Minimum-Free-Energy Distribution of RNA Secondary Structures: Entropic and Thermodynamic Properties of Rare Events

S Wolfsheimer
*Department of Applied Mathematics,
Université Paris Descartes,
45 rue des Saint-Pères
F-75270 Paris Cedex 06, France*

AK Hartmann
*Institut für Physik,
Carl von Ossietzky Universität Oldenburg,
D-26111 Oldenburg*

(Dated: June 15, 2010)

We study the distribution of the minimum free energy (MFE) for the Turner model of pseudoknot free RNA secondary structures over ensembles of random RNA sequences. In particular, we are interested in those rare and intermediate events of unexpected low MFEs. Generalized ensemble Markov-chain Monte Carlo methods allow us to explore the rare-event tail of the MFE distribution down to probabilities like 10^{-70} and to study the relationship between the sequence entropy and structural properties for sequence ensembles with fixed MFEs. Entropic and structural properties of those ensembles are compared with natural RNA of the same reduced MFE (z-score).

I. INTRODUCTION

Biopolymers such as DNA, RNA or proteins are heteropolymers meaning that they consist of different types of monomers connected through a "backbone" in a linear order. In the case of RNA, which is considered here, the monomers (called "nucleotides") consist of one out of four nitrogenous bases (adenine (A), cytosine (C), guanine (G) or uracil (U)), a ribose sugar, and a phosphate connected through phosphodiester bonds. The sequences of bases are referred as "primary structure"

In the last two decades fundamental knowledge about RNA has be achieved, in particular the fact that the transport of genetic information (via messenger RNA, or mRNA), where the relevant description is the primary structure, is only one out of many functions of RNA.

Nowadays it is well established that RNA also work as catalyst [1, 2] and regulator [3]. In particular in biochemical processes in the ribosomes so called ribosomal RNA (rRNA) plays a leading role in the translation process [4]. In a recent review [5] Bartel discussed the regulatory functions of so called microRNA. They are a non-coding RNA meaning that they are not translated directly to proteins. Instead, they bind to the transcribed mRNA and prevent it from beiing translated (*posttranscriptional gene silencing* or *RNA interference*).

Together with the change of the viewpoint of RNA playing an active biochemical role instead of a passive information carrier, the spacial conformation of the molecule has become of particular interest, because, in analogy to proteins, the three-dimensional structure, or tertiary structure, determines the molecule's function. Interestingly, the RNA structure prediction problem (the prediction of higher order structures from primary sequences) is conceptional simpler than protein folding, because the formation of secondary structures (the topology of the folded molecule in terms of paired bases) is energetically separated from the full three-dimensional structure [6]. This implies that the primary structure determines the secondary structure and, in contrast to the protein folding problem, the tertiary structure can then be seen as a perturbation to the secondary structure. For this reason the RNA secondary structure is already a meaningful description of the molecule.

Physically, RNA secondary structures can be seen as a disordered system (e.g. physical systems with random interaction) with rugged free-energy landscape [7–12]. In this context the sequence is considered as a random object and each particular realization induces a Gibbs-ensemble of possible structures. The low-temperature properties of the simple "pair-energy model" [13, 14], is suitable to understand the fundamental low-temperature properties of RNA. The energy landscape in such systems depends strongly on the sequence, and there is much evidence, that its ruggedness is due to the randomness [8].

However, from the biological point of view, the folding process is much more complex than described by those simplified models, and more sophisticated free-energy models had been introduced. Much effort had been made to adjust the parameters in order to increase accuracy in the secondary-structure prediction. Fortunately efficient algorithms for RNA secondary-structure prediction are not only available for the pair-energy model [13–15], but also for more realistic models [16] which, equipped with empirical free-energy parameters, are able to predict structures to an accuracy of $60 - 90\%$ in terms of correctly predicted base pairs [17]. Those algorithms neglect so called pseudoknots (see below) which is harmless in the case of RNA secondary structures. The software have been made available in the public domain. Two popular implementations are the program `mfold`, maintained by Michael Zuker [18] and the `vienna` package [19], maintained by Ivo Hofacker.

It turned out that the consideration of RNA sequences as purely random objects is not valid for natural biological sequences. In most cases, natural sequences have a lower minimum free energy than random sequences drawn from ensembles with similar statistical properties as the natural one (for example the same composition) [20–23]. Higgs has illustrated that natural tRNA sequences have a lower minimum free energy than purely random ones with the same composition [20] and also that the probability to find the minimum free energy (among all states) is larger for natural sequences at realistic physiological temperatures.

In the case of mRNA this issue has been discussed controversially. Where Seffens and Digby found evidence that natural mRNA are more stable than random ones [22], Workman and Krogh found contrary results [21] when considering random sequences with the same dinucleodic distribution as natural mRNA. This could be explained by the strong dependency of free-energy contributions on small local structures in the neighborhood of stacked base pairs [21, 23].

Another important observation [17, 20, 22] is that the minimum free energy is strongly correlated with the $C + G$ content of the sequence.

The evidence that a natural RNA sequence has a lower free energy than purly random ones was obtained by measuring the so called *z-score* of the minimum free energy of the natural sequence against the random ensemble. This quantity measures the distance of the observed free energy value $G_{\min}$ from the mean μ of the free-energy distribution over an ensemble in terms of standard deviations σ,

$$\text{z-score} := \frac{G_{\min} - \mu}{\sigma}\,. \tag{1}$$

The free-energy distribution is determined by a randomization of the natural sequence [20–23].

Here, we approach the problem from a different direction. Instead of comparing natural RNA against a reference ensemble characterized by the statistical properties (like the composition) we keep the (normalized) free energy fixed and compare entropic properties of natural RNA sequences against those of microcanonical sequence ensembles.

For example, one may ask how likely natural sequences are modeled by an i.i.d. (identically and independent distributed) sequence with uniform composition (each letter occurs with equal probability) given that the random and the natural sequences have the same minimum free energy (MFE). Since each sequence in a microcanonical ensembles occurs equally likely, one may check how likely a natural sequence is compatible with a maximum entropy principle.

To address this problem we have generated random sequences in generalized ensembles where each MFE occurs equally likely. This allows us to access the tail of the MFE distribution and to analyze properties of rare events by reweighting techniques. These techniques have been recently applied to study the tails of distribution of properties of random objects for several models [24–30].

Furthermore we compare these rare-event properties to those of natural rRNA sequences taken from a current database.

The article is organized as follows. We introduce the sequence and structure models in Sec. II. The simulation and analysis method are discussed in Sec. III followed by the results in Sec. IV.

II. MODELS

A. Sequence models

The space of RNA sequence is the set of all possible sequences of length L over the alphabet $\Sigma = \{A, C, G, U\}$. This space will be denoted as Σ^L.

For random sequences we have chosen a simple model of i.i.d. sequences. This means each letter $a \in \Sigma$ occurs with a fixed probability f_a ($f_a = 1/|\Sigma| = 1/4 \quad \forall a \in \Sigma$ here) independent of the other letters and of the position in the sequence. Hence the sequence $\mathbf{a}$ occurs with probability

$$p(\mathbf{a}) = p(a_1, \dots a_L) = \prod_{i=1}^{L} f_{a_i} = \frac{1}{|\Sigma|^L}\,.$$

Later on, we shall compare composition of natural RNA sequences against compositions of sequences in microcanonical ensembles. For this purpose we used the Bhattacharyya distance measure (BDM) [31–33] for two distributions p and q which is defined as

$$B(p||q) = \sum_i \sqrt{p(i)} \cdot \sqrt{q(i)}. \qquad (2)$$

The BDM measure fulfills the properties

- $0 \leq B(p||q) \leq 1$,
- $B(p||q) = 1$, if and only if $p = q$, and
- $B(p||q) = B(q||p)$.

This allows one to measure the distance of an observed normalized composition $\hat{f}(a) = \frac{1}{L}\sum_{j=1}^{L} \delta_{a,a_j}$ of a given sequence $\mathbf{a} = a_1 \ldots a_L$ to a "null" distribution $f_0(a)$

$$\hat{B} \;=\; B(\hat{f}||f_0) = \sum_{a=1}^{|\Sigma|} \sqrt{\hat{f}(a)}\sqrt{f_0(a)}.$$

The BDM alone does not provide a statistical interpretation in the spirit of a statistical hypothesis test [34]. Such statistical tests address significance of a certain observation with respect to a reference model. Under the assumption that the empirical distribution $\hat{f}$ is described asymptotically by f_0, the BDM deviates from 1 more likely for short sequences than for longer ones. In classical statistics one usually relies on so called p-values to assess the significance of a certain observation which allows for a statistical interpretation. That is the probability that an observed event (characterized by a test statistic, i.e. the BDM here) is at least as extreme as one would expect under the conditions of a *null model*. In the case of the interpretation of the BDM, the null model is given by an uniform histogram and the p-value of an observed BDM $\hat{B}_0$ is the probability that a BDM of $\hat{B}_0$ or smaller occurred by pure chance under the assumption that the null model is true. A p-value for a given $\hat{B}_0$, sample size and number of bins can be determined numerically [35] by generating independent histograms with fixed L and $|\Sigma|$ according to the null model (e.g. an uniform composition) and counting the fraction of events, where the BDM is smaller than $\hat{B}_0$. However, in our case the p-values might be very small, which means that the interesting events occur very unlikely. Therefore we implemented to method Wilbur's method [36] to compute very low p-values in combination with the BDM measure. Note that this method is very similar to *successive umbrella sampling* to compute free energy differences [37].

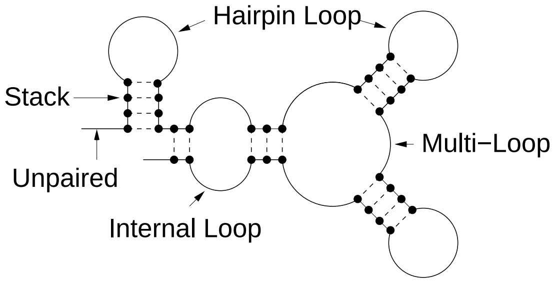

FIG. 1: RNA secondary-structure elements. Dots indicate bases, the backbone is illustrated by the solid line and and hydrogen bonds by broken lines. Each secondary structure can be decomposed into different elements.

B. RNA secondary structures

A secondary structure S of $\mathbf{a} = a_1 \ldots a_L \in \Sigma^L$ is a set of pairings pairs $\{(i_1, j_1), \ldots, (i_N, j_N)\}$, where each $i_k, j_k \in \{1, \ldots, L\}$ and $i_k < j_k$. $N = |\mathcal{S}|$ denotes the number of pairs. In the following, we frequently just write "structure", when we refer to secondary structures. The state space of all possible structures for a given realization $\mathbf{a}$ will be denoted as $\mathcal{S}_{\mathbf{a}}$.

For any two pairs $i < j$ and $k < l$ there are in principle three possible cases of order, choosing $i < k$ without loss of generality. They are either *nested* ($i < k < l < j$), *separated* ($i < j < k < l$), or crossing, also known as *pseudoknot* ($i < k < j < l$). In most studies of RNA secondary structures, it is assumed that pseudoknots occur on a different energy scale and hence are rather an element of the tertiary structure [6], thus being neglected. This means only nested or separated pairs are considered here.

Generally one may decompose every secondary structure into different *elements* (see Fig. 1). Of particular interest are *stacks*, i.e. a set of consecutive base pairs $(i, j), (i+1, j-1), \ldots, (i+n, j-n)$, where n is the size (or length) of the stack. They are important, because on one side, they stabilize the molecule, on the other side they decrease the entropy for loop formation.

A *free-energy model* assigns each structure $S \in \mathcal{S}_{\mathbf{a}}$ of a given realization $\mathbf{a} \in \Sigma^L$ a free energy $G(S, \mathbf{a})$. Note that the entropy contributions to the free energy do not arise from fluctuations of the structure. Instead they arise from the teritiary structure and are thus implicit in the model. Hence, the actual parameters of a free-energy model are obtained usually from measurements. Finding the structure S_0 that minimizes the free energy among all possible structures is referred as the RNA folding problem. Due to the planarity of RNA secondary structures, when omitting pseudoknots, this can be done in polynomial time $\mathcal{O}(L^3)$ using transfer matrix (called dynamic programming in computer science) methods [13–15].

C. The Turner free-energy model

In order to predict the secondary structure of natural biological sequences, complex models that account for different free-energy contributions from different loop types have been developed in the last two decades.

We used the `vienna` package, which provides an implementation of the Turner free-energy model [16, 17]. It incorporates hundreds of parameters which had been adjusted over the years. The model involves the entropic contributions for general loops

$$\Delta G_{\text{loop}} = -T\Delta S_{37,\text{loop}}$$

and a special Gibbs free-energy contribution for stacks

$$\Delta G_{\text{stacks}} = \Delta H_{37,stack} - T\Delta S_{37,\text{stack}}$$

(H denoting the enthalpy). The resulting minimum free energy of a sequence $\mathbf{a}$ is denoted by $G_{\text{min}}(\mathbf{a})$. The parameters had been determined experimentally (mainly via absorbance versus temperature melting curves [17, 38, 39]) at the standard physiological temperature $37°C$ and a given salt concentration and then improved by comparison of predicted structures with those known from phylogenetic analysis [40].

The free energy contributions of certain structure elements usually depend on the type, size and partially on the base composition. Since multiloops (loops that are bounded by more than two base pairs) are treated effectively and the size of loops are only considered up to a fixed length, the free-energy minimization algorithm is still of cubic time complexity.

III. SIMULATION AND REWEIGHTING METHOD

A. Importance sampling

Importance sampling is a general technique to reduce the variance in the estimation of expectation values [41]. In this framework, "interesting" events are generated more often by sampling from a different distribution and correcting for this bias afterward. This results in a more accurate estimate with a reasonable number of samples. Let $q : \Sigma^L \to [0,1]$ be an alternative distribution over the state space of sequences satisfying $q(\mathbf{a}) > 0$, whenever for the distribution of interest $p(\mathbf{a}) > 0$. Any expectation value with respect to p of an observable $A(\mathbf{a})$ can be estimated as

$$\begin{aligned} \langle A(\mathbf{a})\rangle_p &= \sum_{\mathbf{a}} A(\mathbf{a}) \cdot p(\mathbf{a}) \\ &= \sum_{\mathbf{a}} A(\mathbf{a}) \frac{p(\mathbf{a})}{q(\mathbf{a})} \cdot q(\mathbf{a}) \\ &\approx \frac{1}{z} \sum_{i=1}^{n} A(\mathbf{a}_i) \frac{p(\mathbf{a}_i)}{q(\mathbf{a}_i)} \end{aligned} \quad (3)$$

where each of the n samples $\mathbf{a}_i$ $(1 \leq i \leq n)$ is drawn from the sampling distribution q and z is the normalization constant $z = \sum_{i=1}^{n} \frac{p(\mathbf{a}_i)}{q(\mathbf{a}_i)}$. The variance in estimator Eq. 3 is reduced when the weights q are chosen such, that the probability mass $\frac{p(\mathbf{a}_i)}{q(\mathbf{a}_i)}$ is large in the region of interest [41].

In order to probe the MFE distribution over the sequence space Σ^L, in particular to access the rare-event tail we used the Wang-Landau sampling method [42] to estimate generalized ensemble weights. In the second stage the sequence space is explored with Monte-Carlo simulations with fixed weights.

The quantity of interest is the probability of the MFE of random sequences

$$P(G_{\text{min}}) = \sum_{\mathbf{a}\in\Sigma^L} p(\mathbf{a})\, \delta_{G_{\text{min}},G_{\text{min}}(\mathbf{a})}, \quad (4)$$

where $\delta_{G_{\text{min}},G_{\text{min}}(\mathbf{a})} = 1$ if G_{min} is the MFE of the sequence $\mathbf{a}$ and 0 otherwise.

Via importance reweighting it is also possible to determine mean values of observable depending on the deviation from the mean of the MFE distribution. This means, one considers ensembles, where more probability mass is put in either tail of the distribution. For this purpose one can either choose certain free-energy intervals to define such ensembles, or, in order to avoid binning effects and to obtain better statistics, define canonical-like ensembles with a certain inverse "temperature" Θ, where expectation values are defined as

$$\langle A\rangle_\Theta \approx \frac{1}{z_\Theta} \sum_{i=1}^{n} \frac{A(G^i_{\text{min}})}{q([G^i_{\text{min}}])} \cdot e^{-\Theta G^i_{\text{min}}}, \quad (5)$$

with $z_\Theta = \sum_{i=1}^{n} e^{-\theta G^i_{\text{min}}}/q([G^i_{\text{min}}])$. As a first step the temperature is tuned such that the expectation value of the free energy equals a desired value

$G_{\min} = \langle G^i_{\min}\rangle_\Theta$ and then the "canonical" average of the quantity of interest $\langle A\rangle_\Theta$ is computed. By choosing different values of Θ one may probe the entire free-energy range that has been sampled ($\Theta < 0$ probes the right tail above the mean and $\Theta > 0$ the left one) and relate $\langle A\rangle_\Theta$ to $\langle G_{\min}\rangle_\Theta$ via Θ, for the sake of simplicity denoted as $A(G_{\min})$ below.

B. Markov chain Monte Carlo of sequences

As in any Markov chain Monte Carlo (MCMC) method, the procedure employs an ergodic Markov chain whose stationary distribution converges towards the sampling distribution q. Let $N(\mathbf{a}) \subset \Sigma^L$ denote the *neighborhood* of $\mathbf{a}$ if each $\mathbf{b} \in \mathcal{N}(\mathbf{a})$ can be generated from $\mathbf{a}$ by one of the following operations

a) substitution at position k,

b) insertion at position k with left shift,

c) insertion at position k with right shift,

d) deletion at position k with left shift,

e) deletion at position k with right shift.

For the operation a) we replace the letter a_k with a letter $a \in \Sigma$. The operation b) involves a left shift of the sequence $a_1 \dots a_k$ (a_i is replaced by a_{i+1}) and an replacement of the letter a_k by a new letter and so on. Note that all sequences in $\mathcal{N}(\mathbf{a})$ have the same length and each operation involves a replacement of an existing letter with a newly drawn one, in case a) by a direct substitution and in the cases b)-e) indirectly via a shift operation, i.e., by deleting the first or last letter of the sequence.

In each step of the simulation a new state $\mathbf{b} \in \mathcal{N}(\mathbf{a})$ is proposed from the neighborhood of the current state $\mathbf{a}$. This proposal is accepted with the Metropolis criterion [43]

$$\alpha\left(\mathbf{a} \to \mathbf{b}\right) = \min\left\{1, \frac{q(\mathbf{b})}{q(\mathbf{a})}\right\}. \qquad (6)$$

If the detailed balance condition is fulfilled the chain converges to the stationary distribution q. For the model of i.i.d. sequence, detailed balance can be assured by drawing all new letters according to the probabilities f_a ($a \in \Sigma$) [24, 28].

C. Generalized ensemble Metropolis sampling

In the spirit of multicanoical or generalized ensemble methods [44], q should be chosen in such a way, that all realizations (from very high MFEs that sit far above the mean down to very low MFEs) occur with high probability in the simulation. Ideally, the distribution $P(G_{\min})$ is already known. In that case one can choose q as $q(\mathbf{a}) = q(G_{\min}(\mathbf{a})) \propto 1/P(G_{\min}(\mathbf{a}))$ which yields a "flat histogram" of MFEs over the entire range (denoted as q^{flat} in the following).

There are several ways to approximate q^{flat} iteratively. Here, we use the Wang-Landau method [42] because of its ease of use. Firstly an energy range of interest is chosen. The algorithm basically employs a histogram $H(G_{\min})$ and weights $q(G_{\min})$ defined on the desired range (we have chosen a bin size of 1 kcal/mol). Furthermore, a real valued parameter $\phi_j > 1$ is used in each iteration j. At each time step i the weights q are modified by $q(G^i_{\min})/\phi_j \to q(G^i_{\min})$, where $G^i_{\min}$ denotes the MFE of the sequence $\mathbf{a}_i$. Furthermore the histogram H is updated by one: $H(G^i_{\min}) + 1 \to H(G^i_{\min})$. This is continued until an "approximately flat histogram" is achieved. Let $\overline{H(G_{\min})}$ the average number of counts over the energy range. A possible flatness criterion might be that all $H(G_{\min})$ counts exhibit at least $\epsilon^{\text{WL}}\overline{H(G_{\min})}$, where ϵ^{WL} can be 0.8 for example.

Once the histogram is "flat", ϕ is decreased by the rule $\sqrt{\phi_j} \to \phi_{j+1}$ and all entries of the histogram H are set to 0 while q is kept for the next iteration $j + 1$.

Due to the decreasing rule $\sqrt{\phi_j} \to \phi_{j+1}$ the modification factor ϕ converges towards 1. The simulation is stopped, when ϕ reaches a value which is close to 1. It turned out that in our case the range from $\phi_0 = \exp(0.1) \approx 1.105$ to $\phi_{\text{final}} = \exp(0.0002) \approx 1.0002$ has been proven valuable.

Since detailed balance is violated explicitly the convergence of the algorithm can not be proven. For this reason one should always perform a simulation with $\phi = 1$ for data production, which is in fact the standard Metropolis algorithm with fixed weights q.

IV. RESULTS

A. The minimum-free-energy distributions of the Turner model

In this section we discuss the resulting MFE distributions obtained for the RNA sequences using the `vienna` package. Before presenting the results of the rare-event simulation, first the scaling properties of the mean, standard deviation and the skewness [34] are discussed.

Informal spoken, the skewness measures how much probability mass is located at either side of the mean. A positive (negative) value indicates the distribution to have more mass on the right (left) tail. It is defined as

$$\text{skewness} := \frac{\mu_3}{\sigma^3},$$

where $\mu_3 = \left\langle \left(\langle X - \langle X \rangle \rangle \right)^3 \right\rangle$ is the third moment about the mean and $\sigma = \sqrt{\left\langle \left(X - \langle X \rangle \right)^2 \right\rangle}$ the standard deviation of the distribution.

For this purpose it is sufficient to consider only data generated by simple sampling (randomly drawn i.i.d. sequences without importance sampling) which allows us to sample considerable larger sequences (up to $L = 1280$) than for the rare-event simulations. The sample size varied between $10,000$ for the smallest ($L = 40$) and 1300 for the largest system. The result is shown in Fig. 2.

The first moments $\langle G_{\min} \rangle_L$ and the standard deviation σ scale in analogy to previous studies [22] as

$$\langle G_{\min} \rangle_L = c_1 \cdot L + c_0 \,, \qquad \sigma[G_{\min}]_L = d \cdot L^\nu \,. \quad (7)$$

The resulting fit-parameters of a least-χ^2 fit are summarized in Tab. I and will be used in Sec. IV C.

The small skewness differs from other models with quenched disorder and long range interaction. For example, the long-range spin-glass exhibits ground-state energy distribution that can be described by a modified Gumbel distribution [26], i.e., a skewed distribution. Also for the ground-state-energy distribution of the pair-energy model introduced in [13–15], we found a different behavior (results not shown here). For this model we found positive skewed distributions.

For small sequences and not too small temperature, the skewness is much more negative. In all cases the skewness approaches 0 for large system sizes, which means that the distributions are essential symmetric in the high probability region. This can also be seen in the inset of Fig. 3, where the unscaled free-energy distributions for different temperatures are shown.

	c_0	c_1	d	ν
T=37 °C	8.9(4)	0.331(2)	0.51(1)	0.511(5)
T=0 °C	10.6(5)	0.691(4)	0.75(1)	0.498(3)
T=−100 °C	17.8(7)	1.842(5)	1.29(2)	0.494(3)

TABLE I: Fit parameters of a least square fit of the mean and standard deviation of the MFE distributions to the functional form Eq. 7.

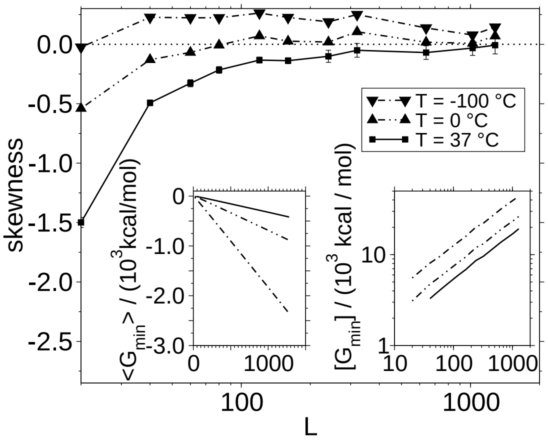

FIG. 2: Skewness of the minimum free-energy distribution over an random ensemble of i.i.d. sequences for different temperatures as a function of sequence length. Insets: scaling of the first moments and widths of these distributions with sequence length.

Since the computation of the minimum free energy is still of $\mathcal{O}(L^3)$ time complexity, the system sizes for the generalized ensemble simulation are restricted to relatively short sequences (in comparison to what can been achieved by simple sampling).

We used system sizes from $L = 20$ up to $L = 160$ [51]. Typically 10 Wang-Landau iterations (starting with $\phi_0 = 0.1$) where enough to achieve a suitable guess for the generalized ensemble simulations.

The main plot of Fig. 3 displays the MFE distributions obtained by the generalized ensemble simulation in a logarithmic scale. Note that using the Wang-Landau approach the distribution can be easily accessed in a region with probabilities as small as 10^{-70}. **Clearly, the distributions differ from a Gaussian distribution, which one would obtain for an energy function which is the sum of independent contributions. For the RNA secondary structures, in particular due to the existence of pseudo knots, each pair has a strong influence on the feasibility of other pairs, explaining the long tail of the MFE distribution observed.** The shape of the distributions at different temperatures differ slightly. Interestingly the one for lower temperature seems to be more symmetric, which is again in contrast to other models like the distribution of finite-temperature alignments that is discussed in Ref. [30]. **A simple explanation for the asymmetry of the distribution will be given in the next section.**

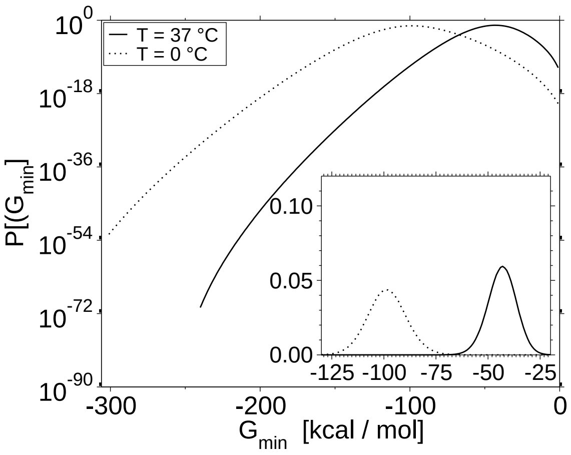

FIG. 3: Raw minimum-free-energy distributions at different temperatures for the largest system $L = 160$. The inset shows the same data with linear scale.

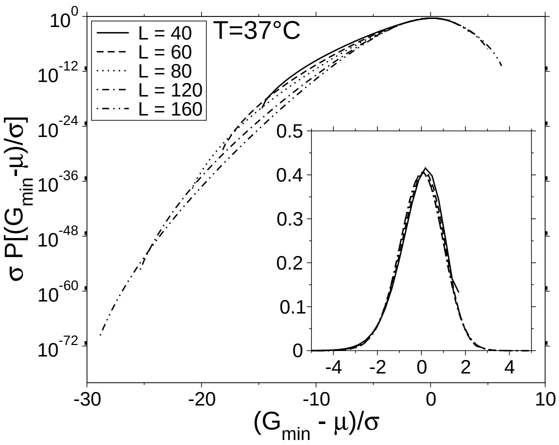

FIG. 4: Rescaled free-energy distributions at $T = 37°C$ for different system sizes. The inset shows the same data with linear scale.

In order to better understand the finite-size effects, the rescaled distributions for different system sizes and $T = 37°$ are shown in Fig. 4. For large probabilities and for the short tail the distributions collapse quite well. In the long tail some effects show up. Long sequences seem to have a given rescaled free energy less likely than short sequences (for intermediate values of the rescaled free energy $(G_{\min} - \mu)/\sigma$).

B. Entropy and thermodynamics of rare events

For the pair-energy model [7] we observed (not shown here), that the sequence composition is uniform in the left tail and highly nonuniform in the far right tail. This can be understood by entropic arguments: In order to achieve a low energy the sequence requires to have many complementary bases. Ideally the second half of the sequence consists of complementary partners of the first one in the same linear order. In this case the ground-state is just a single stack of size $L/2$. Such sequences exhibit an uniform composition, because one may choose the letters of the first half freely. In contrast, for a large ground-state energy, the sequence composition requires a huge amount of non-complementary bases, because the presence of a certain letter requires its complementary partner to occur rarely in the sequence.

In the same spirit, we analyzed the sequence ensembles that are biased towards very rare events of the MFE distribution. Here, in contrast to the simplified pair-energy model, the observed letter distributions were nonuniform in both tails, which is shown in the bottom of Fig. 5.

Also in Fig. 5 the functional dependence of $B(\hat{f}||f_0)$ with $f_0(a) = 1/|\Sigma| \quad \forall a \in \Sigma$ on $G_{\min}$ is shown. That means for each sample $\mathbf{a}_i$ the empirical composition $\hat{f}^i$ and the corresponding value of the BDM $B^i \equiv B(\hat{f}^i||f_0)$ was estimated. Then the canonical averages for different Θ's were determined using Eq. 5 and identified with $G_{\min}$, as explained in Sec. III A.

Close to the mean of the distribution the value B is also close to 1 as would be also expected from simple sampling. Far in the left tail the value shrinks, what is also supported by the form of the histograms that are shown in the bottom of the figure. In the right tail also nonuniform compositions are observed, implying B to deviate from 1.

The plots in Fig. 5 are labeled with the medians of the p-values of a BDM test of the observed microcanonical sequence ensembles (depending on $G_{\min}$) against an uniform letter composition.

Note, that to determine the histograms and the p-values we used binned free-energy intervals instead of the reweighting procedure. For that purpose the free energy range was divided into 50 bins for the largest system $L = 160$.

Sequences at the left end of the distribution essentially only consist of the bases C and G. This pair forms three hydrogen bonds, which explains why the resulting structures are very stable [17, 20, 22].

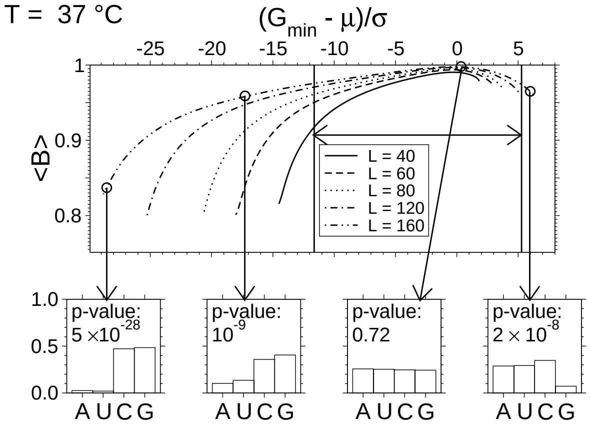

FIG. 5: top: Observed BDM as a function of the rescaled minimum free energy, for different system sizes. Nonuniform compositions are found in both tails. Vertical lines indicate the rescaled minimum-free-energy range of the selection of natural rRNA sequences (see Sec. IV C)
bottom: Histograms of observed compositions in different bins ($L = 160$), very far and far from the mean on the left side, close to the mean and far in the right tail. The medians of the corresponding p-values for the BDM-test (against a perfectly uniform composition) are written in the plots of the histograms.

The composition in the right tail seems to be unexpected at the first glance, in particular as it not only describes the average composition, but it also turned out that individual sequences in this region have a similar empirical letter frequency. Even though there are many $A - U$ Watson-Crick pairs available, the minimum free energy is relatively large. This is so because a loop needs to be closed by a stable pair, ideally by $C - G$.

The presence of Cs without the complementary partner G seems to *destabilize* the structure which can be supported by the following simple computer experiment on a sequence of length $L = 160$. First the sequence is initialized as $A^{L/2}U^{L/2}$, yielding a low minimum free-energy structure ($G_{\rm min} = -63.50$kcal/mol) consisting of a single large stack. Then the sequence is modified by randomly replacing letters with Cs. The minimum free energy increases rapidly with the concentration of C's and reaches $G_{\rm min} = 0$, when approximately every third letter is modified. On the other side, when repeating the experiment by replacing the letters with G instead of C a much higher fraction of replacements (approximately 70%) is necessary in order to achieve $G_{\rm min} = 0$.

By looking in the standard free-energy reference material, which was summarized by Mathews et.al. [17], this effect can be explained by penalty terms to the overall free energy for certain unstable secondary-structure motives. Noticeable are so called "olgio-C loops" and "tandem mismatches" (see Table 6. and Table 11. in ref. [17]). Olgio-C loops are hairpin-loops, in which all unpaired bases are C. Tandem mismatches are internal loops with two unpaired bases on each strand. Free-energy contributions of loops of this kind have different values depending on the types of the mismatches (unpaired letters) and on closing base pairs. Some combinations have negative contributions others have positive penalties. Cases, where tandem mismatches are closed by $A-U$ pairs and that contain $C - A$, $C - U$ or $C - C$ mismatches are penalized most. $A + U$ rich sequences that are "dotted" with C are entropically more favorable than sequences that contain only few complementary letters, which is the condition to achieve a large ground-state energy in the pair-energy model.

By the same entropic arguments we may also explain the fact that the tail for large MFEs is shorter than the one for low MFEs. In particular a MFE, of 0 is achieved with larger probability than the minmal possible MFE, because it is more likely to find a composition where one letter occurs rarely and the remaining equally likely **(see composition histograms in Fig. 5)**, than a compostion where two letter occur equally likely and the remaining ones very rarely.

The thermodynamics of rare sequences can be studied by looking not only at the sequences and values of the free energy but also at properties of the minimum-free-energy structures, which are also reported by the program `RNAfold`. Fontana et. al. [45] studied various of such quantities using simple sampling of random RNA sequences and compared the statistics of this ensembles with natural RNA sequences. One quantity that was considered in [45] is the distribution of stack sizes over the ensemble of MFE structures. Please remember that the stack size is the number of consecutive base pairs minus one, see Sec. II B. This is also used here in the biased ensembles.

Three typical structures that occur in the biased sequence ensemble are shown in Fig. 6. The underlying sequence of structure A has a typical $C + G$ rich composition, which occurs in the left tail of the MFE distribution. Large stabilizing stacks are characteristic for those sequences. Such structures are typical examples for so called precurson miRNA, pre-miRNA which are produced by from the primary transscript before leaving the nucelus. Although these structures are most stable, for some biological functions, they lack of important structural elements. The sequence with B

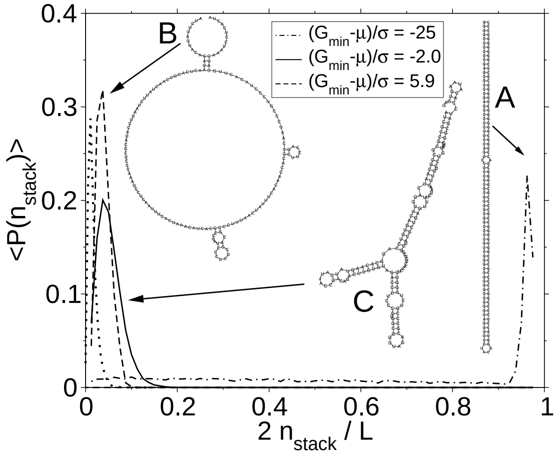

FIG. 6: Reweighted stack-size distribution as a function of $G_{\min}$ for $L = 160$ and typical structures in the generalized ensemble. The dotted line show the averaged stack-size distribution for the collection of natural rRNA sequences discussed in Sect IV C.

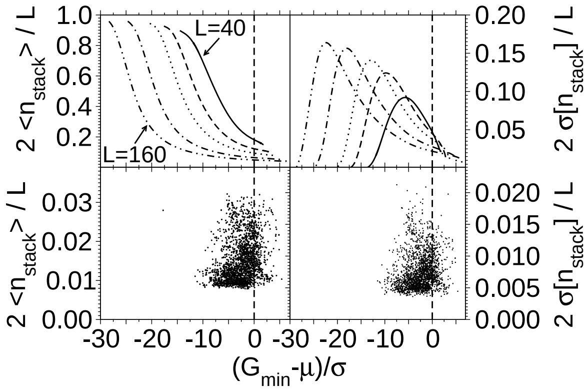

FIG. 7: Top: mean and width of the stack-size distribution normalized to sequence length as a function of the rescaled minimum free energy.
Bottom: Scatter plot of the mean and width of the stack-size distributions of 2078 natural rRNA sequences (see Sec. IV C) taken from the SilvaDB [46].

as minimum-free-energy structure was drawn from the rare event tail on the right side and consists of large loops, that are usually very unstable. More attractive is structure C, which has a free-energy of 2.0 standard deviations below the mean of the minimum-free-energy distribution.

Reweighted stack-size distributions (based on the method described in Sec. III A) for three values of the MFE is also shown in Fig. 6. In the ensemble of large minimum free energies only short stacks occur. For those sequences that have an extremely low minimum free energy, stack sizes on all length scales occur equally likely. Additionally a strong peak for stack sizes that are of the order of the half of the sequence length is observed. This reflects the observation of structure A, where a large stack is interrupted by a small internal loop.

Interestingly, the difference between the biological interesting free-energy range (slightly below the mean) and the extreme unstable region is not significant. However deviations up to $n_{\text{stack}} = 15$ become not as unlikely as for those sequences from the far right tail of the MFE distribution. The loop-size distribution (note shown here) seems to be a better description in order to characterize differences between the right tail and sequences from the left tail in an intermediate probability range, whereas the stack-size distribution distinguishes better very rare events from the left tail and typical sequences.

The mean stack size and the width of the stack-size distribution as function of the $G_{\min}$ is shown in the upper row of Fig. 7. The left plots indicate that only a small fraction of sequences have minimum free-energy structures that consist of a single stack in the order of the sequence length. Fontana et.al. [45] observed that the mean stack size converges to a length independent value of approximately 3 base pairs. By studying the width of the stack size distribution one also learns that the greatest variety of stack lengths occurs in very rare sequences.

Both, the composition of the sequences and the stack-size distribution is discussed under the viewpoint of natural biological sequences in the following.

C. Comparison between random and natural RNA sequences

The distribution of random RNA sequences allows one to gain more insight in the question in which sense natural RNA sequences differ from random i.i.d. sequences. Under the viewpoint of rare events in the sequence space we want to study thermodynamic and entropic aspects for natural ribosomal RNA sequences. For that purpose we randomly selected 2078 large subunit rRNA sequences from different species up to lengths $L = 1000$ from the SILVA database [46]. First of all, the minimum free energies of all sequences were obtained. In order to make the values of sequences of different lengths more comparable the MFE values have been rescaled by subtracting the average value and

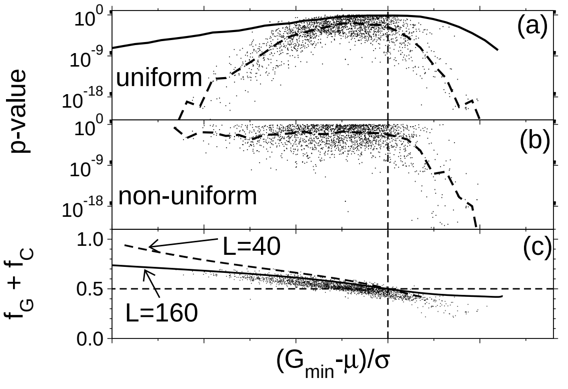

FIG. 8: (a) Dependence of the p-values of Bhattacharyya tests against an uniform letter composition on the rescaled minimum free energy using 2078 natural rRNA sequences (dots) taken from the SilvaDB [46]. The thin vertical dashed line indicates $g_{\min} = \mu$. The thick (non-linear) dashed line marks the median of the p-value of natural rRNA sequences. The solid Line indicates the median of the p-value of the random sequence model in the generalized ensemble (L=160).
(b) The p-values of a Bhattacharyya test of the composition of the natural sequences against compositions that occur at the same rescaled minimum free energies in the random sequence model. The thick dashed line indicates the median. The observed p-values are much smaller towards large free energies.
(c) The observed frequencies of $G + C$ as a function of minimum free energy.

then dividing by the standard deviation which are given by the scaling relations Eq. 7, using the fit parameters that are listed in Tab. I.

These rescaled free energies are the z-scores, see Eq. (1), with respect to the i.i.d. sequence ensemble for each sequence. Even for the simple assumption of uniform letter frequencies here, one observes (see Fig. 7 and Fig. 8) that most of the sequences are located below the mean in agreement with previous observations [17, 20–22].

In a similar way as for the random sequence ensemble, we performed Bhattacharyya test against an uniform letter distribution $f_0(a) = 1/|\Sigma|$ for each individual sequence and we found the relationship between p-values of the test and rescaled free energy energy as it is shown in Fig. 8(a).

Natural sequences that have a minimum free energy below the mean down to about 5 standard deviations (a z-score of -5), exhibit intermediate and large p-values (dots in Fig. 8(a)). This indicates that there is some evidence that all letters of those sequences occur (more or less) equally frequently. However in this region there are also realizations with relatively small p-values (down to $\sim 10^{-9}$), but these values are large, in comparison to sequences that are more than 5 standard deviations below the mean, where p-values down to $\sim 10^{-26}$ occur. Since the distribution of p-values is broad, we included their medians as a function of the rescaled free energy (dashed line).

Sequences above the mean are also very unlikely modeled by an uniform i.i.d. letter distribution which is also indicated by very small p-values of the natural sequences. We compared this with the random sequence model by calculating the dependence of the median of the p-values as function of deviation of the free energy from the mean, which is shown as solid lines in Fig. 8(a). The qualitative behavior resembles those of natural sequences. Numerical deviations are probably due the fact that the largest system for the random-sequence model was $L = 160$, whereas the natural sequences are explicitly longer.

The stabilizing effect of $C - G$ pairs shows up in the clear correlation between free energies and $C + G$ content, as shown in Fig. 8(c). In addition, the mean of the $C + G$ content of the random ensemble, shown by lines, tells us that the random model is suitable to explain low free energies due stabilizing $C - G$ pairs over a broad free energy range, as it is also observed in previous studies of natural RNA [17, 20, 22]. In order to support this argument, the statistical test of the sequence composition of the collection of natural sequences was repeated under the assumption of a different null hypothesis. That is the assumption, that the composition of a natural biological sequence is given by the mean composition of the random sequence model given the same rescaled MFE (called “microcanonical ensemble” here). Note that the histograms in Fig. 5 are 4 out of 50 different reference compositions. The test was performed by using frequency tables, obtained by binning the MFE range for $L = 160$ into 50 bins. These the empirical frequencies of the natural sequences were tested against those distributions. The corresponding p-values, see Fig. 8 (b), show a significant increasing of the values for low free energies in comparison to the original test against a perfectly uniform composition. Hence, for MFEs below the mean, the composition of the sequences in the rare-event simulations give a fairly good description of the composition of natural RNA with the same MFE. On the other side, for large free energies, no such observation could be made. Hence the assertion, that low free energies are strongly related to the $C + G$ content, is further confirmed.

At this point a few statements about the approximation of this test should be made. It is assumed that the composition is determined by the rescaled free energy alone and not on the sequence lengths (expect the scaling of the mean and the width). The sequence lengths are much larger for the collection of natural sequences. This assumption becomes reasonable, when comparing Fig. 5 with the scatter plot in Fig. 8. The rescaled free energies of the natural sequences (z-scores) range from -10 to 5. At least in the left tail, the finite size effects of the BDM are relatively small for lengths $L > 120$ in the biological relevant range of the rescaled free energies.

Note that the free energy parameters rely on the nearest neighbor model [47]. This means the $C+G$ content alone is only the leading effect to obtain a low free energy.

Obviously, natural sequences with relatively large MFEs do not compare well with random sequence model in terms of composition. In the latter one $A+U+C$ rich sequences are entropically favorable.

Regarding the stack sizes we find, in agreement with [45], no correlation between the value of the MFE and mean and standard deviation of the stack-size distribution, as shown in the bottom in Fig. 7. The biological relevant free-energy region is above the sequence length dependent threshold value, where stacks sizes are of the order of the sequence length. Also the maximum of the standard deviation, where the greatest variety of stack sizes is expected, sits below this region. **From Fig 6, where we compare the normalized averaged stack-size distribution of all natural rRNA sequences (dotted lines) with the reweighted stack-size distribution from the simulations, and from Fig 7 we also learn that finite size effects lead the normalized stack size $2n_{\mathbf{stack}}/L$ to be length dependent. Larger normalized stack-sizes seem to be more probable for shorter sequences. Qualitatively, the shape of the averaged stack-size distribution for natural rRNA agrees with the ones for intermediate low (say less than 10 standard deviations below the mean) or positive free energies.**

In analogy, we also checked for a possible correlation between the minimum free energy and other thermodynamic quantities, for example a measure for the non-extensive character of the free energy [10, 48, 49]. That is the difference between free energy of the entire sequence and the sum of the free energies of the first and the second half of the sequence, when it is broken exactly in the middle, $\Delta G_{\min} = G_{\min}(r_1, \ldots r_L) - G_{\min}(r_1 \ldots, r_{L/2}) - G_{\min}(r_{L/2+1} \ldots, r_L)$. Again, $\Delta G_{\min}$ is largest for very low free energies, but in the biological relevant region it remains small and is not correlated to the free energy of natural sequences. Also the mean loop size of structures of natural sequences does not correlate with the minimum free energy (not shown).

V. DISCUSSION AND OUTLOOK

To our knowledge, we have presented the first Monte-Carlo study of the rare-event tail of the MFE distribution of RNA secondary structures down to very small probabilities ($\approx 10^{-70}$).

Large-deviation properties of random RNA sequences are discussed. We illustrated how they can provide an additional classification of "randomness" of natural RNA sequences.

Properties of large deviations can be explained by entropic and thermodynamic arguments (Sec. IV B). As an entropic measure on the sequence level, the Bhattacharyya distance measure was used in order to discriminate observed sequences against the null-model with perfectly uniform composition, which is expected in the high probability region close to the mean. For the pair-energy model, the composition is uniform, even in the far left tail (low energies), whereas the composition deviates significantly from an uniform distribution in the right tail.

For the free-energy model, nonuniform compositions occur in both tails. The leading effect for stable structures in the left tail (low MFEs) is due to $G+C$ rich sequences. The destabilizing effect of $A+U+C$ rich sequences are responsible for unexpected large MFEs. These sequences are entropically favorable over such sequences that have many non-complementary bases.

In comparison to natural biological sequences, $G+C$ rich sequences also have the lowest minimum free energies, whereas many $A+U+C$ rich sequences are not found. One expects that all sequences in a microcanonical-like ensemble occur equally likely, due to the maximum entropy principle. ¿From the statistical tests of the natural sequences against those in the microcanonical ensemble one, which agree very well in our study constrained on low minimum free energies, one may infer that natural sequences are (more or less) compatible with entropy maximization. For large free energies this assumption seems not to be the case.

There is a plenty of room for further studies of the z-score statistics from this microcanonical perspective. Even though, at least in the left tail, the p-values have increased significantly when going from the uniform null distribution to the one obtained from microcanonical ensemble, they are still relatively small. For example, the median of p-values changes from 10^{-20} to 0.02 for the free-energy bin $(G_{\min} - \mu)/\sigma \approx -10$. One may change the sequence model from i.i.d. to model dinucleodic distribution, like in Ref. [21] or even more complicated shuffling procedures [50]. Possibly one would observe even larger p-values in the left tail. Eventually these models allow one to better describe the microcanonical sequences from the right tail as well. Similarly one may also modify the test statistics from the BDM to more complicated descriptions like Markov sequences instead of an i.i.d. model.

In a future work, it would also be of interest to not only look at rRNA as biolgical examples. Those sequences are essentially characterized by small MFE. In contrast, for pre-miRNA we would expect even smaller MFEs as their structures exhibits typically a very long stacks, similar to the structure A in Fig 6.

Acknowledgments

This project was supported by the German *VolkswagenStiftung* (program “Nachwuchsgruppen an Universitäten”) and by the European Community DYGLAGEMEM program. SW was partially supported by the University Paris Descartes. The simulations were performed at the GOLEM I cluster for scientific computing at the University of Oldenburg (Germany).

[1] T. R. Cech, A. J. Zaug, and P. J. Grabowski, Cell **27**, 487 (1981).

[2] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman, Cell **35**, 849 (1983), URL `http://www.sciencedirect.com/science/article/B6WSN-4C89C0V-2C/2/fe8fb940635298396e7f45463494bf0d`.

[3] J. E. G. McCarthy and C. Gualerzi, Trends in Genetics **6**, 78 (1990).

[4] H. F. Noller, Ann. Rev. Biochem. **60**, 191 (1991).

[5] D. P. Bartel, Cell **136**, 215 (2009), ISSN 0092-8674, URL `http://www.sciencedirect.com/science/article/B6WSN-4VF5R39-9/2/248c6ece0cad7816cfeb8e4200203c88`.

[6] I. Tinoco,Jr. and C. Bustamante, J. Mol. Biol. **293**, 271 (1999).

[7] P. Higgs, Phys. Rev. Lett. **76**, 704 (1996), URL `http://link.aps.org/abstract/PRL/v76/p704`.

[8] A. Pagnani, G. Parisi, and F. Ricci-Tersenghi, Phys. Rev. Lett. **84**, 2026 (2000).

[9] A. K. Hartmann, Phys. Rev. Lett. **86**, 1382 (2001).

[10] R. Bundschuh and T. Hwa, Phys. Rev. E **65**, 031903 (2002).

[11] M. F. Krzakala, Mézard, and M. Müller, Europhys. Lett. **57**, 752 (2002).

[12] M. Lässig and K. J. Wiese, Phys. Rev. Lett. **96**, 228101 (2006).

[13] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, J. Appl. Math. **35**, 68 (1978).

[14] R. Nussinov and A. B. Jacobson, Proc Natl Acad Sci U S A **77**, 6309?6313 (1980).

[15] P.-G. de Gennes, Biopolymers **6**, 715 (1968).

[16] M. Zuker and P. Stiegler, Nucl. Acids Res. **9**, 133 (1981).

[17] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, J. Mol. Biol. **288**, 911 (1999).

[18] M. Zuker, Nucl. Acids Res. **31**, 3406 (2003).

[19] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, Monatshefte für Chemie **125**, 167 (1994).

[20] P. Higgs, J. Phys. I France **3**, 43 (1993).

[21] C. Workman and A. Krogh, Nucl. Acids Res. **27**, 4816 (1999).

[22] W. Seffens and D. Digby, Nucl. Acids Res. **27**, 1578 (1999).

[23] P. Clote and R. Backofen, *Computational Molecular Biology: An Introduction* (John Wiley & Sons, Ltd., 2005).

[24] A. Hartmann, Phys. Rev. E **65**, 056102 (2002).

[25] A. Engel, R. Monasson, and A. Hartmann, J. Stat. Phys. **117**, 387 (2004).

[26] M. Körner, H. Katzgraber, and A. Hartmann, Stat.Mech. p. P04005 (2006).

[27] C. Monthus and T. Garel, Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) **74**, 051109 (pages 11) (2006), URL `http://link.aps.org/abstract/PRE/v74/e051109`.

[28] Wolfsheimer, S. and Burghardt, B. and Hartmann, A. K., Algor. Mol. Biol. **2**, 9 (2007).

[29] K. Hukushima and Y. Iba, Journal of Physics: Conference Series **95**, 012005 (2008).

[30] S. Wolfsheimer, O. Melchert, and A. K. Hartmann, Phys. Rev. E **80**, 061913 (2009).

[31] A. Bhattacharyya, Bull. Cal. Math. Soc. **35**, 99 (1943).

[32] F. C. Porter (2008), arXiv:0804.0380v1.

[33] S. Wolfsheimer, B. Burghardt, A. Mann, and A. K. Hartmann, Journal of Statistical Mechanics: Theory and Experiment **2008**, P03005 (2008), URL `http://stacks.iop.org/1742-5468/2008/i=03/a=P03005`.

[34] A. Hartmann, *Practical Guide to Computer Simulations* (World Scientific, Singapore, 2009).

[35] M.L.J.Scott, Tech. Rep. 2004-010, TINA (2004), URL `http://www.tina-vision.net`.

[36] W. Wilbur, Comp.Stat. **13**, 153 (1998).

[37] P. Virnau and M. Muller, The Journal of Chemical Physics **120**, 10925 (2004).

[38] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner, Proc Natl Acad Sci USA **83**, 9373?9377 (1986).

[39] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Mller, D. H. Mathews, and M. Zuker, Proc Natl Acad Sci U S A **27**, 9218 (1994).

[40] J. A. Jaeger, D. H. Turner, and M. Zuker, Proc Natl Acad Sci USA **86**, 7706?7710 (1989).

[41] J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, 2002).

[42] F. G. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

[43] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J.Chem.Phys. 21 **21**, 1087 (1953).

[44] B. A. Berg and T. Neuhaus, Phys.Rev.Lett. **68**, 9 (1992).

[45] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster, Biopolymers **33**, 1389 (1993).

[46] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner, Nucl. Acids Res. **35**, 7188 (2007).

[47] T. Xia, J. SantaLucia, M. Burkard, R. Kierzek, S. Schroeder, X. Jiao, C. Cox, and D. Turner, Biochemistry **37**, 14719 (1998).

[48] R. Bundschuh and T. Hwa, Europhys. Lett. **59**, 903 (2002).

[49] S. Hui and L.-H. Tang, Euro. Phys. J. B **53**, 77 (2006).

[50] P. Clote, F. Ferré, E. Kranakis, and D. Krizanc, RNA **11**, 578 (2005).

[51] Note that in studies [24, 28] the feasible system systems are much larger ($L \approx 800$) due to $\mathcal{O}(L^2)$ time complexity of the model that was considered there