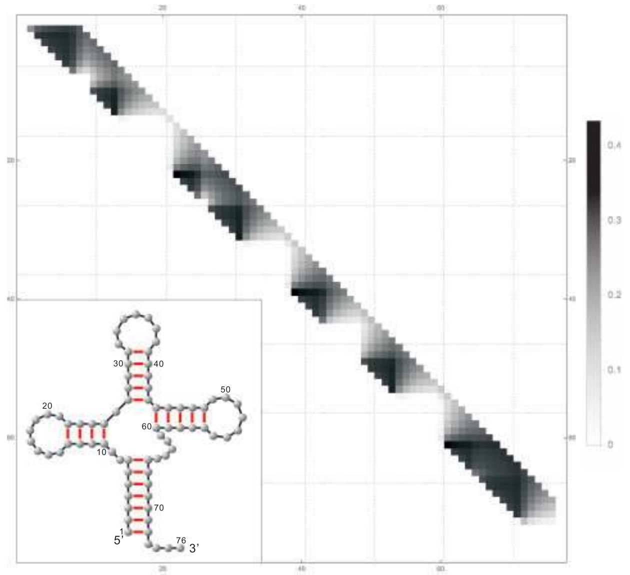


Fig. 11. The secondary structure of 1EHZ and its heat-map. We display the most frequent sampled pattern for the largest interval having $R_{i,j} > 0.52$. The sample size is 10^4

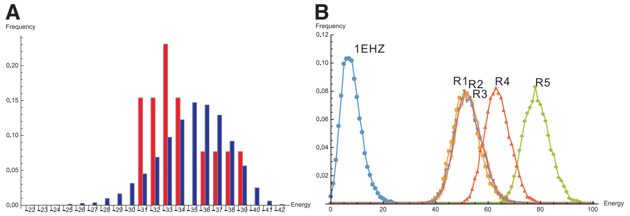


Fig. 12. (**A**) The energy distribution of the Boltzmann sampled sequences. The frequency of sequences having a particular energy level (right bars), the frequency of sequences folding into 1EHZ (left bars). (**B**) $\Delta\eta(\sigma)$-data of 1EHZ versus $\Delta\eta(\sigma)$-data of five random structures

In Figure 12(B) we display the $\Delta\eta(\sigma)$-data and contrast them with those obtained by the Boltzmann samples of five random structures. We observe a significant difference between the $\Delta\eta(\sigma)$-distribution of the 1EHZ sample and those of the random structures.

The three above examples indicate that sequence–structure correlations can be used to locate regions where specific embedded patterns arise. Furthermore we observe that studying $Q(S)$ has direct implications for inverse folding. This is in agreement with the findings in Busch and Backofen (2006), Levin, A., *et al.* (2012) and Reinharz,V., *et al.* (2013), but leads to deriving alternative, unbiased starting sequences for inverse folding. Although at present we can only estimate the mutual information, we can conclude that there are sequences that cannot be aligned but obtain almost identical mutual information.

We observe that biological relevant sequences exhibit a $\Delta\eta(\sigma)$-signature distinctive different from that of random structures. Therefore, the $\Delta\eta(\sigma)$-signature is capable of distinguishing biological relevant structures from random structures. In Miklós *et al.* (2005), the expected free energy and variance of the Boltzmann ensemble of a given sequence has been employed in order to distinguish biologically functional RNA sequences from random sequences. This result is in terms of the pairing $\varepsilon : \mathcal{Q}_4^n \times \mathcal{S}_n \rightarrow \mathbb{R}^+$, dual (the flip side of the coin, so to speak) to our approach. Our $\Delta\eta(\sigma)$-signature characterize the naturality of a fixed structure and Miklós *et al.* (2005) the naturality of a fixed sequence. Accordingly, $Q(S)$ augments the analysis of $Q(\sigma)$ in a natural way, capturing the correlation between RNA sequences and structures.

As a result, sequences carry embedded patterns that cannot be understood by considering the sequence of nucleotides. At this point we have no concept of what these patterns are and provided in Section 2.4 a rather conventional notion of 'embedded pattern'. However, even when considering specific nucleotide patterns in hairpin loops, we observe significant context dependence on the structure. Other loops affect the energy of the hairpin loop and thus determine this particular subsequence. We observe that the embedded patterns can, for certain structures, be quite restricted, possibly elaborate and are not entirely obvious. In any case, the analysis cannot be reduced to conventional sequence alignment. The heat-maps introduced here identify the regions for which only a few select patterns appear and computed the *a priori* probabilities of their occurrence.

This type of analysis will be carried out for the far more advanced MC-model (Parisien and Major, 2008), incorporating non-canonical base pairs, SHAPE-directed model for long RNAs (Deigan *et al.*, 2009; Hajdin *et al.*, 2013). This will in particular enable us to have a closer look at the hairpins of the tRNA structure. In addition we believe that this line of work may enable us to arrive at non-heuristic inverse foldings.

Folding of RNA secondary structures including pseudoknots is studied in Rivas and Eddy (1999) by extending the dynamic programming paradigm introducing substructures with a gap. The framework generates a particular, somewhat subtle class of pseudoknot structures, discussed in detail in Rivas and Eddy (2000). A specific, multiple context-free grammar (MCFG) for pseudoknotted structures is designed (Rivas and Eddy, 1999), employing a vector of nonterminal symbols referencing a substructure with a gap.

Our results facilitate the Boltzmann sample RNA sequences for pseudoknotted structures. Let $S_{i,j;r,s}$ denote a substructure with a gap where (i, j), (r, s) are base pairs and $Q(\sigma_i, \sigma_j; \sigma_r, \sigma_s)$ denote the partition function of $S_{i,j;r,s}$, then one can compute $Q(\sigma_i, \sigma_j)$ following the MCFG given by Rivas and Eddy (1999).

A different approach was presented in Penner and Waterman (1993) and Penner (2004), where topological RNA structures have been introduced (Penner and Waterman, 1993; Penner, 2004). In difference to Rivas and Eddy (1999), which was driven by the dynamic programming paradigm, topological structures stem from the intuitive idea to just 'draw' their arcs on a more complex topological surface in order to resolve crossings. Random matrix theory (von Neumann and Goldstine, 1947) facilitates the classification and expansion of pseudoknotted structures in terms of topological genus (Bon *et al.*, 2008; Orland and Zee, 2002) and in Reidys *et al.* (2011) a polynomial time, loop-based folding algorithm of topological RNA structures was given. The results in this paper are for representation purposes formulated in terms of loops. However they were originally developed in the topological framework, in which loops become topological boundary components. This means that we can extend our framework to pseudoknot structures. The key then is of course to be able to recursively compute the novel partition function, i.e. an unambiguous grammar. Recent results (Huang and Reidys, 2016) associate a topological RNA structure with a certain, arc-labeled secondary structure, called λ-structure. The resulting disentanglement gives rise to a context free grammar for RNA pseudoknot structures (Huang and Reidys, 2016). (More precisely, a λ-structures corresponds one-to-one to a pseudoknotted structure together with some additional information, i.e. a specific permutation of its backbone.) We illustrate this correspondence in Figure 13. This finding facilitates to extend all our results to pseudoknotted structures and offers insight in patterns and inverse folding of more general RNA structure classes as well as RNA-RNA interaction complexes.

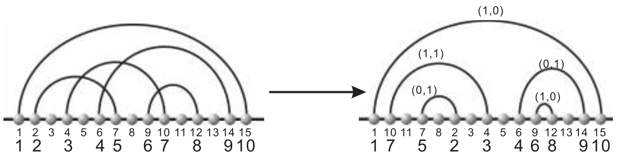


Fig. 13. Disentanglement: by means of permuting the backbone of a pseudoknotted structure one resolves all crossings

As mentioned above, the present analysis is just a first step and discusses embedded patterns in the sense of subsequent nucleotides. However our framework can deal with any embedded pattern. We think a deeper, conceptual analysis has to be undertaken aiming at identifying how a collection of structures provides sequence semantics. Quite possibly this can be done in the context of formal languages. We speculate that advancing this may lead to a novel class of embedded pattern recognition algorithms beyond sequence alignment.

Acknowledgments

Special thanks to Michael Waterman and Peter Stadler for their input on this manuscript. We gratefully acknowledge the help of Kevin Shinpaugh and the computational support team at VBI, Rebecca Wattam, Henning Mortveit, Madhav Marathe and Reza Rezazadegan for discussions. Many thanks for the constructive feedback of the anonymous reviewers and their suggestions.

Conflict of Interest: none declared.

References

Bon,M. *et al.* (2008) Topological classification of RNA structures. *J. Mol. Biol.*, **379**, 900–911.

Bonneau,E. *et al.* (2015) The NMR structure of the II-III-VI three-way junction from the neurospora VS ribozyme reveals a critical tertiary interaction and provides new insights into the global ribozyme structure. *RNA*, **21**, 1621–1632.

Busch,A. and Backofen,R. (2006) INFO-RNA–a fast approach to inverse RNA folding. *Bioinformatics*, **22**, 1823–1831.

Cevec,M. *et al.* (2008) Solution structure of a let-7 miRNA:lin-41 mRNA complex from *C. elegans*. *Nucleic Acids Res.*, **36**, 2330–2337.

Cheng,J. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.

Deigan,K.E. *et al.* (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 87–102.

Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.

Do,C. *et al.* (2006) Contrafold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.

Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.

Fekete,M. *et al.* (2000) Prediction of RNA base pairing probabilities on massively parallel computers. *J. Comput. Biol.*, **7**, 171–182.

Hajdin,C.E. *et al.* (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 5498–5503.

Hofacker,I.L. (2003) The vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.

Huang,F. and Reidys,C.M. (2016) Topological language for. *RNA*, arXiv:1605.02628.

Koonin,E.V. *et al.* (1989) Tentative identification of RNA-dependent RNA polymerases of dsRNA viruses and their relationship to positive strand RNA viral polymerases. *FEBS Lett.*, **252**, 42–46.

Levin,A. *et al.* (2012) A global sampling approach to designing and reengineering RNA secondary structures. *Nucl. Acids Res.*, **40**, 10041–10052.

Lorenz,R. *et al.* (2016) SHAPE directed RNA folding. *Bioinformatics*, **32**, 145–147.

Mathews,D. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mathews,D. *et al.* (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 7287–7292.

Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.

McCarthy,B.J. and Holland,J.J. (1965) Denatured DNA as a direct template for in vitro protein synthesis. *Proc. Natl. Acad. Sci. U. S. A.*, **54**, 880–886.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Miklós,I. *et al.* (2005) Moments of the Boltzmann distribution for RNA secondary structures. *Bull. Math. Biol.*, **67**, 1031–1047.

Mount,D.M. (2004). *Bioinformatics: Sequence and Genome Analysis*. 2nd edn. Cold Spring Harbor Laboratory Press, New York.

Nussinov,R. *et al.* (1978) Algorithms for loop matching. *SIAM J. Appl. Math.*, **35**, 68–82.

Orland,H. and Zee,A. (2002) RNA folding and large *n* matrix theory. *Nuclear Phys. B*, **620**, 456–476.

Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.

Penner,R.C. (2004). Cell decomposition and compactification of Riemann's moduli space in decorated Teichmüller theory. In: Tongring,N. and Penner,R.C. (eds.) *Woods Hole Mathematics-Perspectives in Math and Physics*, pp. 263–301. World Scientific, Singapore. arXiv: math. GT/0306190.

Penner,R.C. and Waterman,M.S. (1993) Spaces of RNA secondary structures. *Adv. Math.*, **101**, 31–49.

Reidys,C.M. *et al.* (2011) Topology and prediction of RNA pseudoknots. *Bioinformatics*, **27**, 1076–1085.

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

Rivas,E. and Eddy,S.R. (2000) The language of RNA: A formal grammar that includes pseudoknots. *Bioinformatics*, **16**, 334–340.

Reinharz,V. (2013) A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution. *Bioinformatics*, **29**, 308–315.

Schuster,P. (1997) Genotypes with phenotypes: adventures in an RNA toy world. *Biophys. Chem.*, **66**, 75–110.

Schuster,P. *et al.* (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Biol. Sci.*, **255**, 279–284.

Shi,H. and Moore,P.B. (2000) The crystal structure of yeast phenylalanine tRNA at 1.93 a resolution: a classic structure revisited. *RNA*, **6**, 1091–1105.

Smith,T. and Waterman,M. (1978) RNA secondary structure. *Math. Biol.*, **42**, 31–49.

Tacker,M. *et al.* (1996) Algorithm independent properties of RNA structure prediction. *Eur. Biophys. J.*, **25**, 115–130.

Temin,H. and Mizutani,S. (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, **226**, 1211–1213.

The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Turner,D. and Mathews,D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38**, 280–282.

Uzawa,T. *et al.* (2002) Polypeptide synthesis directed by DNA as a messenger in cell-free polypeptide synthesis by extreme thermophiles, *Thermus thermophilus* HB27 and *Sulfolobus tokodaii* strain 7. *J. Biochem.*, **131**, 849–853.

von Neumann,J. and Goldstine,H.H. (1947) Numerical inverting of matrices of high order. *Bull. Am. Math. Soc.*, **53**, 1021–1099.

Waterman,M.S. (1978) Secondary structure of single-stranded nucleic acids. *Adv. Math.*, **1**, 167–212.

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