Statistics of noncoding RNAs: alignment and secondary structure prediction

2011 J. Phys. A: Math. Theor. 44 195001

(http://iopscience.iop.org/1751-8121/44/19/195001)

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **44** (2011) 195001 (18pp) doi:10.1088/1751-8113/44/19/195001

Statistics of noncoding RNAs: alignment and secondary structure prediction

S K Nechaev[1,2,3], **M V Tamm**[4] **and O V Valba**[1,5]

[1] LPTMS, Université Paris Sud, 91405 Orsay Cedex, France
[2] P.N. Lebedev Physical Institute of the Russian Academy of Sciences, 119991 Moscow, Russia
[3] J.-V. Poncelet Laboratory, Independent University, 119002 Moscow, Russia
[4] Physics Department, Moscow State University, 119992 Moscow, Russia
[5] Moscow Institute of Physics and Technology, 141700 Dolgoprudny, Russia

E-mail: nechaev@lptms.u-psud.fr

Received 21 December 2010, in final form 22 March 2011
Published 14 April 2011
Online at stacks.iop.org/JPhysA/44/195001

Abstract

A new statistical approach to alignment (finding the longest common subsequence) of two random RNA-type sequences is proposed. We have constructed a generalized 'dynamic programming' algorithm for finding the extreme value of the free energy of two noncoding RNAs. In our procedure, we take into account the binding free energy of two random heteropolymer chains which are capable of forming the cloverleaf-like spatial structures typical for RNA molecules. The algorithm is based on two observations: (i) the standard alignment problem can be considered as a zero-temperature limit of a more general statistical problem of binding of two associating heteropolymer chains; (ii) this last problem can be generalized naturally to consider sequences with hierarchical cloverleaf-like structures (i.e. of RNA type). The approach also permits us to perform a 'secondary structure recovery'. Namely, we can predict the optimal secondary structures of interacting RNAs in a zero-temperature limit knowing only their primary sequences.

PACS numbers: 05.40.−a, 87.14.gn

(Some figures in this article are in colour only in the electronic version)

1. Introduction: noncoding RNAs and associating heteropolymers

According to a common definition, the noncoding RNA (ncRNA) is an RNA molecule that is not translated into a protein [1]. The ncRNAs either regulate the gene expression directly, for example by occupying the ribosome binding site, or indirectly providing RNA targeting specificity for a protein-based regulatory mechanism [2]. In general, regulatory RNAs act in the cell by one of the two basic mechanisms: by base-pairing interactions with other nucleic acids or by binding to proteins [1, 3, 4]. Thus, the base pairing with target molecules constitutes the typical mechanism, by which the ncRNA regulates the gene expression.

Since base-pairing of noncoding and target RNAs plays such important biological role, it is a worthy task to construct an algorithm, which, knowing the primary structures of each macromolecule, allows us to estimate theoretically the binding free energy of an ncRNA–target RNA complex. This problem resembles the one of alignment (or comparison) of two DNA sequences (or, more generally, two given sequences of letters) with one principal difference: in the ncRNA case one should align the sequences of nucleotides which constitute pairs between two RNAs, and also take into account the secondary structure of each RNA which comes into play by additional contribution to the total cost function.

In brief, the main goal of this work is to develop a constructive method to build a 'cost function', which characterizes a matching (alignment) of two noncoding RNAs with arbitrary primary sequences. To put the problem of alignment of ncRNAs into context of conventional statistical physics, it seems desirable to specify the basic features of ncRNAs which would play the major role in our analysis.

The ncRNAs are the specific examples of a wide class of so-called associating heteropolymers. Generally speaking, we call a polymer 'associating' if, besides the strong covalent interactions responsible for the frozen primary sequence of monomer units, it is capable of forming additional weaker reversible temperature-dependent (thermoreversible) bonds between different monomers. For associating polymers the variety of possible thermodynamic states ('secondary and ternary structures', in biological terminology) is determined by the interplay between the following three major factors: (i) the energy gain due to the direct 'pairing', i.e. formation of thermoreversible contacts; (ii) the combinatoric entropy gain due to the choice of which particular monomers (among those able to participate in bonds formation) do actually create bonds; (iii) the loss of conformational entropy of the polymer chain due to pairing (and in particular, the entropic penalty of loop creation between two paired monomers).

The RNA molecules differ from other biologically active associating polymers, such as, for instance, proteins (see [5] for a review), by a capability of forming mostly hierarchical 'cloverleaf-like' (or 'cactus-like') secondary structures. In other words, the formation of a thermoreversible contact between two distant monomers in an RNA (or in a single-stranded DNA) molecule imposes a nonlocal constraint on a number of possible thermoreversible bonds formed by other monomers: all bonds in an RNA chain are known to be arranged in a way to allow only hierarchical cactus-like folded conformations topologically isomorphic to a tree. The pairs of bonds, which do not obey such a structure are called 'pseudoknots' (see figure 2(*b*)); in most cases their formation in RNA molecules is highly suppressed. We shall not discuss here the reason why it happens, but rather accept the absence of pseudoknots as a matter of fact. Let us note however that in the work [6] the dynamic programming algorithm has been developed for predicting optimal RNA secondary structure, including pseudoknots.

Being formulated in statistical terms, the main goal of our work is as follows. We propose a new statistically justified algorithm for the determination of the binding free energy of two primary heteropolymer sequences allowing for the possibility for each sequence to form a hierarchical cactus-like secondary structure, typical for RNA molecules. Using this algorithm we can also predict optimal secondary structures of each interacting ncRNAs by knowing their primary sequences.

2. Theoretical background

Let us reveal the similarities and differences between computations of the free energy of associating heteropolymer complexes and standard matching algorithms.

The matching (or 'alignment') problem, even for linear structures is one of the key tasks of computational evolutionary biology. In particular, one of the most important applications of longest common subsequence (LCS) search in linear structures is a quantitative definition of a 'closeness' of two DNA sequences. Such a comparison provides information about how far, in evolutionary terms, two genes of one parent have deviated from each other. Also, when a new DNA molecule is sequenced *in vitro*, it is important to know whether it is really new or is it similar to already existing molecules. This is achieved quantitatively by measuring the LCS of the new molecule with other ones available from databases.

The task of this work consists of extending the statistical approach developed for alignment of linear sequences to the computation of pairing free energy of two RNA-type structures. The target object of our approach is the ground state free energy of complexes ncRNA–target RNA, or ncRNA–DNA.

2.1. Alignment of linear sequences

The problem of finding the LCS of a pair of linear sequences drawn from the alphabet of c letters is formulated as follows. Consider two sequences $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ (of length m) and $\beta = \{\beta_1, \beta_2, \ldots, \beta_n\}$ (of length n). For example, let α and β be two random sequences of $c = 4$ base pairs A, C, G, T of a DNA molecule, e.g. $\alpha = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{C}, \mathsf{T}, \mathsf{A}, \mathsf{C}\}$ with $m = 7$ and $\beta = \{\mathsf{C}, \mathsf{T}, \mathsf{G}, \mathsf{A}, \mathsf{C}\}$ with $n = 5$. Any subsequence of α (or β) is an ordered sublist of α (β) entries which need not be consecutive, e.g. it could be $\{\mathsf{C}, \mathsf{G}, \mathsf{T}, \mathsf{C}\}$, but not $\{\mathsf{T}, \mathsf{G}, \mathsf{C}\}$ for the α sequence. A common subsequence of two sequences α and β is a subsequence of both of them. For example, the subsequence $\{\mathsf{C}, \mathsf{G}, \mathsf{A}, \mathsf{C}\}$ is a common subsequence of both α and β. There are many possible common subsequences of a pair of initial sequences. The aim of the LCS problem is to find the longest of them. This problem and its variants have been widely studied in biology [7–10], computer science [11–14], probability theory [16–21] and more recently in statistical physics [15, 22–25].

The basis of dynamic programming algorithms for comparing genetic sequences has been formulated for the first time in [26] (see also [27]). In most general settings, this algorithm takes into account the number of perfect matches in the pair of sequences and also distinguishes between 'mismatches' and 'gaps'. Being formulated in statistical terms, it consists in constructing a 'cost function', F, which has a meaning of energy (see, for example, [28, 29] for details):

$$F = N_{\mathrm{match}} + \mu N_{\mathrm{mis}} + \delta N_{\mathrm{gap}}. \tag{1}$$

In equation (1), N_{match}, N_{mis} and N_{gap} are, respectively, the numbers of matches, mismatches and gaps for a given alignment of two sequences, and μ and δ are respectively the energies of mismatches and gaps (without a loss of generality, the energy of matches can be set to 1). Besides equation (1) we have an obvious conservation law

$$n + m = 2N_{\mathrm{match}} + 2N_{\mathrm{mis}} + N_{\mathrm{gap}} \tag{2}$$

which allows one to exclude N_{gap} from equation (1) and rewrite it as follows:

$$\begin{aligned} F &= N_{\mathrm{match}} + \mu N_{\mathrm{mis}} + \delta(n + m - 2N_{\mathrm{match}} - 2N_{\mathrm{mis}}) \\ &= (1 - 2\delta)N_{\mathrm{match}} + (\mu - 2\delta)N_{\mathrm{mis}} + \mathrm{const}. \end{aligned} \tag{3}$$

In equation (3) the irrelevant constant $\delta(n + m)$ can be dropped out.

Adopting $(1 - 2\delta)$ as a unit of energy, we arrive to the expression

$$\tilde{F} = N_{\mathrm{match}} + \gamma N_{\mathrm{mis}} \tag{4}$$

where

$$\gamma = \frac{\mu - 2\delta}{1 - 2\delta}, \tag{5}$$

and $\gamma \leqslant 1$ by definition. The interesting region is $0 \leqslant \gamma \leqslant 1$, otherwise there are no mismatches at all in the ground state (i.e. there is no difference between $\gamma = 0$, which corresponds to the simplest version of the LCS problem, and $\gamma < 0$).

The maximal cost function

$$\tilde{F}^{\max} = \max\left[N_{\text{match}} + \gamma N_{\text{mis}}\right] \tag{6}$$

can be computed recursively using the 'dynamic programming' algorithm [26–29]

$$\tilde{F}^{\max}_{m,n} = \max\left[\tilde{F}^{\max}_{m-1,n}, \tilde{F}^{\max}_{m,n-1}, \tilde{F}^{\max}_{m-1,n-1} + \zeta_{m,n}\right] \tag{7}$$

with

$$\zeta_{m,n} = \begin{cases} 1 & \text{for match} \\ \gamma & \text{for mismatch.} \end{cases} \tag{8}$$

It is known [28–30] that the statistical behavior of the matching cost function (7)–(8) in *random linear* sequences is substantially non-Gaussian. In particular, for a pair of linear sequences of lengths $m = n$ ($n \gg 1$) the fluctuations of maximal cost function (averaged over different realizations of sequences) grow as $n^{1/3}$. Moreover, in the previous study of one of us (SN) it was shown [30] that properly normalized asymptotic distribution of the LCS in a simplified version of the problem, known in the literature as a 'Bernoulli model', is given by the so-called Tracy–Widom (TW) distribution. The TW law has been derived first for the distribution of the highest eigenvalues of random matrices belonging to the Gaussian unitary ensemble [31][6].

2.2. *Matching versus pairing of two random linear heteropolymers*

The main goal of this paper consists in developing an algorithm for the computation of a cost function, which characterizes the similarity of two given RNA-type sequences. To succeed, we should incorporate the energetic and entropic contributions coming from the different rearrangements of intra-molecular bonds typical for RNAs into the conventional cost function discussed above. It is not obvious how to do that directly in the frameworks of the dynamic programming approach formalized in the recursion relation (7)–(8). To proceed, we exploit some trick (formulated for the first time in [33]), which consists of two consecutive steps.

First, we reformulate the recursion relation (7) in terms natural for statistical mechanical consideration and show that (7)–(8) can be regarded as a relation for the free energy of some statistical model describing the formation of a complex of two random linear heteropolymer chains in a zero-temperature limit. Second, we take into account the possibility for these heteropolymers of forming complex spatial cactus-like structures and write the corresponding recursion relations for the *partition function* (but not for the free energy) at some nonzero temperature T. By taking the limit $T \to 0$ at the very end we arrive at the desired cost function.

To accomplish the first of these two tasks, consider the following auxiliary statistical model describing the formation of a complex of two heteropolymer linear chains with arbitrary primary sequences. Let the chains be of lengths $L_1 = ma$ and $L_2 = na$, respectively. In what follows, we measure the lengths of the chains in numbers of monomers, m and n, which implies $a = 1$. Every monomer can be chosen from a set of c different types A, B, C, D, ... Monomers

[6] For a recent review of the appearance of Tracy–Widom distribution in several physics problems, see [32].

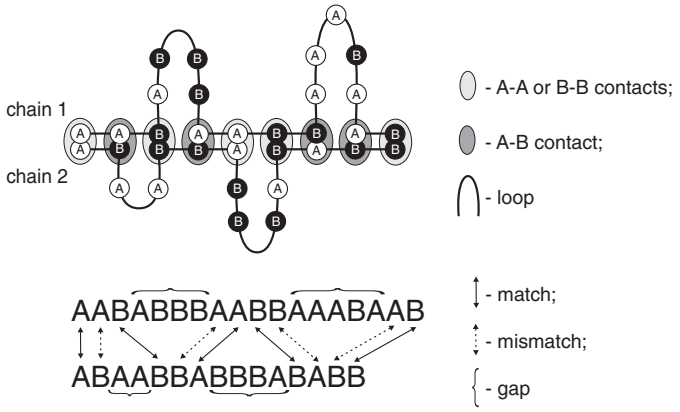

Figure 1. Schematic picture of a complex of two random linear heteropolymer chains with two types of letters ($c = 2$).

of the first chain could form saturating thermoreversible bonds with monomers of the second chain (the term 'saturating' here means that any monomer can form a bond with at most one monomer of the other chain). The bonds between similar types (like A–A, B–B, C–C, etc) have the attraction energy u and are called below 'matches', while the bonds between different types (like A–B, A–D, B–D, etc) have the attraction energy v and are called 'mismatches'[7]. Suppose also that some parts of the chains can form loops. These loops obviously produce 'gaps' since the monomers inside the loops of one chain have no matching (or mismatching) counterparts in the other chain. We give an example of a configuration of such a system with $c = 2$ in figure 1.

Our goal is to compute the free energy of the described model at sufficiently low temperatures, which allows us to assume that the entropic contribution of the loop formation is negligible compared to the energetic part of direct interactions between monomers. Let $G_{m,n}$ be the partition function of such a complex, i.e. the sum of contributions from all arrangements of bonds. In the low-temperature limit $G_{m,n}$ satisfies a recursion

$$\begin{cases} G_{m,n} = 1 + \sum_{i,j=1}^{m,n} \beta_{i,j} G_{i-1,j-1} \\ G_{m,0} = 1;\ G_{0,n} = 1;\ G_{0,0} = 1. \end{cases} \tag{9}$$

The meaning of equation (9) is straightforward. Starting from, say, left ends of chains (see figure 1) we find the first actually existing contact between the monomers i (of the first chain) and j (of the second chain) and sum over all possible arrangements of this first contact. The first term '1' in (9) corresponds to the arrangement with no contacts at all. The entries $\beta_{i,j}$ ($1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n$) are the statistical weights of bonds; they are encoded in a contact map $\{\beta\}$:

$$\beta_{m,n} = \begin{cases} \beta^+ \equiv e^{u/T} & i \text{ and } j \text{ match} \\ \beta^- \equiv e^{v/T} & i \text{ and } j \text{ do not match.} \end{cases} \tag{10}$$

[7] This general description covers both cases (DNA and RNA) by a straightforward redefinition of letters.

Straightforward computation shows that the partition function $G_{m,n}$ (9) obeys the following exact *local* recursion[8]:

$$G_{m,n} = G_{m-1,n} + G_{m,n-1} + (\beta_{m,n} - 1)G_{m-1,n-1}. \tag{11}$$

Now, write the partition function $G_{m,n}$ as $G_{m,n} = \exp\{F_{m,n}/T\}$, where $-F_{n,m}$ and T are the free energy and the temperature of the complex of two heterogeneous chains of lengths m and n. Considering the $T \to 0$ limit in equation (11), we get

$$F_{m,n} = \lim_{T\to 0} T \ln(e^{F_{m-1,n}/T} + e^{F_{m,n-1}/T} + (\beta_{m,n} - 1)e^{F_{m-1,n-1}/T}) \tag{12}$$

which can be regarded as an equation for the ground state energy of a chain. Equation (12) can be rewritten as

$$F_{m,n} = \max[F_{m-1,n}, F_{m,n-1}, F_{m-1,n-1} + \eta_{m,n}] \tag{13}$$

where

$$\begin{aligned} \eta_{m,n} &= T \ln(\beta_{m,n} - 1) \\ &= \begin{cases} \eta^+ = T \ln(e^{u/T} - 1) & \text{match} \\ \eta^- = T \ln(e^{v/T} - 1) & \text{mismatch.} \end{cases} \end{aligned} \tag{14}$$

Indeed, the ground state energy (13) may correspond either (i) to last two monomers connected, and then $F_{m,n}$ equals to $F^{\max}_{m-1,n-1} + \eta_{m,n}$, or (ii) to the unconnected end monomer of first (or second) chain, and then $F_{m,n}$ is $F^{\max}_{m,n-1}$ (or $F^{\max}_{m-1,n}$).

Taking η^+ as the unit of energy, one can rewrite (13) in a form identical to the dynamic programming equation (7):

$$\tilde{F}_{m,n} = \max[\tilde{F}_{m-1,n}, \tilde{F}_{m,n-1}, \tilde{F}_{m-1,n-1} + \tilde{\eta}_{m,n}] \tag{15}$$

with

$$\tilde{\eta}_{m,n} = \begin{cases} 1 & \text{in case of match} \\ a = \dfrac{\eta^-}{\eta^+} & \text{in case of mismatch} \end{cases} \tag{16}$$

(compare to (8)). The parameter a has a simple expression in terms of coupling constants u and v:

$$a = \frac{\eta^-}{\eta^+} = \left.\frac{\ln(e^{v/T} - 1)}{\ln(e^{u/T} - 1)}\right|_{T\to 0} = \frac{v}{u}. \tag{17}$$

The initial conditions for $\tilde{F}_{m,n}$ are transformed into $\tilde{F}_{0,n} = \tilde{F}_{n,0} = \tilde{F}_{0,0} = 0$.

Note that the model of heteropolymer binding described above is an auxiliary one and it bears only vague resemblance to the formation of real polymer–polymer complexes of linear geometry (on can think of a formation of a double-stranded DNA as, probably, the most familiar example). Indeed, in the partition function described by (9), a series of important features of real-life DNAs are neglected, namely

(i) the 'loop factors', i.e. the entropic penalty in the partition function due to forcing ends of side-loops (see figure 1) to meet again;
(ii) the cooperativity of bond formation, meaning that it is much easier to form a bond if there is another one between two adjacent monomers;

[8] Note that if $\beta_{i,j} = 2$ for all $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$, the recursion relation (11) generates the so-called Delannoy numbers [34].

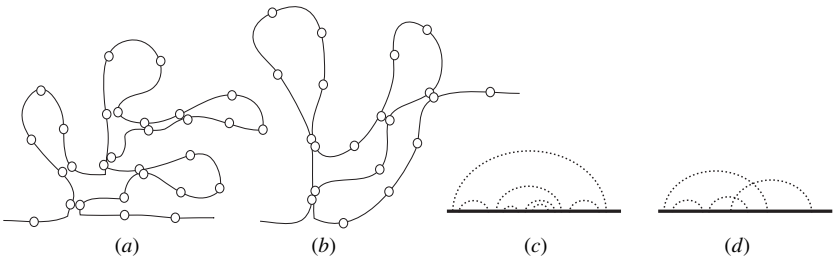

Figure 2. Schematic picture of allowed (*a*) cactus-like and prohibited (*b*) pseudoknot configurations of the bonds; (*c*), (*d*): arc diagrams corresponding respectively to configurations (*a*) and (*b*) (note the intersection of arcs in (*d*)).

(iii) the restriction on the minimal loop size, which takes into account the finite flexibility of polymer chains;
(iv) the fact that different matches (i.e. A–A versus B–B) can have different energies.

All these factors are known to affect the formation of double-stranded DNA structures (see [35], for example) but we would like to emphasize on the following essential point. The procedure leading from (9) to (15) is weakly sensitive to an exact form of equation (9). Only two crucial properties should be taken into account: (1) equation (9) is linear in G, which reflects the very fact that we are considering the binding of linear chains, and (2) the factor $\beta_{i,j}$ describing the bond formation is local. Therefore, since the cooperativity, the minimal loop weight and the variations in binding energies influence only local properties of chains, the procedure described by equations (9)–(16) can be easily generalized in a way to account for all these facts, and the corresponding expressions are straightforward if rather cumbersome. A more complicated problem is that of accounting for loop factors, since the loop factors are essentially nonlocal (they depend on the distance between two adjacent bonds) and in this case one cannot preserve a local dynamic programming algorithm similar to (15). Fortunately, though, the loop factors are of entropic nature and therefore become negligible in the low-temperature limit.

In turn, the construction of a dynamic programming algorithm for alignment of sequences that have an internal secondary structure of RNA-like type is a rather more tricky problem. Indeed, there is an energetic contribution associated with this secondary structure which survives even in the low-temperature limit. Nevertheless it is still possible to develop a nonlocal matching algorithm for this case. This is exactly the problem we address in the forthcoming sections of this paper.

2.3. Matching versus pairing of two random RNA-type heteropolymers

Here, we generalize the theory of heteropolymer binding developed above for the case of RNA molecules, whose monomers, apart from forming inter-chain bonds, are capable also of creating intra-chain links. As mentioned in the introduction, we assume that the structures formed by thermoreversible bonds of each chain are always of a cactus-like type, as shown in figure 2(*a*). It means that we restrict ourselves to the situation in which the chain conformations with 'pseudoknots' shown in figure 2(*b*) are prohibited. The difference between allowed and not allowed structures becomes more transparent, being redrawn in the following way. Represent a polymer under consideration as a straight line with active monomers situated along it in the natural order and depict bonds by dashed arcs connecting the corresponding

Figure 3. Diagrammatic form of the Dyson-type equation (19) for the partition function of an individual chain g_n with cactus-like topology.

monomers. Now, the absence of pseudoknots means the absence of intersection of arcs—see figures 2(*c*) and (*d*).

Similarly to the previous section, we neglect for simplicity the cooperativity effect and the fact that different pairs of matching nucleotides have different matching energies. These assumptions are known to be false for real RNA molecules (see [36] for the theory of secondary structure formation in RNA-like homopolymers where these effects are taken into account). However, as we have mentioned above, our model allows a fairly straightforward generalization to account for all these factors.

Following [37] we write the partition function $G_{m,n}$ of a complex of two heteropolymers capable of forming a cactus-like structure (compare equation (9)):

$$\begin{cases} G_{m,n} = g^{(1)}_{1,m} g^{(2)}_{1,n} + \sum\limits_{i,j=1}^{m,n} \beta_{i,j} G_{i-1,j-1} g^{(1)}_{i+1,m} g^{(2)}_{j+1,n} \\ G_{m,0} = g^{(1)}_{1,m}; \quad G_{0,n} = g^{(2)}_{1,n}; \quad G_{0,0} = 1, \end{cases} \tag{18}$$

where $g^{(1)}_{i,j}$ and $g^{(2)}_{i,j}$ are the partition functions of individual chains. They satisfy the self-consistent equation [38, 39]

$$\begin{cases} g^{(a)}_{k,n} = 1 + \sum\limits_{i=k}^{n-1} \sum\limits_{j=i+1+\ell}^{n} \beta'_{i,j} g^{(a)}_{i+1,j-1} g^{(a)}_{j+1,n}; \\ g^{(a)}_{0} = 1, \quad a = 1, 2. \end{cases} \tag{19}$$

This equation (the corresponding diagrammatic representation is shown in figure 3) generates the secondary structures of RNA-like (cactus) topology and it has frequently appeared in the RNA context (see, for example, [25, 36, 37, 40]). Here, $g^{(a)}_{i,j}$ is the statistical weight of the loop from the nucleotide i till the nucleotide j in the first ($a = 1$) or second ($a = 2$) sequence. The Boltzmann weights $\beta'_{i,j}$ are the constants of self-association, which are, similarly to $\beta_{m,n}$, encoded by the contact map. The summation over j runs from $i + 1 + \ell$ till n ensuring the absence of loops of lengths smaller than ℓ monomers; in what follows we mostly consider $\ell = 3$. Note also that since in this paper we are interested in the low-temperature behavior of the partition function, we neglect here the aforementioned 'loop weights', i.e. entropic factors due to the formation of intra-chain loops.

Equations (18) and (19) constitute the analytical basis of our numerical studies, for the problem of RNA-like matching (i.e. matching of sequences with RNA-type architecture). These equations replace the dynamic programming algorithm (7)–(8) valid for linear sequences.

3. Matching algorithm for two noncoding RNAs

In this section, we describe an algorithm for computing the binding free energy (which plays a role of the cost function) for a pair of two noncoding RNAs. If the cactus-like secondary

AUCGAUGUAGGGUCACCGGGCUUAUGUUACGACGAGAUGUCUUGUUCGAUCAUGCGCUUCCGCGGAGAGUGGAAA - s1
AGUUGCACCGCCAGACUACUUAACUAAACGUCGGCCAAGACAAUUCGCAUCGACCUAGUUAGCACGCACCAUCGA - s2

Figure 4. Two trial sequences of $m = n = 75$ nucleotides.

structure within the RNAs constituting the complex is not allowed (this case corresponds to inserting $\beta'_{i,j} = 0$ into (19)), the extrapolation of the free energy to zero temperature leads to the well-known standard dynamic programming algorithm described in (7) and (8). For brevity, we call this case 'linear'. When cactus-like structures exist (we call this case 'RNA-type' matching) our algorithm is not reduced (even at zero temperature) to any local recursive scheme.

For clarity we formulate the sequential steps of our algorithm using a specific example of two trial sequences of nucleotides of lengths m and n with $m = n = 75$ taken from [41]. These sequences are depicted in figure 4. We are going to show how these sequences are aligned according to both 'linear' and 'RNA-like' algorithms. The corresponding binding free energies (cost functions) of these alignments

$$F_{m,n} = T \ln G_{m,n}$$

are calculated.

3.1. Linear matching

To compute the cost function for the 'linear' alignment, proceed as follows. Construct the matrix G whose elements $G_{i,j}$ $(1 \leqslant i \leqslant m; 1 \leqslant j \leqslant n)$ are the partition functions satisfying the relation (10)–(11) with the boundary conditions $G_{m,0} = G_{0,n} = G_{0,0} = 1$ (see (9)). The matrix element $G_{i,j}$ defines matching of i first nucleotides of the first sequence with j first nucleotides of the second one. Assume that the effective energy of two complimentary nucleotides in (10) is $u = 1$, while for noncomplimentary ones we take $v = 0$. It is clear from (11) that the search of $G_{m,n}$ can be completed in polynomial time $\sim O(mn)$. At $T \to 0$ we recover the standard dynamic programming algorithm [26, 27] (see (7)).

3.2. RNA-type matching

Suppose now that both sequences in figure 4 can form hierarchical cactus-like (i.e. 'RNA-type') structures. At finite temperature the computation of the free energy of the complex built by the pair of RNA-type sequences can be accomplished as follows.

Find the matrices $g^{(1)}$ and $g^{(2)}$ (of sizes $m \times m$ and $n \times n$) of statistical weights of first and second sequences separately. To do that note that on the rhs of (19) the difference in lower indices of corresponding $g^{(a)}$s is always smaller than on the lhs. This allows us to solve the system of equations (19) recursively, starting with obvious boundary conditions $g^{(a)}_{i,i} = 1$ for any i. Indeed, from (19) equipped by defined boundary conditions one immediately finds the elements $g^{(a)}_{i,i+1}$; on the next step, one calculates $g^{(a)}_{i,i+2}$. For definiteness, one can set $g^{(a)}_{i,j} = 0$ for all $i > j$. Now, knowing the matrices $g^{(1)}$ and $g^{(2)}$ find the elements $G_{i,j}$ of the matrix G by solving (18). Obviously, calculation of each $G_{i,j}$ takes not more than $O(m \times n)$ time steps. Therefore, the whole matrix G can be determined in time $\sim O(m^2 \times n^2)$.

Now, the ground state free energy $F_{m,n}$ (i.e. the binding free energy at zero temperature) for RNA-like structures can be explicitly computed by extending the approach developed in

section 2.3. Indeed, taking a zero-temperature limit in (18) (compare to equations (15)–(16)) we get

$$F_{m,n} = \max_{\substack{i=1,\dots,m \\ j=1,\dots,n}} \left[f^{(1)}_{1,m} + f^{(2)}_{1,n}, Q^{m,n}_{i,j} \right], \tag{20}$$

where $f^{(a)}_{i,j} = \lim_{T\to 0} T \ln g^{(a)}_{i,j}$ $(a = 1, 2)$ are the free energies of individual subsequences from the nucleotide i till the nucleotide j, and $Q^{m,n}_{i,j}$ is the zero-temperature limit of the (i, j)th term in equation (18):

$$Q^{m,n}_{i,j} = F_{i-1,j-1} + f^{(1)}_{i+1,m} + f^{(2)}_{j+1,n} + \tilde{\eta}_{i,j}. \tag{21}$$

Clearly, $Q^{m,n}_{i,j}$ has a meaning of a ground state free energy of a complex which is forced to have a bond in position (i, j). In turn, the ground state energy of a single chain satisfies the following equation:

$$f^{(a)}_{i,j} = \max_{\substack{r=1,\dots,i \\ s=i+1+\ell,\dots,j}} \left[f^{(a)}_{r+1,s-1} + f^{(a)}_{s+1,j} + \tilde{\eta}'^{(a)}_{r,s} \right]. \tag{22}$$

Here the values $\tilde{\eta}_{i,j}$ are the inter-sequence matching constants (compare to (16)), while $\tilde{\eta}'^{(a)}_{i,j}$ are the intra-sequence matching constants (compare to $\beta'_{i,j}$ in (19)).

The boundary conditions for the ground state free energy follow from the boundary conditions of the partition function (18):

$$\begin{cases} F_{0,0} = 0; \\ F_{i,0} = f^{(1)}_{1,i}; \quad 1 \leqslant i \leqslant m \\ F_{0,j} = f^{(2)}_{1,j}; \quad 1 \leqslant j \leqslant n. \end{cases} \tag{23}$$

Thus, to compute the ground state free energy of the complex of two RNA-like sequences, we should first reconstruct the matrices $f^{(1)}$ and $f^{(2)}$ for individual chains by applying equation (22) and then find the matrix F using equation (20). The boundary conditions (23) together with equation (21) allow us to compute the elements of the matrices $Q^{1,j}$ for $m = 1$ and any $1 \leqslant j \leqslant n$. Knowing the corresponding matrix $Q^{1,j}$ we define the elements $F_{1,j}$ $(1 \leqslant j \leqslant n)$ of the free energy matrix by using equation (20). Then we proceed recursively and determine the matrices $Q^{2,j}$, compute $F_{2,j}$ $(1 \leqslant j \leqslant n)$, etc. Clearly, this algorithm can be completed in time of order $O(m^2 \times n^2)$.

3.3. Statistical analysis of a pair of random sequences

We have performed the statistical analysis of the ground state energy for pairs of random sequences with linear and RNA-type matching. For simplicity, we considered the chains of same length n. It has been shown in [30] that for *linear matching* the ground state free energy in the so-called Bernoulli matching approximation has the following behavior at $n \gg 1$:

$$\begin{aligned} &\langle F \rangle \approx \frac{2}{1+\sqrt{c}} n + f(c)\langle \chi \rangle n^{1/3} \\ &\sigma \equiv \sqrt{\langle F^2 \rangle - \langle F \rangle^2} \approx \sqrt{\langle \chi^2 \rangle - \langle \chi \rangle^2} f(c) n^{1/3}, \end{aligned} \tag{24}$$

where $f(c) = \frac{c^{1/6}(\sqrt{c}-1)^{1/3}}{\sqrt{c}+1}$ (see [30] for details), c is the number of different letters in the sequence (in our case it is just the number of nucleotides, $c = 4$) and χ is some random variable with known n-independent distribution (the so-called Tracy–Widom distribution, $\langle \chi \rangle = -1.7711\dots$ and $\langle \chi^2 \rangle - \langle \chi \rangle^2 = 0.8132\dots$).

The corresponding numerical results are presented in figure 5. The slope $k_1 \approx 0.65$ in figure 13(*a*) is in very good agreement with the value $k_1 = \lim_{n\to\infty} \frac{\langle F \rangle}{n} \to \frac{2}{3}$ computed from the

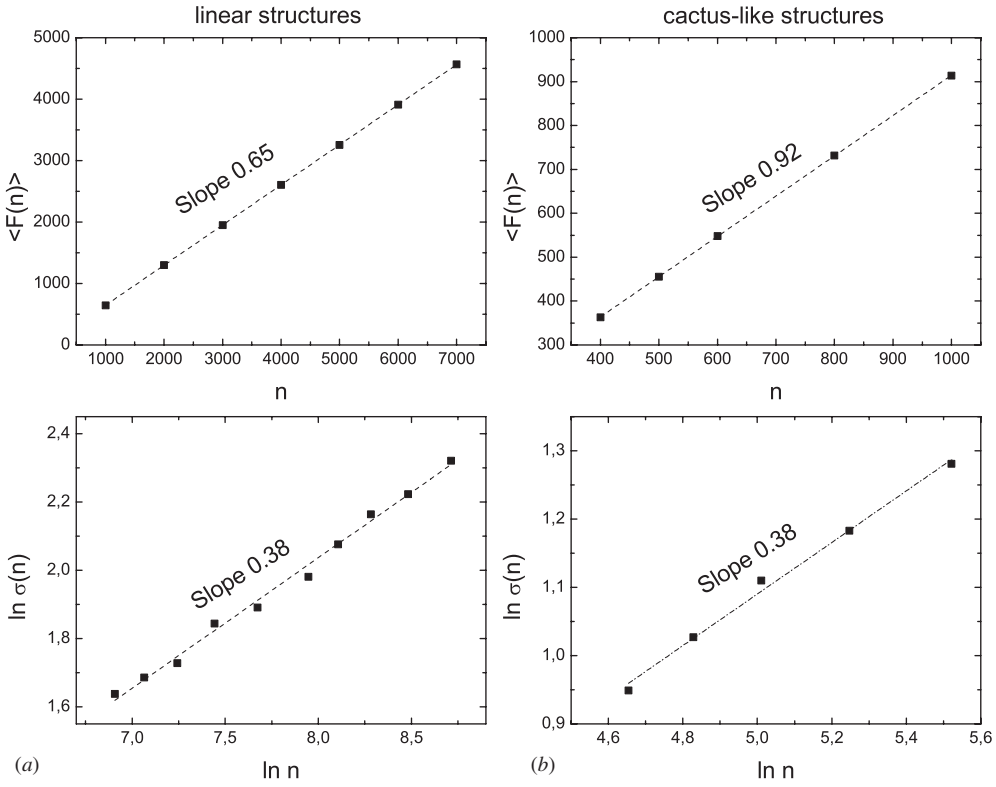

Figure 5. Plots of the ground state energy, $\langle F(n, T = 0)\rangle$ (linear scale), and its fluctuations, $\sigma(n)$ (double logarithmic scale) for (*a*) linear matching and (*b*) RNA-like matching.

first of equations (24), while the slope 0.38 in figure 13(*b*) is close to the exponent $\frac{1}{3}$ in the second line of (24). The averaging has been performed over 200 different randomly chosen structures with a uniform distribution of $c = 4$ nucleotides.

The similar analysis have been done for the same trial sequences but with RNA-type matching rules with $\ell = 0$. The plots of $\langle F(n)\rangle$ and $\sigma(n)$ are shown in figures 5(*c*). One sees that again, similarly to the linear matching, $\langle F(n)\rangle = k_c n$ for large n, but the coefficient $k_c \approx 0.92$ is larger than k_l. This signals the large number of pairs in the ground state, leading to the loop creation. The slope in figure 5(*d*) allows one to conclude that the loop creation does not affect the universality class of fluctuations and it remains the same as for linear sequences. The details of this statistical analysis will be published separately [42].

4. Secondary structure recovery

In this section, we show how the algorithm used for the computation of the ground state energy of the RNA-like complexes can help to recover the details of the ground state secondary structure. We start with recalling the corresponding procedure for linear matching [28, 29, 33], and then we pass on to the more complicated RNA-like case.

S2 →

S1 ↓

$\tilde{\eta} =$

	C	G	C	A	A	G	C	U	C
U	0	0	0	1	1	0	0	0	0
C	0	1	0	0	0	1	0	0	0
G	1	0	1	0	0	0	1	0	1
A	0	0	0	0	0	0	0	1	0
A	0	0	0	0	0	0	0	1	0
A	0	0	0	0	0	0	0	1	0
U	0	0	0	1	1	0	0	0	0
G	1	0	1	0	0	0	1	0	1
G	1	0	1	0	0	0	1	0	1

(*a*)

$F =$

	C	G	C	A	A	G	C	U	C
U	0	0	0	1	1	1	1	1	1
C	0	1	1	1	1	2	2	2	2
G	1	1	2	2	2	2	3	3	3
A	1	1	2	2	2	2	3	4	4
A	1	1	2	2	2	2	3	4	4
A	1	1	2	2	2	2	3	4	4
U	1	1	2	3	3	3	3	4	4
G	1	1	2	3	3	3	4	4	5
G	1	1	2	3	3	3	4	4	5

(*b*)

Figure 6. (*a*) Incidence matrix $\tilde{\eta}$ and (*b*) ground state free energy matrix F.

4.1. Finding the longest common subsequence for linear chains

Sequence matching problem for linear structures consists in finding the longest common subsequence of two given sequences of nucleotides. In other words, we are interested not just in the *length* of the LCS (which is provided by the dynamic programming algorithm (7), (8)) but in the complete set of matched nucleotides. Note that a degeneracy in the ground state can exist, meaning that the 'best' common subsequence is not unique. Correspondingly, the algorithm of LCS recovery described below may have 'branching points', and following different branches one can recover different best subsequences.

Let us demonstrate the recovery algorithm on a simple example. Take two sequences of nucleotides, say

$$\begin{array}{ll} \mathsf{U\,C\,G\,A\,A\,A\,U\,G\,G} & -\mathsf{S1} \\ \mathsf{U\,C\,G\,A\,A\,A\,U\,G\,G} & -\mathsf{S2} \end{array} \tag{25}$$

(in this case $m = n = 9$). Construct the incidence matrix $\tilde{\eta}$ with $\tilde{\eta}_{i,j} = 1$ if monomers i of the first sequence and j of the second one match, and $\tilde{\eta}_{i,j} = 0$ otherwise—see figure 6(*a*). Then—see figure 6(*b*)—construct the matrix of ground state free energies F using the recursion algorithm (15)–(16). The lower right element of this matrix $F_{9,9} = 5$ is the ground state free energy of the whole linear complex.

Now, in order to see which particular nucleotides do form links, one should pay attention to which particular option is realized on each step of algorithm (7). All this information is preserved in the F matrix. Indeed, take an element $F_{i,j}$ with any i, j and compare its value to the values of three neighboring matrix elements $F_{i-1,j-1}$, $F_{i-1,j}$, $F_{i,j-1}$. Then we proceed as follows.

(1) If $F_{i-1,j-1} = \max[F_{i-1,j-1}, F_{i-1,j}, F_{i,j-1}]$, then to get the desired value of $F_{i,j}$ one should link the ith nucleotide in the first sequence to the jth one in the second one.
(2) If $F_{i-1,j} = \max[F_{i-1,j-1}, F_{i-1,j}, F_{i,j-1}]$, then the ith nucleotide in the first sequence does not influence the $F_{i,j}$ and can be skipped.
(3) If $F_{i,j-1} = \max[F_{i-1,j-1}, F_{i-1,j}, F_{i,j-1}]$, then the ith nucleotide in the first sequence does not influence the $F_{i,j}$ and can be skipped.

If several of these options happen at once, one gets the aforementioned branching point which ultimately would lead to different ways of realizing the ground state.

	C	G	C	A	A	G	G	G	G
U	0	0	0	1	1	0	0	0	0
C	0	1	1	1	1	2	2	2	2
G	1	1	2	2	2	2	3	3	3
A	1	1	2	2	2	2	3	4	4
A	1	1	2	2	2	2	3	4	4
A	1	1	2	2	2	2	3	4	4
U	1	1	2	3	3	3	3	4	4
G	1	1	2	3	3	3	4	4	5
G	1	1	2	3	3	3	4	4	5

(*a*)

(*b*)

Figure 7. Structure recovery algorithm for linear chains: (*a*) free energy matrix; (*b*) recovered matching.

Starting this procedure with the element at the low right corner of F ($F_{9,9}$ in our case—see figure 7(*a*)) and repeating the algorithm recursively one gets the desired longest common subsequence shown in figure 7(*b*).

4.2. Secondary structure recovery for RNA-type matching

Structure recovery for the chains capable of forming cactus-like architecture is a much more involved problem; however, it can also be described recursively. In this case the algorithm consists of the following successive steps.

To recover the *inter-chain* contacts we proceed as follows. Construct the matrices F, $f^{(1),(2)}$ and $Q^{i,j}$ for $i = 1, \ldots, n$, $j = 1, \ldots, m$ as prescribed by the algorithm of finding the ground state free energy. Take the element $F_{m,n}$ and check if $F_{m,n} = f^{(1)}_{1,m} + f^{(2)}_{1,n}$. If yes, then the ground state corresponds to two unconnected RNA structures, and one should go immediately to the intra-chain structure recovery. If not, consider the matrix $Q^{m,n}$ (equation (21)) and choose its maximal element $Q^{m,n}_{p,q}$. This element corresponds to pairing between the nucleotide p of the first sequence and nucleotide q of the second one; this pairing should exist in the ground state secondary structure. (Once again, the existence of several maximal elements of $Q^{m,n}$ should be considered as a branching point of the algorithm.) Now, substitute $m \to (p-1)$, $n \to (q-1)$ and repeat this procedure. Proceeding recursively until $\min(p, q) \leqslant 1$, one gets all the inter-chain contacts in the primary structure.

Knowing all pairs of nucleotides linking two chains together, one should reconstruct the secondary structure of *intra-chain* loops between these connections. To do that, proceed as follows. Assume, for definiteness, that we are interested in the internal structure of the first match between ith and jth nucleotides (these nucleotides are involved in the inter-chain connections). Build the matrix $S^{i,j}$ consisting of elements $S^{i,j}_{r,s} = f^{(1)}_{r+1,s-1} + f^{(1)}_{s+1,j} + \tilde{\eta}_{r,s}$ with $i < r < s < j$. Find the largest element of this matrix; let it be, say, $S^{i,j}_{p,t}$. Then the nucleotides p and t do form a bond, and there are no more bonds between nucleotides with numbers i and p. Now repeat the same procedure substituting $p \to i$ and $t \to j$, and also substituting $t \to i$ and $j \to j$. Since on each step the difference between the starting and ending indices decreases, this procedure converges to the desired secondary structure (list of bonds).

To understand the algorithm better let us demonstrate it by example (25). The corresponding incidence matrix $\tilde{\eta}$ (for cross-sequence matching S1–S2) is shown in

$\widetilde{\eta}' =$

S1 ↓ \ S1 →	U	C	G	A	A	A	U	G	G
U	0	0	0	1	1	1	0	0	0
C	0	0	1	0	0	0	0	1	1
G	0	1	0	0	0	0	0	0	0
A	1	0	0	0	0	0	1	0	0
A	1	0	0	0	0	0	1	0	0
A	1	0	0	0	0	0	1	0	0
U	0	0	0	1	1	1	0	0	0
G	0	1	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	0	0

(*a*)

$\widetilde{\eta}'' =$

S2 ↓ \ S2 →	C	G	C	A	A	G	G	G	G
C	0	1	0	0	0	1	0	0	0
G	1	0	1	0	0	0	1	0	1
C	0	1	0	0	0	1	0	0	0
A	0	0	0	0	0	0	0	1	0
A	0	0	0	0	0	0	0	1	0
G	1	0	1	0	0	0	1	0	1
C	0	1	0	0	0	1	0	0	0
U	0	0	0	1	1	0	0	0	0
C	0	1	0	0	0	1	0	0	0

(*b*)

Figure 8. Incidence matrices for pairs of chains with possible clover-leaf structures in each sequence: (*a*) internal matching S1–S1; (*b*) internal matching S2–S2.

$F =$

	C	G	C	A	A	G	G	G	G
U	0	1	1	2	2	2	3	4	4
C	0	1	1	2	2	3	3	4	4
G	1	2	2	3	3	3	4	5	5
A	2	3	3	3	3	4	4	5	5
A	2	3	3	3	3	4	4	5	5
A	2	3	3	3	3	4	4	5	5
U	3	4	4	4	4	5	5	6	6
G	4	4	5	5	5	5	6	7	7
G	4	4	5	5	5	5	6	7	8

(*a*)

$f^{(1)} =$

	U	C	G	A	A	A	U	G	G
U	0	0	1	2	2	2	3	3	3
C	0	0	1	1	1	1	2	2	2
G	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	1	1	1
U	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0

(*b*)

$f^{(2)} =$

	C	G	C	A	A	G	C	U	C
C	0	1	1	1	1	2	2	3	3
G	0	0	1	1	1	1	2	3	3
C	0	0	0	0	0	1	1	2	2
A	0	0	0	0	0	0	1	2	2
A	0	0	0	0	0	0	1	2	2
G	0	0	0	0	0	0	1	1	1
C	0	0	0	0	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0

(*c*)

Figure 9. Algorithm description: energies corresponding to incidence matrices in figure 8: (*a*) ground-state free energy matrix; statistical weights of the first (*a*) and second (*b*) sequences.

figure 6(*a*), while the incidence matrices η' (for internal matching of the first sequence), η'' (for internal matching of the second one) are shown in figures 8(*a*) and (*b*), respectively.

The ground-state free energy matrices *F*s, as well as the matrices of effective statistical weights $f^{(1)}$ and $f^{(2)}$ of first and second sequences, are shown in figures 9(*a*)–(*c*) (for simplicity, we assume here $\ell = 0$, but note that the algorithm is insensitive to the value of ℓ).

We begin with the reconstruction of the optimal set of contacts between first and second chains (the *inter-chain* structure recovery).

Step 1. Since $F_{9,9} > f^{(1)}_{1,9} + f^{(2)}_{1,9}$ *consider the matrix Q for* $F^{9,9}$*—see figure 10(a).*[9] *The maximal element of the matrix Q is* $Q_{\max} = Q_{9,9} = 8$*, meaning that the ninth nucleotide of* S1 *forms a bond with the ninth nucleotide of* S2.

Step 2. Since $F_{8,8} > f^{(1)}_{1,8} + f^{(2)}_{1,8}$*, take the matrix Q for* $F^{8,8}$*—see figure 10(b). The maximal element of the current matrix Q is* $Q_{\max} = \{Q_{8,1}$ or $Q_{8,3}\} = 7$*, meaning that the eighth*

[9] For brevity we present only those matrices $Q^{i,j}$ which are used for the structure recovery. For each (i, j) the matrix $Q^{m,n}$ is of size $m \times n$.

Q for $F_{9,9}$

5	4	5	6	5	3	4	4	5
4	4	4	4	4	4	3	4	5
5	3	5	4	4	3	5	4	6
5	4	5	5	5	4	4	6	6
6	5	6	6	5	4	5	6	6
5	4	5	6	5	3	4	4	5
5	4	5	6	5	3	4	4	5
7	5	7	6	5	4	6	5	7
7	6	7	7	6	5	6	6	8

(*a*)

Q for $F_{8,8}$

5	4	5	6	5	3	4	4
4	4	4	4	4	4	3	4
5	3	5	4	4	3	5	4
5	4	5	5	5	4	4	6
6	5	6	6	5	4	5	6
5	4	5	5	4	3	4	5
5	4	5	6	5	3	4	4
7	5	7	6	5	4	6	5

(*b*)

Figure 10. Algorithm description for inter-chain contacts: matrices Q corresponding to (*a*) $F_{9,9}$; (*b*) $F_{8,8}$.

$f^{(1)} =$

	U	C	G	A	A	A	U
U	0	0	1	2	2	2	3
C	0	0	1	1	1	1	2
G	0	0	0	0	0	0	1
A	0	0	0	0	0	0	1
A	0	0	0	0	0	0	1
A	0	0	0	0	0	0	1
U	0	0	0	0	0	0	0

(*a*)

$S =$

	C	G	A	A	A	U
U	1	1	3	3	2	1
C	2	1	1	0	0	0
G	1	1	0	0	0	0
A	1	0	1	0	0	0
A	0	1	0	0	0	0
A	1	0	0	0	0	0

(*b*)

(*c*)

Figure 11. Algorithm description for inter-chain contacts: matrices $f^{(1)}$ (*a*) and S (*b*); the corresponding loop structure (*c*).

nucleotide of S1 *forms a bond with the first nucleotide of* S2 *or the eighth nucleotide of* S1 *forms a bond with the third nucleotide of* S2*—see the structures shown in figure 12.*

Step 3. (Only for the structure 2 in figure 12(b)). Since $F_{7,2} > f^{(1)}_{1,7} + f^{(2)}_{1,2}$ *there are no more interacting inter-chain pairs.*

To finalize we should now reconstruct the loop structures of both ground states. So we proceed with the *intra-chain* structure recovery.

The loop is reconstructed using the corresponding values of $f_{p,q}$. Consider the structure recovery on example of the loop with $f^{(1)}_{1,7}$: U C G A A A U. The matrices $f^{(1)}$ and S corresponding to the element $f^{(1)}_{1,7} = 3$ are depicted in figures 11(*a*) and (*b*). The maximal element of the matrix S for the statistical weight $f_{1,7}$ is shown in figure 11(*b*) and is $S_{\max} = S_{1,3} = S_{1,4} = 3$. If we choose the element $S_{1,3}$, the nucleotides with numbers 1 and 4 form a bond. Considering now $f^{(1)}_{2,3} = 1$ and $f^{(1)}_{5,7} = 1$ and developing the corresponding matrices S (we do not write them because of their simplicity), we arrive at the internal loop structure shown in figure 11(*c*).

The overall structures with inter- and intra-chain optimal matches are shown in figures 12(*a*) and (*b*).

We have applied this algorithm to the longer trial sequences shown in figure 4. We have performed the structure recovery for three different cases: linear matching (a), RNA-like

Figure 12. Algorithm description: recovered structures for the pair of sequences shown in (25): (*a*) and (*b*) have identical ground state free energies.

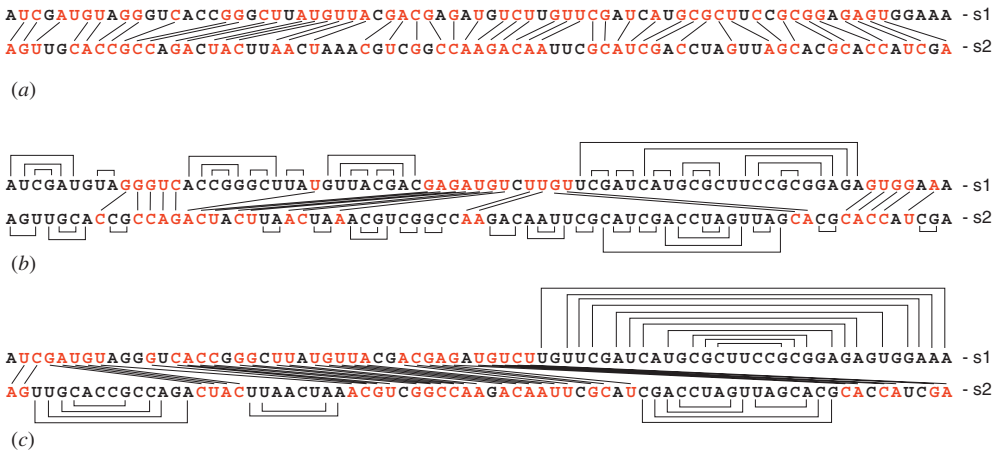

Figure 13. Structures recovered from the pair of sequences shown in figure 4: (*a*) linear structure; (*b*) branching structure with $\ell = 0$; (*c*) branching structure with $\ell = 3$.

matching with $\ell = 0$ (b), and RNA-like matching with $\ell = 3$ (c). The resulting structures are depicted in figure 13.

5. Conclusion

In this paper, we have developed and implemented a new statistical algorithm for quantitative determination of the binding free energy of two heteropolymer sequences under the supposition that each sequence can form a hierarchical cactus-like secondary structure, typical for RNA molecules.

In section 3, we offered a constructive way to build a 'cost function' characterizing the matching of two *noncoding* RNAs with arbitrary primary sequences. The substantial difference of this procedure from the convenient sequence comparison is that in the ncRNA case we not only align the sequences of nucleotides which constitute pairs between two RNAs but also take into account the secondary structure of the parts of RNA between the aligned nucleotides. Our algorithm is based on two facts: (i) the standard alignment problem can be reformulated as a zero-temperature limit of a more general statistical problem of binding two associating heteropolymer chains; (ii) the last problem can be straightforwardly generalized onto the sequences with hierarchical cactus-like structures (i.e. of RNA type). Taking the zero-temperature limit at the very end we arrive at the desired ground state free energy with the account for entropy of side cactus-like loops.

We have demonstrated in detail (see section 4) how our algorithm enables us to solve the *secondary structure recovery* problem. In particular, we can predict in zero-temperature limit the secondary structure of each ncRNA (without pseudoknots) by knowing only their primary sequences.

Let us emphasize that the structure recovered turns out to be very sensitive to the details of the model: compare figures 12(*b*) and (*c*), which differ only by the minimal allowed size of a loop. The theoretical question which still remains open concerns the quantitative description of the change in the topology of RNA pairing when ℓ is changed. Yet it is not clear whether this change in topology is phase transition or just a cooperative effect. We plan to attack this question in a separate publication [42].

Such a strong sensitivity of the secondary structure to the details of the model means that to get the experimentally verifiable secondary structures, one should plug into the model as much information about exact values of loop factors, cooperativity parameters and interaction energies as possible. As we have mentioned in the discussion at the end of section 2.2, the procedure developed in this paper can be straightforwardly generalized to allow for all these factors. The corresponding results concerning the prediction of real RNA complexes as well as comparison of our algorithm with the ones existing in the literature will be provided in the forthcoming paper [42].

Let us emphasize that in this contribution we were guided by an attempt to avoid as much as possible some heuristic 'cookings' remaining in the framework of statistical physics. So we have exploited the similarity in mathematical description of matching (alignment) problem and finding the free energy of a complex of two mutually 'adsorbed' (i.e. paired) heteropolymer chains. We believe that this similarity could lead for mutual enrichment of both dynamic programming approach to alignment of sequences and investigation of conformational properties of adsorbed heteropolymers[10].

Acknowledgments

We are very grateful to A A Mironov for opening for us the world of ncRNAs and to V A Avetisov for numerous encouraging discussions concerning the biophysical and statistical aspects of the problem. This work has been supported partially by the grant ERARSysBio+ #66; MVT and OVV acknowledge the warm hospitality of LPTMS where this work has been completed.

References

[1] Eddy S R 2002 *Cell* **109** 137
[2] Ambros V 2001 *Cell* **107** 862
[3] Storz G 2002 *Science* **296** 1260
[4] Navarro P, Pichard S, Ciaudo C, Avner P and Rougeulle C 2005 *Genes Dev.* **19** 1474
[5] Pande V, Grosberg A and Tanaka T 2000 *Rev. Mod. Phys.* **72** 259
[6] Rivas E and Eddy S R 1999 *J. Mol. Biol.* **285** 205
[7] Needleman S B and Wunsch C D 1970 *J. Mol. Biol.* **48** 443
[8] Smith T F and Waterman M S 1981 *J. Mol. Biol.* **147** 195
Smith T F and Waterman M S 1981 *Adv. Appl. Math.* **2** 482
[9] Waterman M S, Gordon L and Arratia R 1987 *Proc. Natl Acad. Sci. USA* **84** 1239
[10] Altschul S F *et al* 1990 *J. Mol. Biol.* **215** 403
[11] Sankoff D and Kruskal J 1983 *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (Reading, MA: Addison-Wesley)

[10] In particular, in [42] we are going to imply this similarity to address the question of gaps' statistics in a pair of aligned RNA-type chains versus the gaps' in the chains with linear matching.

[12] Apostolico A and Guerra C 1987 *Alogorithmica* **2** 315
[13] Wagner R and Fisher M 1974 *J. Assoc. Comput. Mach.* **21** 168
[14] Gusfield D 1997 *Algorithms on Strings, Trees, and Sequences* (Cambridge: Cambridge University Press)
[15] Boutet de Monvel J 1999 *Eur. Phys. J.* B **7** 293
Boutet de Monvel J 2000 *Phys. Rev.* E **62** 204
[16] Chvátal V and Sankoff D 1975 *J. Appl. Probab.* **12** 306
[17] Deken J 1979 *Discrete Math.* **26** 17
[18] Steele J M 1982 *SIAM J. Appl. Math.* **42** 731
[19] Dancik V and Paterson M 1994 *STACS94* (*Lecture Notes in Computer Science* vol 775) (New York: Springer) p 306
[20] Alexander K S 1994 *Ann. Appl. Probab.* **4** 1074
[21] Kiwi M, Loebl M and Matousek J 2004 *Lecture Notes in Computer Science* vol 2976 (Berlin: Springer) p 302
[22] Zhang M and Marr T 1995 *J. Theor. Biol.* **174** 119
[23] Hwa T and Lassig M 1996 *Phys. Rev. Lett.* **76** 2591
[24] Bundschuh R 2001 *Eur. Phys. J.* B **22** 533
[25] Bundschuh R and Hwa T 1999 *Phys. Rev. Lett.* **83** 1479
[26] Waterman M S 1984 *Bull. Math. Biol.* **46** 473
[27] Waterman M S and Vingron M 1994 *Stat. Sci.* **9** 387
[28] Bundschuh R and Hwa T 2000 *Discrete Appl. Math.* **104** 113
[29] Drasdo D, Hwa T and Lassig M 2000 *J. Comput. Biol.* **7** 115
[30] Majumdar S N and Nechaev S 2004 *Phys. Rev.* E **69** 011103
[31] Tracy C A and Widom H 1994 *Commun. Math. Phys.* **159** 151
see also Tracy C A and Widom H 2002 *Proc. ICM (Beijing)* vol 1 p 587
[32] Majumdar S N 2007 Les Houches Lecture Notes on 'Complex Systems' arXiv:cond-mat/0701193
[33] Tamm M V and Nechaev S K 2008 *Phys. Rev.* E **78** 011903
[34] Comtet L 1974 *Advanced Combinatorics: The Art of Finite and Infinite Expansions* (Dordrecht: Reidel)
[35] Grosberg A and Khokhlov A 1994 *Statistical Physics of Macromolecules* (New York: AIP) chapter 7
[36] Müller M 2003 *Phys. Rev.* E **67** 021914
[37] Tamm M V and Nechaev S K 2007 *Phys. Rev.* E **75** 031904
[38] de Gennes P 1968 *Biopolymers* **6** 715
[39] Erukhimovich I Ya 1978 *Vysokomol. Soedin.* B **20** 10 (in Russian)
[40] Müller M, Krzakala F and Mezard M 2003 *Eur. Phys. J.* E **9** 67
[41] Morgan S R and Higgs P G 1998 *J. Phys. A: Math. Gen.* **31** 3153–70
[42] Nechaev S, Valba O and Tamm M in preparation