

Research

Open Access

Sequence-structure relations of pseudoknot RNA

Fenix WD Huang, Linda YM Li and Christian M Reidys*

Address: Center for Combinatorics, LPMC-TJKLC, Nankai University, Tianjin 300071, PR China

Email: Fenix WD Huang - fenixprotoss@163.com; Linda YM Li - liyanmei@cfc.nankai.edu.cn; Christian M Reidys* - reidys@nankai.edu.cn

* Corresponding author

from The Seventh Asia Pacific Bioinformatics Conference (APBC 2009)
Beijing, China. 13–16 January 2009

Published: 30 January 2009

BMC Bioinformatics 2009, 10(Suppl 1):S39 doi:10.1186/1471-2105-10-S1-S39

This article is available from: <http://www.biomedcentral.com/1471-2105/10/S1/S39>

© 2009 Huang et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The analysis of sequence-structure relations of RNA is based on a specific notion and folding of RNA structure. The notion of coarse grained structure employed here is that of canonical RNA pseudoknot contact-structures with at most two mutually crossing bonds (3-noncrossing). These structures are folded by a novel, *ab initio* prediction algorithm, cross, capable of searching all 3-noncrossing RNA structures. The algorithm outputs the minimum free energy structure.

Results: After giving some background on RNA pseudoknot structures and providing an outline of the folding algorithm being employed, we present in this paper various, statistical results on the mapping from RNA sequences into 3-noncrossing RNA pseudoknot structures. We study properties, like the fraction of pseudoknot structures, the dominant pseudoknot-shapes, neutral walks, neutral neighbors and local connectivity. We then put our results into context of molecular evolution of RNA.

Conclusion: Our results imply that, in analogy to RNA secondary structures, 3-noncrossing pseudoknot RNA represents a molecular phenotype that is well suited for molecular and in particular neutral evolution. We can conclude that extended, percolating neutral networks of pseudoknot RNA exist.

Background

Three decades ago, Michael Waterman pioneered the combinatorics and *ab initio* prediction of the at that time rather exotic ribonucleic acid (RNA) secondary structures [1-5]. The motivation for this work was coming from a fundamental dichotomy represented by RNA. On one hand RNA is described by its primary sequence, a linear string composed of the nucleotides A, G, U and C. The primary sequence embodies the genotypic legislative. On the

other hand, RNA, being less structurally constrained than its chemical relative DNA, does fold into 3D-structures, representing the phenotypic executive. Therefore one molecule stands for both: geno- and phenotype.

Indeed, a vast variety of RNA activities was found: the discovery of catalytic RNAs, or ribozymes, in 1981 proved that RNA could catalyze reactions just as proteins. RNA can act also as a messenger between DNA and protein in

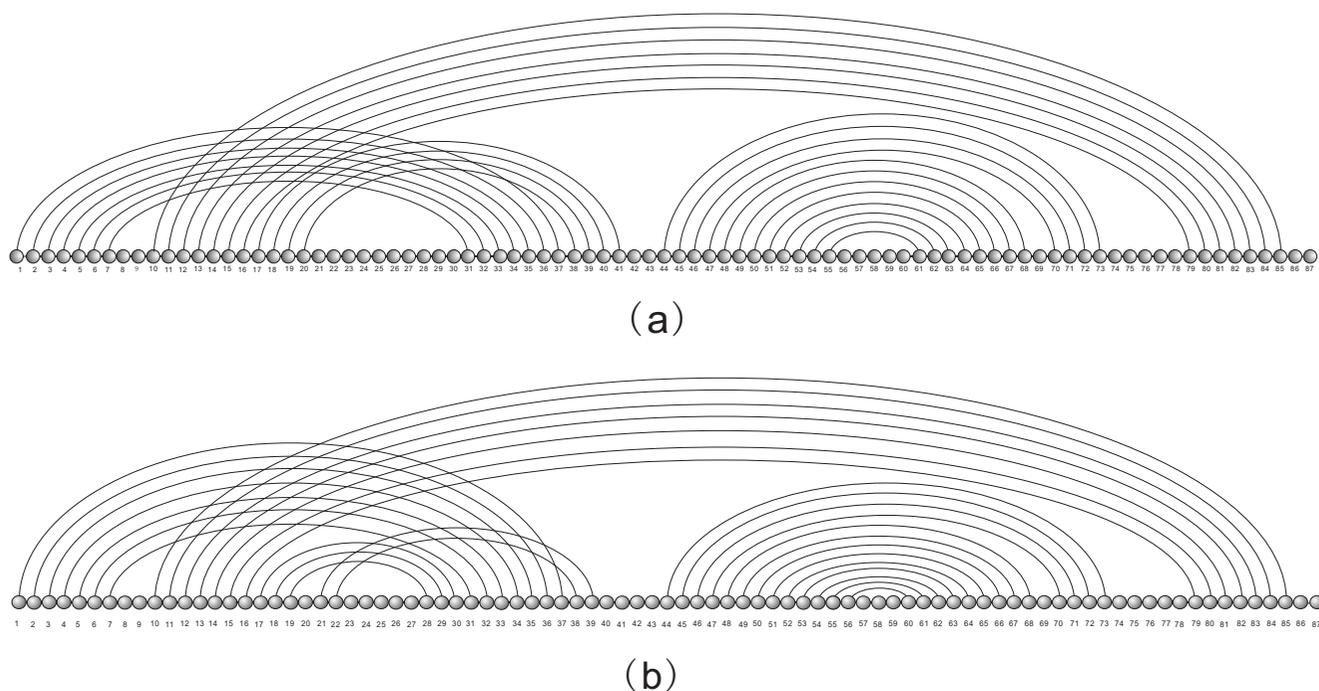


Figure 2
HDV structure. (a) Diagram representation of Hepatitis Delta Virus structure folded by our algorithm. (b) Diagram representation of natural Hepatitis Delta Virus.

pairs, respectively. A diagram is k -noncrossing if it contains at most $k - 1$ mutually crossing arcs. Diagrams have the following three key parameters: the maximum number of mutually crossing arcs, $k - 1$, the minimum arc-length, λ , and minimum stack-length, τ . The length of an arc (i, j) is $j - i$ and a stack of length τ is a sequence of "parallel" arcs of the form

$$((i, j), (i + 1, j - 1), \dots, (i + (\tau - 1), j - (\tau - 1))),$$

see Figure 3. We call an arc of length λ a λ -arc. Biophysical constraints on the base pairings imply that in all RNA

structures λ is greater than or equal to four. We call diagrams with a minimum stack-length τ , τ -canonical and if $\lambda \geq 4$ we refer to diagrams as structures. To reiterate, in the simplest case we have 2-noncrossing RNA structures, i.e. the secondary structures in which no two arcs cross, see Figure 4. The noncrossing of arcs has far-reaching consequences. It implies that RNA secondary structures form a context free language and allow for the DP algorithms [20], predicting the loop-based mfe-secondary structure in $O(n^3)$ -time and $O(n^2)$ -space.

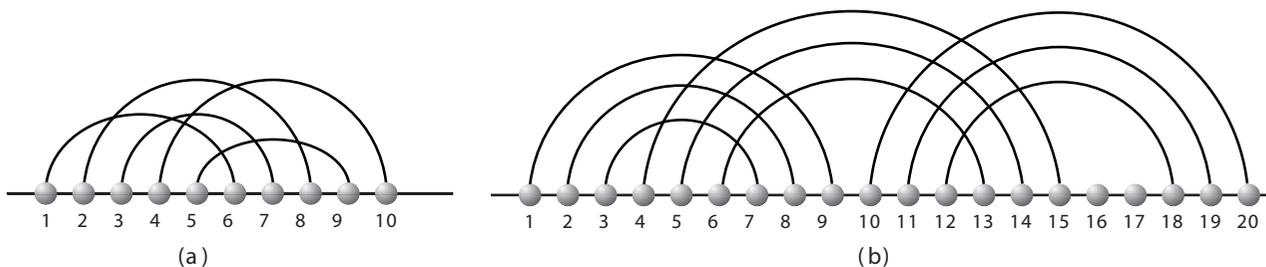


Figure 3
 k -noncrossing diagrams. We display a 4-noncrossing diagram with arc-length $\lambda \geq 4$ and stack-length $\tau \geq 1$ (a) and a 3-noncrossing, $\lambda \geq 4$ and $\tau \geq 3$ diagram (b).

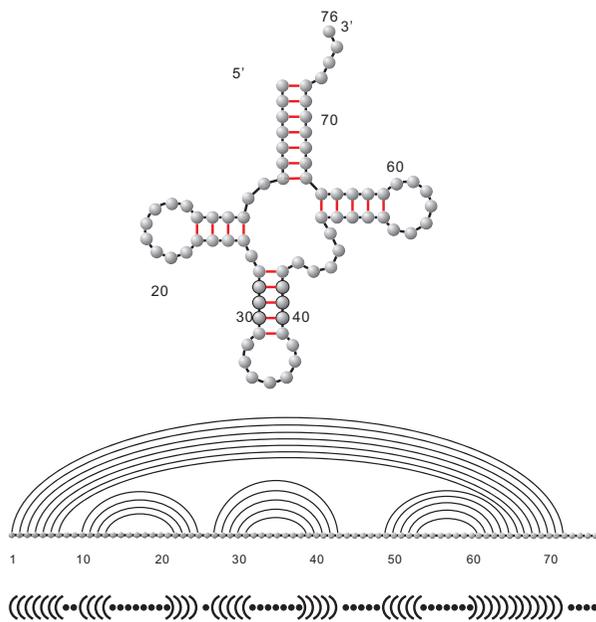


Figure 4
RNA secondary structures: Here we give three representations of the phenylalanine secondary (2-noncrossing) structure. First the outer-planar graph representation (top), second the diagram representation (middle) and finally the '()' and ')' representation.

Let us now, having some background on RNA structures return to the RNA-world. Around 1990 Peter Schuster and his coworkers initiated a paradigm shift. They began to study evolutionary optimization and neutral evolution of RNA via the relation between RNA genotypes and phenotypes. The particular mapping from RNA sequences into RNA secondary structures was obtained by the algorithm *ViennaRNA* [21], an implementation of the folding routine [6,22], mentioned above. Two particularly prominent results of this line of work were the existence of neutral networks, i.e. vast, extended networks, composed of sequences folding into a given secondary structure [23] and the Intersection Theorem [23]. The latter guarantees for any two secondary structures the existence of at least one sequence which simultaneously satisfies all constraints imposed by their Watson-Crick and G-U base pairs. For the implication of the latter with respect to molecular switches, see [24]. It became evident that the "statistical" properties of this mapping played a central role in the molecular evolution of RNA.

But, there is more. Two discoveries suggested that RNA might not just be a stepping stone towards a DNA/protein world. They show that RNA plays an active role in vital cell processes. A large number of very small RNAs of about 22 nucleotides in length, called microRNAs (miRNAs), were

discovered. They were found in organisms as diverse as the worm *Caenorhabditis* organs and humans, and their particular relationship to certain intermediates in RNA interference (RNAi). These findings have put RNA in particular noncoding RNA into the spotlight. In addition, RNA's conformational versatility and catalytic abilities have been identified in the context of protein synthesis and RNA splicing. More and more parallels between RNA and protein are currently being revealed [25].

Let us next briefly overview what we know about the combinatorics of our phenotypes, ultimately allowing for the computation of biophysically relevant pseudoknot structures [26]. The key result comes from a seemingly unrelated field, the combinatorics of partitions. Chen *et al.* proved in a seminal paper [27] a bijection between walks in Weyl chambers and k -noncrossing partitions. This bijection has recently been generalized to tangled diagrams [28]. Now, a k -noncrossing diagram is a special type of k -noncrossing tangle and the relevance of Chen's result lies in the fact that the walks in question can be enumerated via the reflection principle. In fact, the reflection principle facilitated the computation of the generating function of k -noncrossing canonical pseudoknot RNA [19,26,29]. Subsequent singularity analysis [26,29], showed, that the exponential growth rates of canonical pseudoknot RNA structures are surprisingly small, see Table 1, [26]. For instance, the number of 3-noncrossing, 3-canonical RNA structures with arc-length greater than or equal to four is asymptotically given by

$$cn^{-5} 2.0348^n,$$

where c is some (explicitly known) constant. This exponential growth rate is very close to Schuster *et al.*'s finding [30] for 2-canonical RNA secondary structures with arc-length greater than or equal to four

$$1.4848 n^{-3/2} 1.8444^n. \tag{1}$$

For the analysis presented here, we use the algorithm cross [28], which produces a transparent output. This algorithm

Table 1: Exponential growth rates of (k, τ) -structures. We have k -noncrossing structures with minimum stack-length greater than or equal to three.

k	3	4	5	6	7	8	9
$\tau = 3$	2.0348	2.2644	2.4432	2.5932	2.7243	2.8414	2.9480
$\tau = 4$	1.7898	1.9370	2.0488	2.1407	2.2198	2.2896	2.3523
$\tau = 5$	1.6465	1.7532	1.8330	1.8979	1.9532	2.0016	2.0449
$\tau = 6$	1.5515	1.6345	1.6960	1.7457	1.7877	1.8243	1.8569
$\tau = 7$	1.4834	1.5510	1.6008	1.6408	1.6745	1.7038	1.7297
$\tau = 8$	1.4319	1.4888	1.5305	1.5639	1.5919	1.6162	1.6376
$\tau = 9$	1.3915	1.4405	1.4763	1.5049	1.5288	1.5494	1.5677

does not follow the DP paradigm and generates the mfe- k -noncrossing τ -canonical structure via a combination of branch and bound, as well as DP techniques. cross inductively constructs k -noncrossing, τ -canonical RNA structures via motifs. Currently full loop-based energy models are derived and implemented for $k = 3$ and $\tau \geq 3$.

Therefore, cross finds the mfe-RNA pseudoknot structure in which there are at most two *mutually* crossing arcs, which has minimum arc-length four and in which each stack has size at least three. While cross is an exponential time algorithm it allows to fold sequences of length 100 with an average folding time of 4.5 minutes.

Methods

While it is beyond the scope of this paper to present the algorithm cross in detail, the objective of this section is first to sketch its key organization and second to discuss some basic properties of RNA pseudoknot structures. These combinatorial properties enable us to assign a unique, loop-based energy. In the course of our analysis we show that an RNA pseudoknot structure can be con-

structed via simpler substructures. These serve as the building blocks via which cross derives the mfe-pseudoknot structure. At present time we do not have an algorithm computing the partition function version of cross. For RNA secondary structures, the partition function was obtained 1990 [31], three decades after the first mfe-folding algorithms were derived [32-34]. The partition function is based on a fixed sequence and contains vital statistical information on the probabilities of specific structural configurations of the latter. For any inductively constructed structure class, it allows to compute the base pairing probabilities. In analogy to similar studies in the case of RNA secondary structures [17,35-37,37-45], the partition function is for the type of analysis presented here not of key importance. We shall derive statistical information on the sequence-structure relation by mfe-folding a large number of sequences instead of considering the ensemble of structural configurations of a *single* sequence.

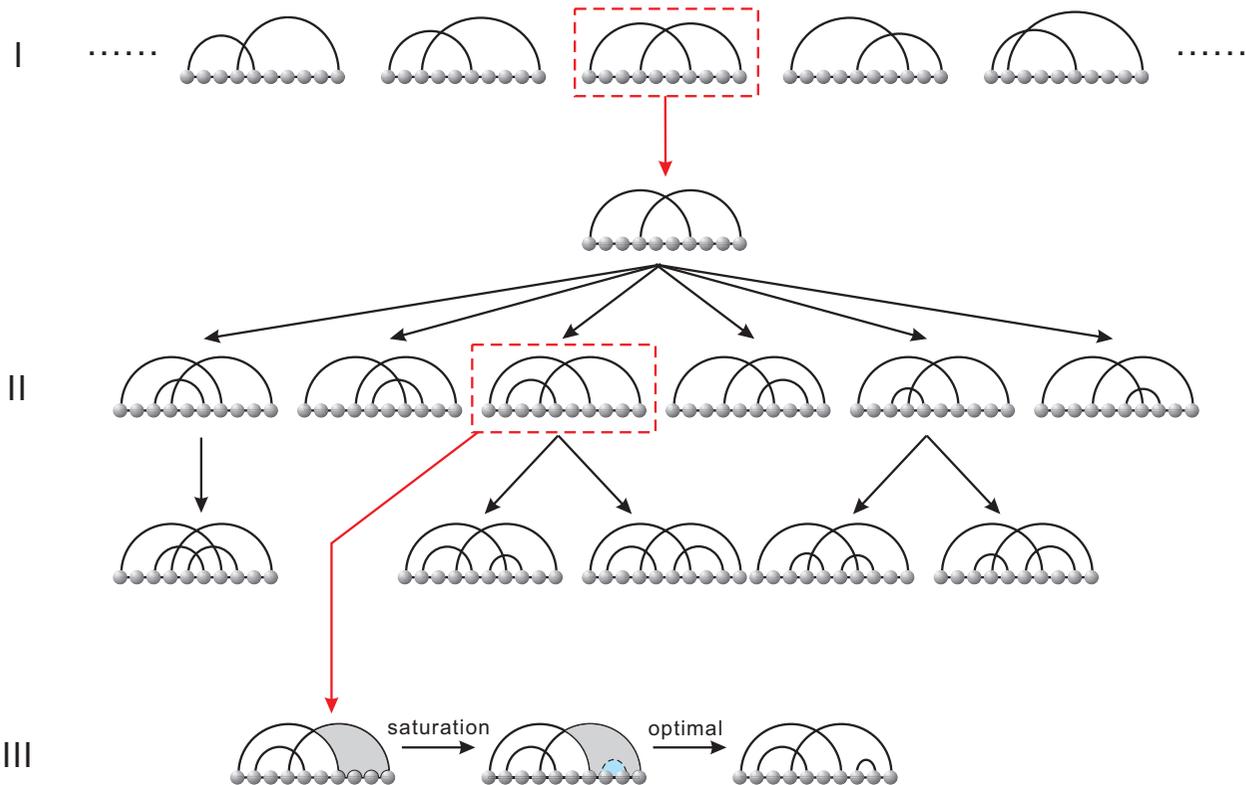


Figure 5
An outline of cross. The figure visualizes the three main phases of cross: the generation of motifs (I), the construction of skeleta-trees, rooted in irreducible shadows (II) and the saturation (III), during which, via DP-routines, optimal fillings of skeleta-intervals are derived.

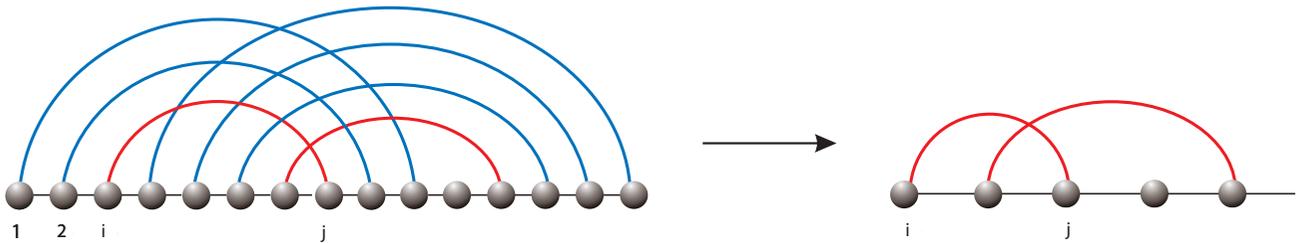


Figure 6
Cores will in general have 2-arcs. The structure δ (lhs) is mapped into its core $c(\delta)$ (rhs). Clearly d has arc-length ≥ 4 and as a consequence of the collapse of the stack $((1, j + 2), (2, j + 1), (i, j))$ into the arc (i, j) , $c(\delta)$ contains the arc (i, j) , which is, after relabeling, a 2-arc.

Cross

The algorithm cross has three distinct phases: the motif-, skeleton- and saturation-phase, see Figure 5 for an overview. We will here briefly discuss these three parts.

Let $<$ denote the following partial order over arcs

$$(i_1, j_1) < (i_2, j_2) \Leftrightarrow i_2 < i_1 \wedge j_1 < j_2, \tag{2}$$

i.e. an arc α_1 is smaller than α_2 if it is nested in it.

I Motifs

Let us begin by defining core-structures. A k -noncrossing core [29] is a k -noncrossing diagram in which all stacks have size one. The core of a structure is obtained by identifying all its stacks by single arcs, keeping the unpaired nucleotides and finally relabeling, see Figure 6.

A $\langle k, \tau \rangle$ -motif is a $\langle k, \tau \rangle$ -diagram over $[n]$, having the following properties

(M1) it has a nonnesting core

(M2) all its arcs are contained in stacks of length exactly $\tau = 3$ and length $\lambda = 4$.

A m -shadow is a k -noncrossing diagram obtained by successively increasing the stacks of m from top to bottom, see Figure 7.

The key observation about motifs is that they can, despite the fact that they exhibit cross-serial dependencies, be generated inductively [46].

II Skeleta

Skeleta represent the non-inductive "frames" of pseudo-knot RNA, i.e. skeleta entail exactly the cross-serial dependencies, that need to be considered exhaustively. A skeleton, S , is a 3-noncrossing structure, whose core has a connected L-graph. An L-graph is a diagram whose arcs are the vertices and two being adjacent if their corresponding arcs cross [46]. An irreducible shadow, $IS_{i,j}$ over $[i, j]$. $IS_{i,j}$ is a skeleton which has no nested arcs, see Figure 7. Phase II consists in the generation of all skeleta-trees, which are rooted in irreducible shadows.

III Saturation

Given a skeleton, cross saturates or "fills" via context-sensitive DP routines the skeleton-intervals. Note that, while the inserted substructures cannot cross any arc of the skeleton, they will in general contain crossing arcs within themselves.

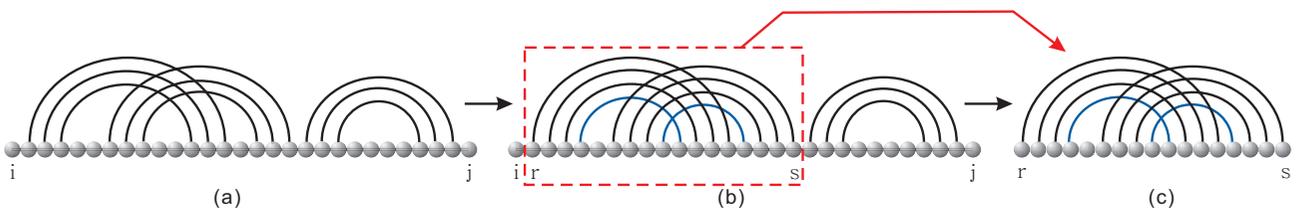


Figure 7
Motifs, shadows and irreducible shadows. We display a $\langle 3, 3 \rangle$ -motif (a) and one of its induced shadows (b), Furthermore we show an irreducible shadow (c), which corresponds to the (b)-substructure contained in the dashed box.

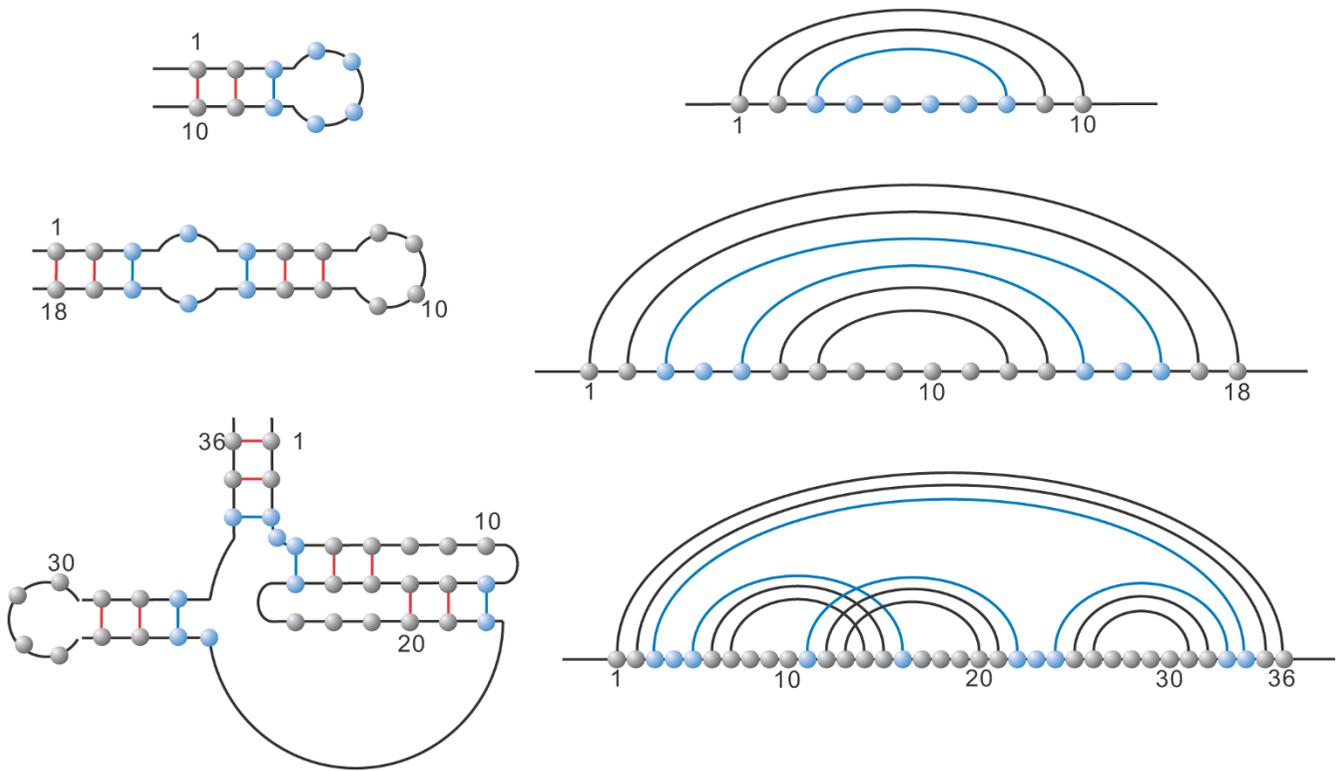


Figure 8
The standard loop-types hairpin-loop (top), interior-loop (middle), multi-loop (lower). The lighter base pairs and lighter unpaired bases represent the particular loop.

To summarize, first cross inductively constructs all roots of the skeleta-trees, second cross generates the skeleta-trees themselves and third it saturates the skeleta.

Loops

We next discuss loops of 3-noncrossing RNA structures. Loops are not only the basic building blocks for the mfe-evaluation but also of importance for the coarse grained notion of pseudoknot-shapes, discussed in Subsection. Let α be an arc in the 3-noncrossing RNA structure, S and denote by $A_S(\beta)$ the set of S -arcs that cross β . Clearly, we have $\beta \in A_S(\alpha)$ if and only if $\alpha \in A_S(\beta)$. An arc $\alpha \in A_S(\beta)$ is called a minimal, β -crossing arc if there exists no $\alpha' \in A_S(\beta)$ such that $\alpha' < \alpha$.

Let the interval $[i, j]$ denote the sequence

$$(i, i + 1, \dots, j - 1, j).$$

It is shown in [46] that any 3-noncrossing RNA structure can be uniquely decomposed into the following four loop-types:

(1) a *hairpin-loop* is a pair

$$((i, j), [i + 1, j - 1])$$

where (i, j) is an arc.

(2) an *interior-loop* is a sequence

$$((i_1, j_1), [i_1 + 1, i_2 - 1], (i_2, j_2), [j_2 + 1, j_1 - 1]),$$

where (i_2, j_2) is nested in (i_1, j_1) .

(3) a *multi-loop*, see Figure 8, is a sequence

$$((i_1, j_1), [i_1 + 1, \omega_1 - 1], S_{\omega_1}^{\tau_1}, [\tau_1 + 1, \omega_2 - 1], S_{\omega_2}^{\tau_2}, \dots)$$

where $S_{\omega_h}^{\tau_h}$ denotes a pseudoknot structure over $[\omega_h, \tau_h]$ (i.e. nested in (i_1, j_1)) and subject to the following condition: if all $S_{\omega_h}^{\tau_h} = (\omega_h, \tau_h)$, i.e. all substructures are simply arcs, for all h , then $h = 2$.

(4) a *pseudoknot*, see Figure 9, consisting of the following data:

(P1) a set of arcs

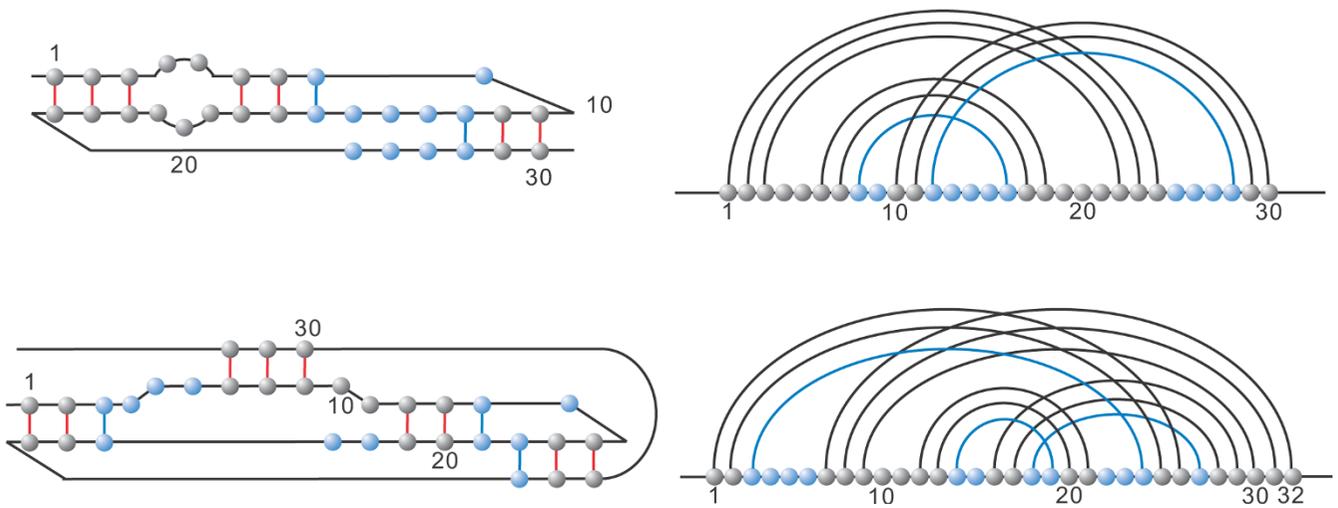


Figure 9
Pseudoknot-loops. The lighter base pairs and lighter unpaired bases represent the particular loop. In the lower structure the arc (3, 24) is lighter, since it is minimal crossing with respect to the arc (9, 30), *not* contained in any pseudoknot-loop.

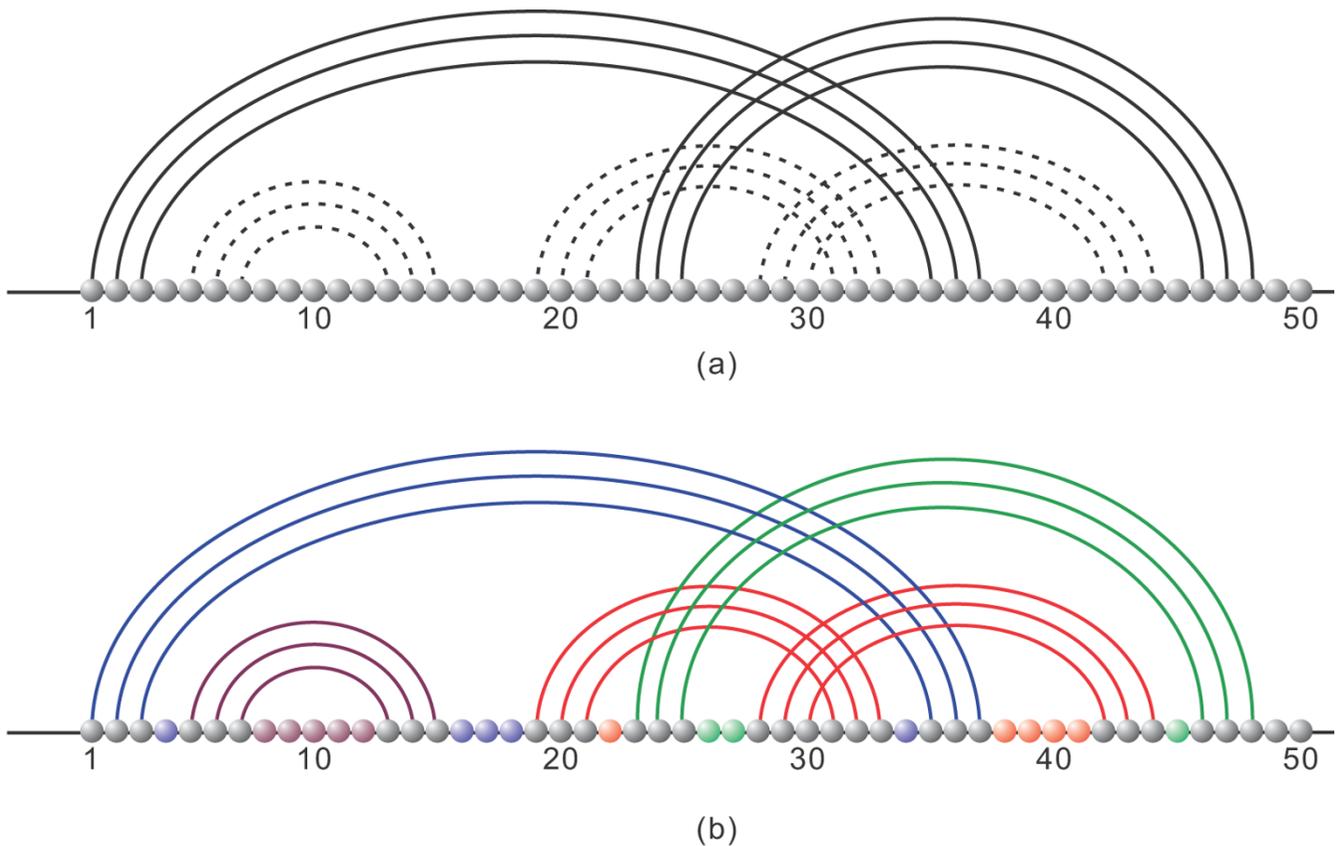


Figure 10
Shadow and loop-decomposition. A pseudoknot structure decomposed in its two shadows (top) and its loop-decomposition (bottom).

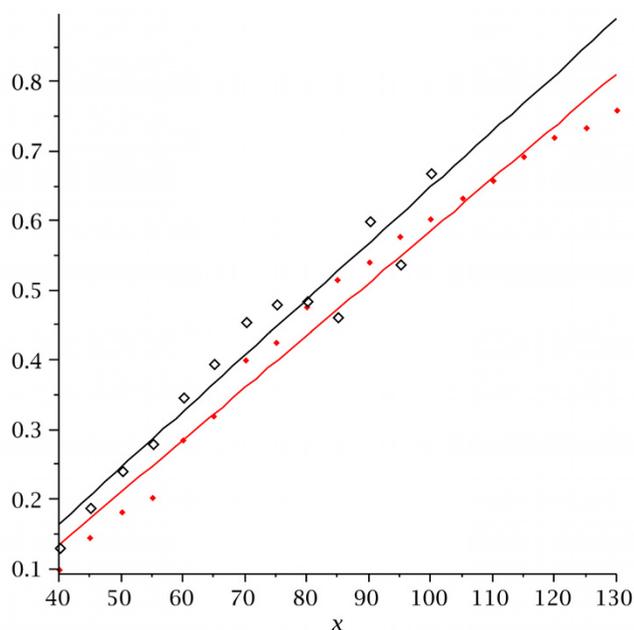


Figure 11
The fraction of pseudoknot structures over sequence length. Pseudoknot fractions for cross₃ (hollow diamonds) and pseudoknot fractions for cross₄ (solid diamonds).

$$P = \{(i_1, j_1), (i_2, j_2), \dots, (i_t, j_t)\},$$

where $i_1 = \min\{i_s\}$ and $j_t = \max\{j_s\}$, such that

(i) the diagram induced by the arc-set P is irreducible, i.e. the line-graph of P is connected and

(ii) for each $(i_s, j_s) \in P$ there exists some arc β (not necessarily contained in P) such that (i_s, j_s) is minimal β -crossing.

(P2) all vertices $i_1 < r < j_t$, not contained in hairpin, interior or multi-loops.

Decomposition

We now show that each 3-noncrossing RNA structure can uniquely be constructed by simpler substructures [46]. Furthermore, each 3-noncrossing RNA structure has a unique loop decomposition-the basis of our energy evaluation. We remark that assertion (b) of the following result remains valid for arbitrary crossing number, k .

Theorem. Suppose $k \geq 2, \tau \geq 3$.

(a) Any k -noncrossing, t -canonical RNA structure corresponds to an unique sequence of shadows.

(b) Any $\langle 3, \tau \rangle$ -structure has an unique loop-decomposition.

In Figure 10 we illustrate how these decompositions work.

Results and discussion

Our results are organized in two sections. First we describe our findings with respect to the statistics of pseudoknot RNA structures and second we present our data with respect to the particular organization of the sequences in neutral networks.

Minimum free energy RNA pseudoknot structures

In this section we present some key statistics on pseudoknotted RNA structures. In order to put our findings into context we consider two variants of cross: first, cross₃, which generates 3-noncrossing, 3-canonical mfe-structures and second, cross₄, which produces 3-noncrossing, 4-canonical mfe-structures.

The fraction of pseudoknots

We next compute the fraction of RNA structures with pseudoknots within all structures for cross₃ and cross₄. Figure 11 displays the fraction of structures with pseudoknots as a function of sequence length. It is evident that the fraction of pseudoknotted structures is monotone with respect to the sequence length. Our data are based on folding 2000 random sequences via cross and suggest an linear relation. In particular, for $n = 100$, approximately 50% of the structures folded by both versions of cross contain pseudoknots.

Pseudoknot-shapes

Next we study the dominant pseudoknot-shapes as a function of sequence length. Our notion of pseudoknot-shape is based on k -noncrossing cores [29] discussed in Subsection. The shape of a structure S , is a subset of the core-arcs, induced by all arcs either contained in pseudoknots or arcs contained in multi-loops which contain nested pseudoknots. In other words, a pseudoknot-shape contains all pseudoknot-arcs and all arcs affecting the energy of pseudoknots, see Figure 12. In Figure 12 we display for cross₃ and cross₄ the dominant types. The shape data are obtained by folding 2000 random sequences. In Figure 13 we display the fraction of sequences on which cross₃ and cross₄ coincide, based on folding 2000 random sequences.

Stack-statistics in pseudoknot RNA

It is wellknown that large stacks contribute to a low mfe of a structure. In this section we relate the distribution of stacks in random structures to the distribution of stacks in mfe-pseudoknot structures generated by cross. This provides insight in what particular spectrum of pseudoknot structures cross produces.

Let us first discuss the distribution of stacks in random pseudoknot structures. The naive approach would be to

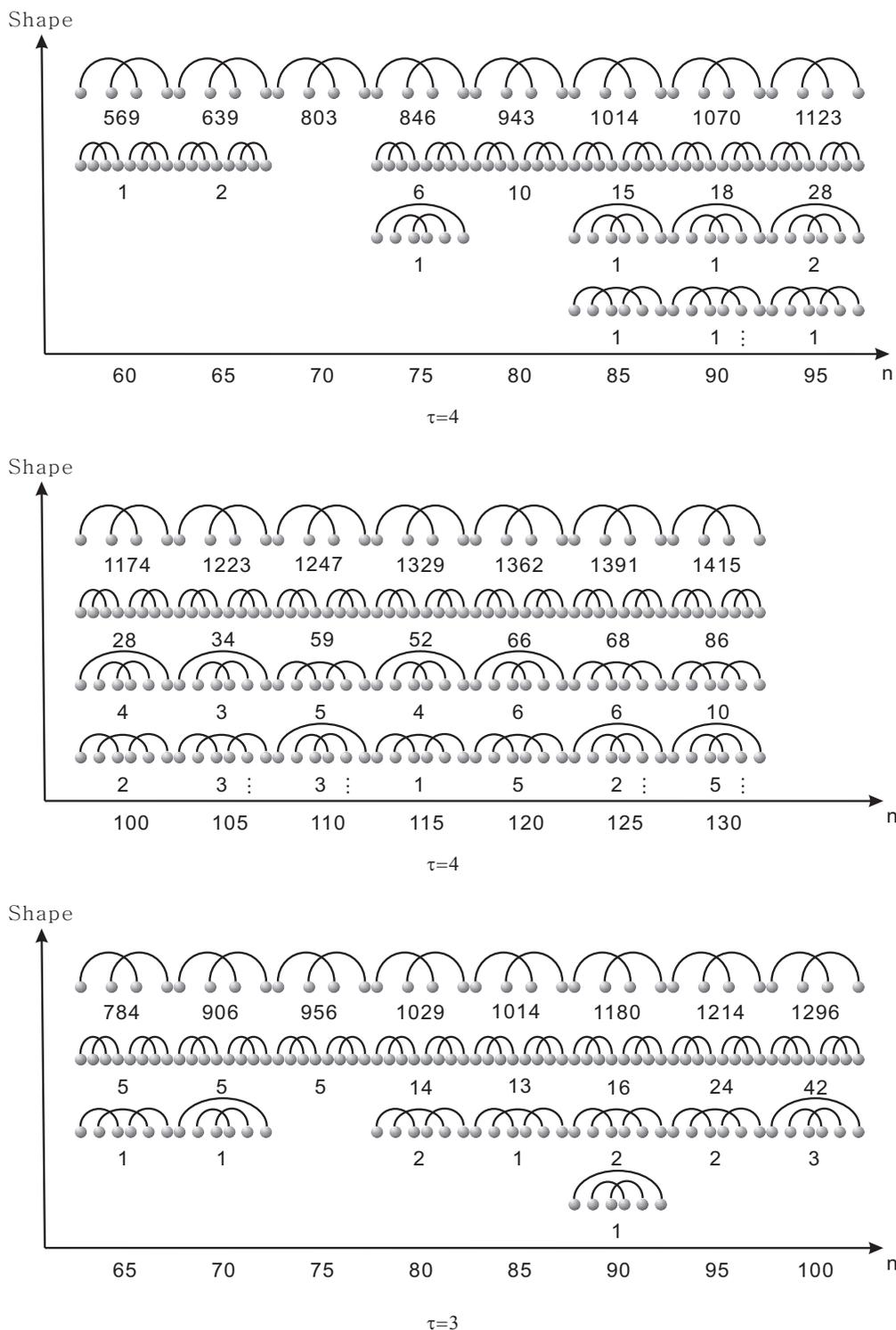


Figure 12
The dominant pseudoknot shapes for $\tau = 4$ and $\tau = 3$. The four dominant shapes displayed from top to bottom according to their frequency. The shapes are obtained sampling 2000 random sequences and labelled by the frequency of their occurrence. For $n = 40, 45, 50, 55$ we only have one shape with frequencies (200, 292, 364, 506), (260, 377, 482, 563), where the coordinates represent cross_4 and cross_3 , respectively. We have only one shape for $n = 60$, cross_3 with frequency 697.

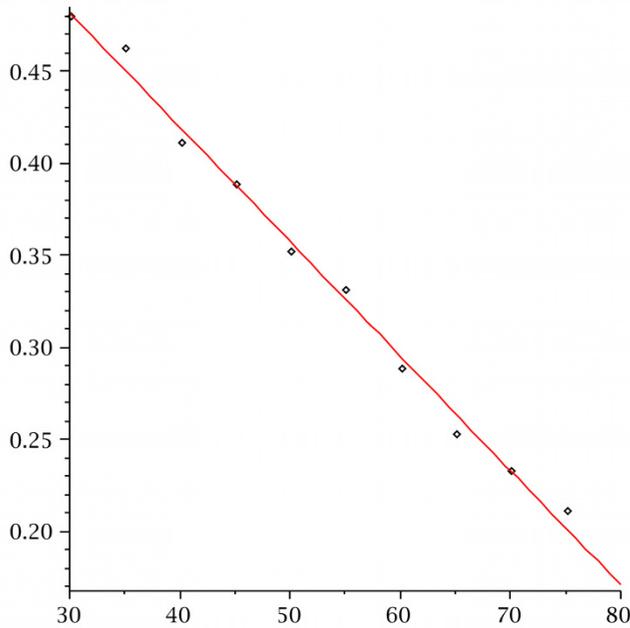


Figure 13
Fraction of same structures over sequence length.
 The fraction of sequences for which cross₄, generating pseudoknot structures with minimum stack-length four. The data are derived by folding 20, 000 random sequences.

generate a random structure and count the number of stacks. However, it is at present time not known how to construct a random pseudoknot structure with uniform probability. Therefore we have to employ a different strategy in order to obtain this distribution for random structures. The key idea [47] is to consider the bivariate generating function

$$T_{k,\tau}(x, u) = \sum_{n \geq 0} \sum_{0 \leq t \leq \frac{n}{2}} T_{k,\tau}(n, t) u^t x^n \tag{3}$$

where $T_{k,\tau}(n, t)$ denotes the number of k -noncrossing, τ -canonical pseudoknot structures having exactly t stacks.

$T_{k,\tau}(x, u)$ can be computed using the cores introduced in Section. The stack-distribution is now given by

$$P(X_{k,\tau}^n = t) = T_{k,\tau}(n, t) / T_{k,\tau}(n) \tag{4}$$

and via singularity analysis one can show that this distribution becomes asymptotically normal with mean $\mu_{k,\tau}$ and variance $\sigma_{k,\tau}^2$ given by

$$\mu_{k,\tau} = -\frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)} \tag{5}$$

$$\sigma_{k,\tau}^2 = \left(\frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)} \right)^2 - \frac{\gamma''_{k,\tau}(0)}{\gamma_{k,\tau}(0)} \tag{6}$$

where $\gamma_{k,t}(u)$ is the unique dominant singularity parameterized by $u = e^s$. In Table 2 we display the values $\mu_{k,\tau}$ and $\sigma_{k,\tau}^2$ for $k = 2, 3, 4$ and $\tau = 3, \dots, 7$. Accordingly the number of stacks scales linearly with sequence length and so does the number of loops, since each loop corresponds to a stack. In Figure 14 we present the stack distributions of 3000 structures of random sequences folded by cross₄ and the normal distribution obtained from Table 2 (lhs). Analogously we present the stack distributions of 3000 structures of random sequences folded by cross₅ and the normal distribution obtained from Table 2 (rhs).

Neutrality and local connectivity

The mapping from sequence to structures plays an important role for evolution [23,43,48]. One of its key roles is to facilitate the search of a sequence-population for better adapted shapes. In tis context, Table 1 contains some non-trivial information about the mapping from RNA sequences into their pseudoknot structures. To be precise, Table 1, in combination with central limit theorems for the number of arcs in k -noncrossing RNA structures [49,50] allows us to conclude that there exist exponen-

Table 2: Mean and variances. Mean and variances of the normal limit distributions of the numbers of stacks in pseudoknot RNA structures for different k and τ . We list mean (μ) and variance (σ^2).

	k = 2		k = 3		k = 4	
	μ	σ^2	μ	σ^2	μ	σ^2
$\tau = 3$	0.090323	0.0189975	0.115473	0.0086760	0.123509	0.0076977
$\tau = 4$	0.071677	0.0131316	0.086554	0.0055685	0.091737	0.0049917
$\tau = 5$	0.059591	0.0098165	0.069467	0.0039688	0.073166	0.0035769
$\tau = 6$	0.051092	0.0077233	0.058149	0.0026885	0.060964	0.0027313
$\tau = 7$	0.044774	0.0062991	0.050083	0.0017584	0.052319	0.0021788

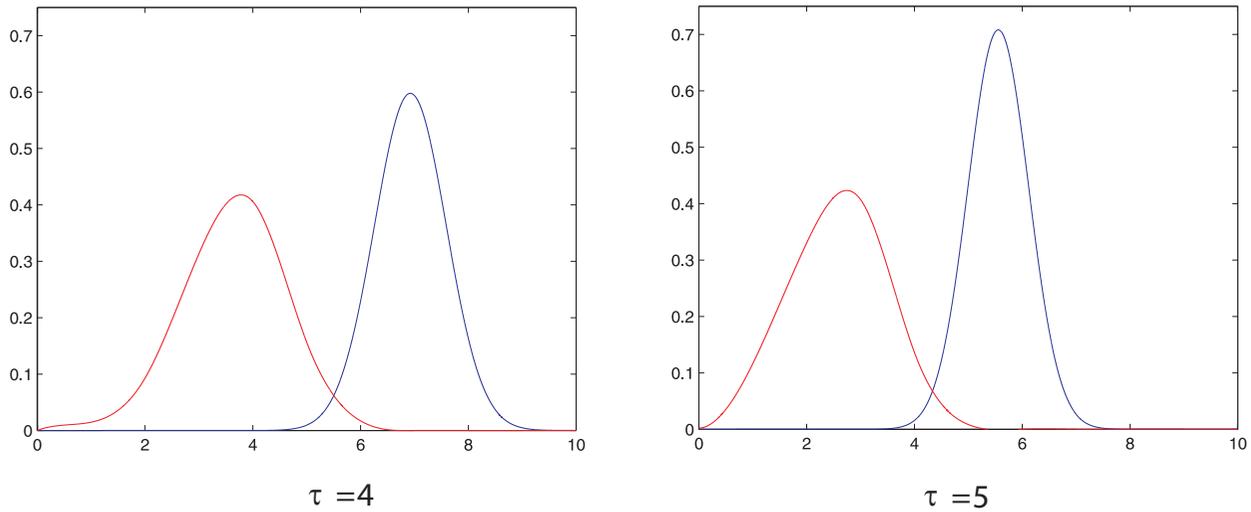


Figure 14

Stack (loop) numbers in pseudoknot RNA. We present the stack distributions based on 3000 random sequences of length 80 for cross_4 (lhs) and cross_5 (rhs). In addition we display the normal distributions (lighter) as implied by theory with the parameters $\mu_4 = 0.086554$, $\sigma_4^2 = 0.0055685$ and $\mu_5 = 0.069467$, $\sigma_5^2 = 0.0039688$.

tially many k -noncrossing canonical structures with exponentially large preimages. Indeed, according to Table 1 the exponential growth rate of the number of k -noncrossing canonical structures, $3 = k = 9$ is strictly smaller than four—the growth rate of the space of all sequences over the natural alphabet.

The central limit theorems for the number of arcs of k -noncrossing, canonical pseudoknot structures [50] exhibit a mean of $0.39 n$ and a variance of $0.041 n$. We conclude from this that sequence to structure maps in pseudoknot RNA structures cannot be trivial, since the preimages of particular structures have exponential growth rates strictly smaller than four. As a result the number of canonical pseudoknot structures grows exponentially. Accordingly, a sequence to structure map in pseudoknot RNA necessarily generates exponentially many canonical structures.

In light of this, the interesting question then becomes how the set of sequences folding into a given structure is "organized" in sequence space. The analysis presented in this section is analogous to the investigations for RNA secondary structures [23,51] and can be viewed as a basic protocol for the local statistics of a genotype-phenotype map. The only exception is Subsection, which elaborates on the novel concept of local connectivity [48].

It is only possible to derive local statistics, since, for instance, exhaustive computations of the set of all sequences over the natural alphabet with fixed pseudoknot structure for $n > 40$ is at present time impossible.

Neutral walks

Let us consider a fixed RNA structure, S . Let furthermore $C[S]$ denote the set of S -compatible sequences, consisting of all sequences that have at any two paired positions one of the 6 nucleotide pairs

$$(A, U), (U, A), (G, U), (U, G), (G, C), (C, G).$$

The structure S motivates to consider a new adjacency relation within $C[S]$. Indeed, we may reorganize a sequence (x_1, \dots, x_n) into the pair

$$((u_1, \dots, u_{n_u}), (p_1, \dots, p_{n_p})), \tag{7}$$

where the u_j denote the unpaired nucleotides and the $p_j = (x_i, x_k)$ all base pairs, respectively, see Figure 15. We can then view $v_u = (u_1, \dots, u_{n_u})$ and $v_p = (p_1, \dots, p_{n_p})$ as elements of the formal cubes $Q_4^{n_u}$ and $Q_6^{n_p}$, implying the new adjacency relation for elements of $C[S]$.

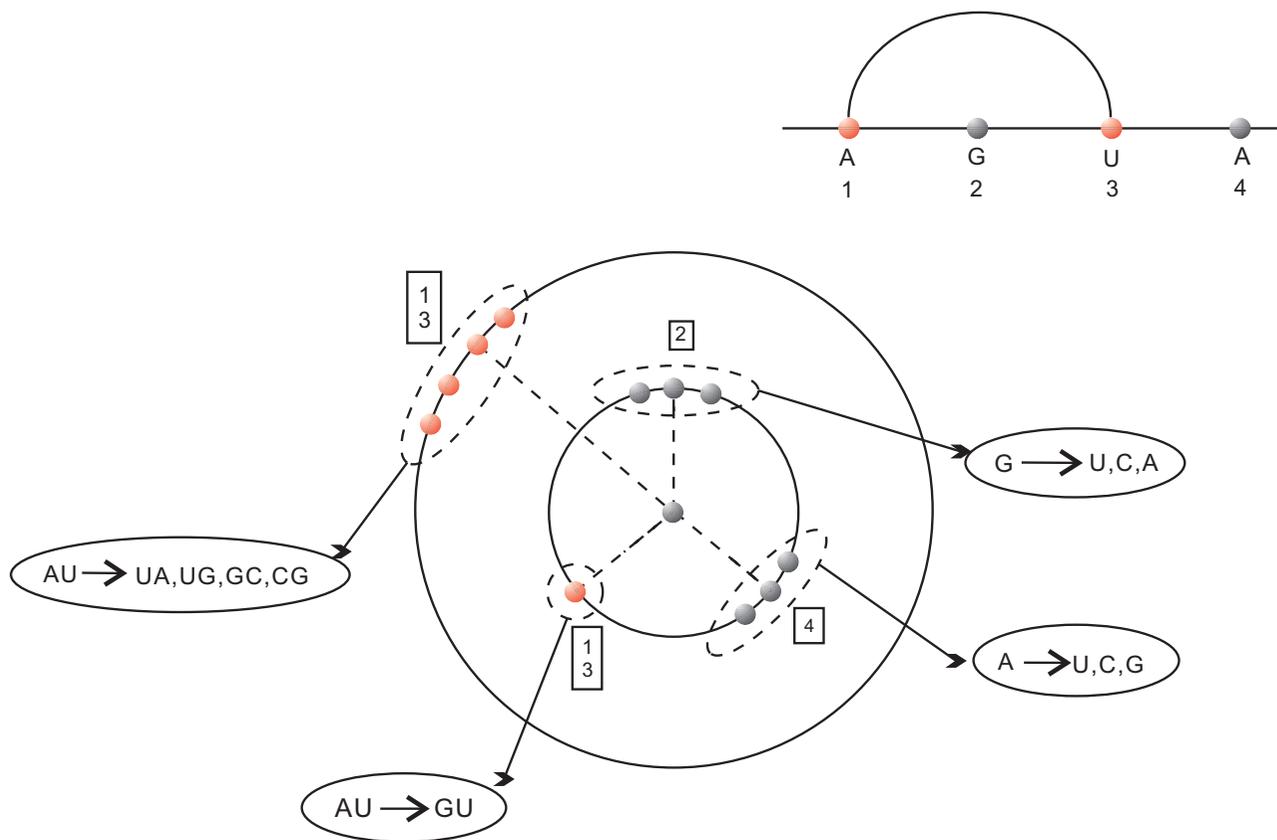


Figure 15
Compatible neighbors in sequence space. Diagram representation of an RNA structure (upper right) and its induced compatible neighbors in sequence space (lower left). Note that each base pair gives rise to 5 compatible neighbors exactly one of which is in Hamming distance one.

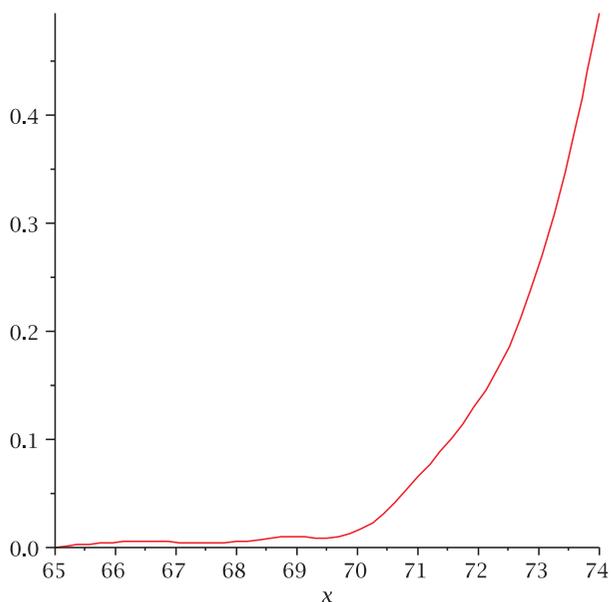
Accordingly, there are two types of compatible neighbors in sequence space: u- and p-neighbors: a u-neighbor has Hamming distance one and differs exactly by a point mutation at an unpaired position. Analogously a p-neighbor differs by a compatible base pair-mutation, see Figure 15. Note however, that a p-neighbor has either Hamming distance one ((G, C) → (G, U)) or Hamming distance two ((G, C) → (C, G)). We call a u- or a p-neighbor, γ , a compatible neighbor. If γ is contained in the neutral network we refer to γ as a neutral neighbor. This gives rise to consider the compatible- and neutral distance, denoted by $C(v, v')$ and $N(v, v')$. These are the minimum length of a C[S]-path and path in the neutral network between v and v' , respectively.

Our basic experiment is as follows: We select a (random) sequence, v and fold it into the structure $S(v)$. We then proceed inductively: assume v_i is constructed. We randomly select some neutral (compatible) neighbor of v_i , denoted by v_{i+1} , subject to the condition $d_H(v, v_{i+1}) > d_H(v,$

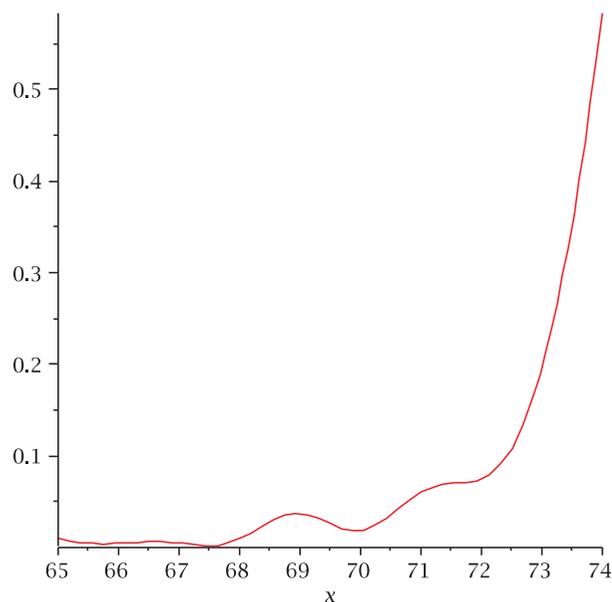
$v_i)$, where $d_H(x, y)$ denotes the Hamming distance. If no such neighbor exists we choose some $v_{i+1} \neq v_i$ with the property $d_H(v, v_{i+1}) = d_H(v, v_i)$. If all neutral v_i -neighbors satisfy $d_H(v, v_{i+1}) < d_H(v, v_i)$ we stop and output the integer $d_H(v, v_i)$. In Figure 16 we study 200 neutral walks for the following four structures: first an *H*-pseudoknot loop structure (a), second a hairpin-loop structure (b), third an interior-loop structure (c) and finally the phenylalanine tRNA structure (d), see Figure 17. Our findings are in accordance with those for RNA secondary structures. One can easily neutrally traverse sequence space, suggesting the picture of vast, connected networks composed by neutral sequences.

Neutral neighbors

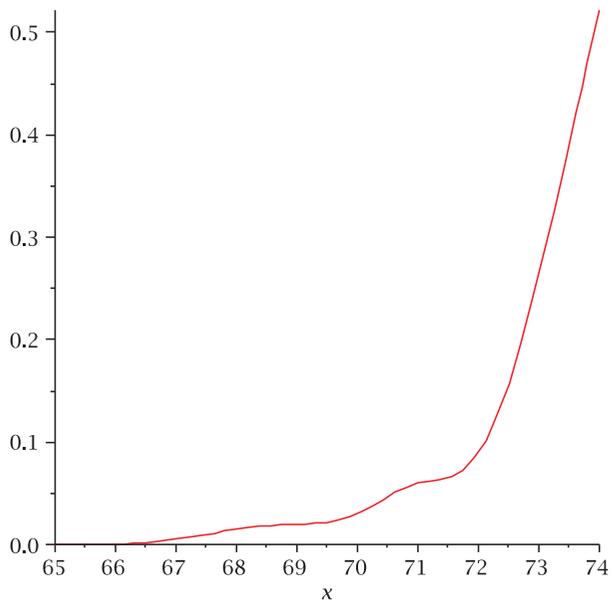
Complementing the analysis of neutral walks, we study now the distribution of neutral neighbors. Recall that a neutral neighbor of a sequence v with respect to the structure $S = S(v)$ is a u- or a p-neighbor, γ , contained in the neutral network of S . It has Hamming distance one or



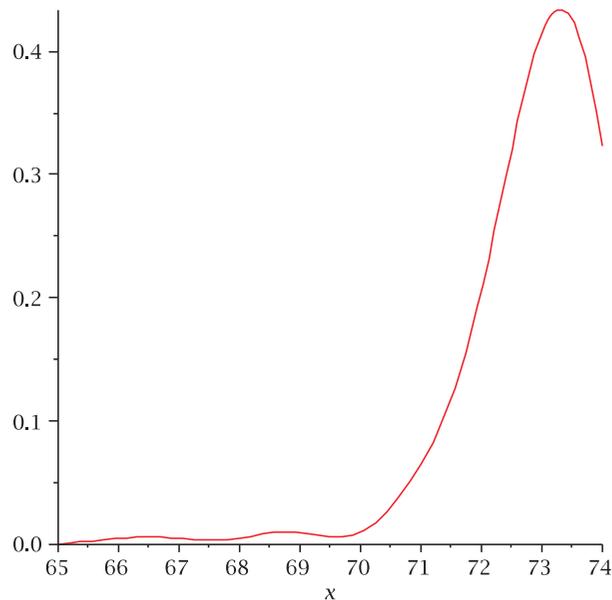
(a)



(b)



(c)



(d)

Figure 16

Neutral walks. Distance distribution of neutral walks for the corresponding four structures of $cross_4$ in Figure 17 based on 200 random paths for each.

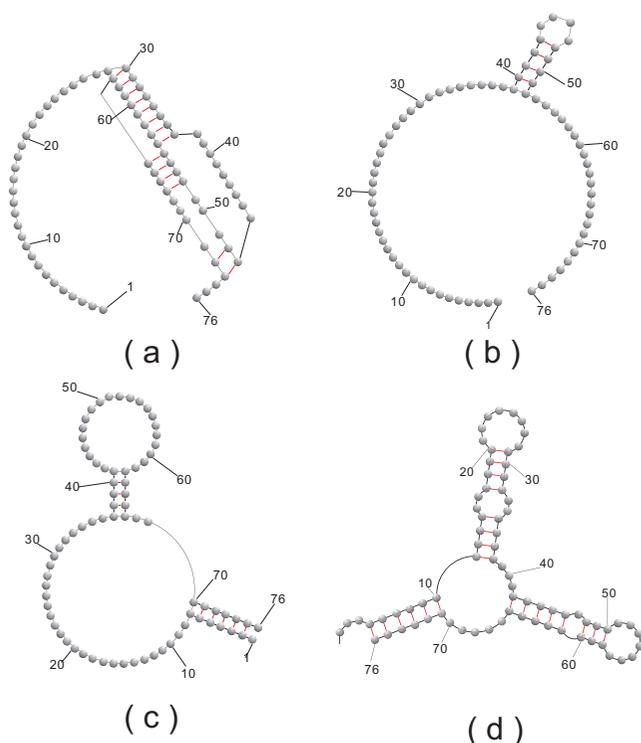


Figure 17
Four particular pseudoknot structures. (a).H-loop (b).Hairpin-loop (c).interior-loop (d).tRNA.

two, depending on whether it is induced by a point or base pair mutation, see Figure 15. The distribution of neutral neighbors provides relevant information about the mutational robustness of the structure S . The data presented here, are obtained in the course of the neutral walk experiments, displayed in Figure 16. They are given in Figure 18. In order to put things into context we also present in Figure 19 the distribution of neutral neighbors for 10000 random sequences folded by cross_4 .

Local connectivity

Connectivity of a subgraph, Γ_n , of an n -cube alone does not imply that a small Hamming distance implies a small distance in Γ_n . For neutral sequences this means that two neutral sequences with Hamming distance less than four, are possibly connected via a neutral path of much greater length. Evidently, for molecular evolution it is therefore not connectivity but the existence of these short paths what matters. Local connectivity is a property which guarantees the existence of these short paths. If Γ_n is locally connected then a small Hamming distance does imply a Γ_n -distance scaled by at most a factor of $\Delta > 0$. We shall begin by studying local connectivity for random induced subgraphs of n -cubes, i.e. where we select sequences with independent probability λ_n . Then we transfer the derived

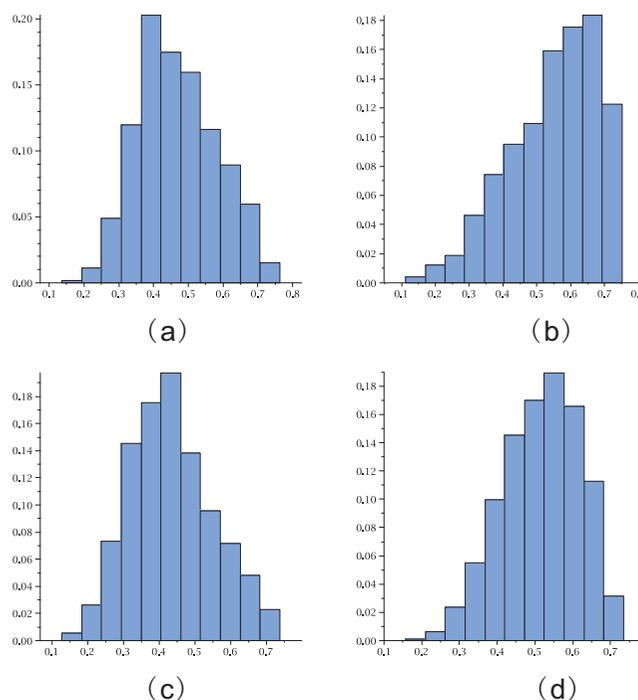


Figure 18
The fraction of neutral neighbors for the pseudoknot structures. (a), (b), (c) and (d) are based on sequences in their random 200 paths for cross_4 .

concepts to neutral networks of RNA pseudoknot structures.

We call Γ_n is locally connected if and only if almost surely (a.s.)

$$(\dagger) \quad \exists \Delta > 0; \quad d_{\Gamma_n}(v, v') \leq \Delta d_{Q_2^n}(v, v'),$$

provided v, v' are in Γ_n . We immediately observe that, trivially, for any finite n such Δ exists. However, the key point is that (\dagger) employs the notion "almost surely", i.e. it holds for arbitrary n .

Random graph theory [48] shows that on the one hand, for λ_n smaller than n^δ / \sqrt{n} , where $\delta > 0$ is arbitrarily small, there exists a.s. no finite Δ satisfying (\dagger) . On the other hand, for λ_n larger than or equal to n^δ / \sqrt{n} , there exists a.s. some finite Δ satisfying (\dagger) . In other words, there exists a threshold value for local connectivity. Since random subgraphs of n -cubes have giant components for $\lambda_n = (1 + \varepsilon) / n$, where $\varepsilon > 0$ [52] we can conclude that local connectivity emerges distinctly later in the evolution of random subgraphs of n -cubes.

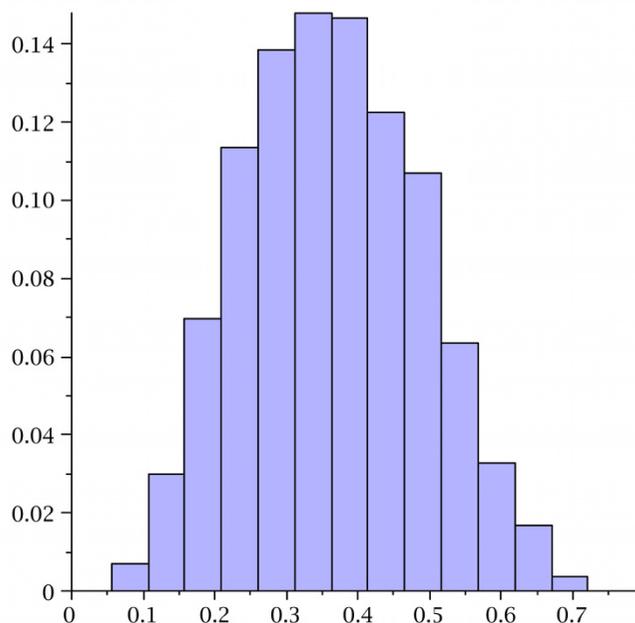


Figure 19
Neutral fraction distribution. Neutral fraction distribution of 20, 000 random sequences for cross₄.

Suppose we are given a structure S and sequence v , contained in its neutral network. By construction, local connectivity refers to the two n -cubes $Q_4^{n_u}$ and $Q_6^{n_p}$ induced by S , see Figure 20. Let

$$C_2 = |\{v' \mid C(v, v') = 2\}|$$

be the cardinality of the set of sequences in compatible distance two. Then the degree of local connectivity of S at v is given by

$$D_S(v) = |\{v' \mid C(v, v') = 2, N(v, v') = 4\}| C_2^{-1}. \tag{8}$$

In other words, $D_S(v)$ is the fraction of locally connected vertices of the compatible distance two neighbors of v , that can be obtained via a neutral path of length at most four.

We perform the following experiment: we consider neutral walks for the UTR-pseudoknot structure of the mouse hepatitis virus displayed in Figure 1, see Subsection. Along these walks we compute the locality degree $D_S(v_i)$ and the total number of locally connected sequences. Our findings are presented in Figure 21. We can report that the degree of local connectivity is, as suggested by random graph theory, almost 100%.

Conclusion

RNA pseudoknot structures-in particular their statistical properties-are a fascinating and new territory. To our knowledge the only statistical data beyond RNA secondary structures were derived for bi-secondary structures in [53,54]. The structural concept of k -noncrossing canonical RNA structures and the resulting sequence to structure map employed for our experiments is new and represents a natural generalization of RNA secondary and bi-secondary structures. To be precise, bi-secondary structures are exactly planar 3-noncrossing RNA structures [19].

It is clear, that for sequence-length less than or equal to 100 we only encounter pseudoknots of limited complexity. Our findings presented in Figure 12 provide a transparent picture of which pseudoknot-shapes dominate for given sequence length. These results, in combination with the data on the fractions of pseudoknotted structures over sequence length show, that for $n = 80$ we have approximately 35% structures with nontrivial pseudoknots. In addition it is striking that basically all folded structures are irreducible, i.e. only a very small fraction can be decomposed into several independent substructures. This is of interest since decomposable structures can be folded much faster. It is known, [55] that Dyck-paths, i.e. path starting at the origin, having only up $(1, 1)$, or down $(1, -1)$ steps which end on the x -axis, decompose on average into three irreducible parts. This is of relevance, since a slight generalization of Dyck-path, the Motzkin-paths, having additional horizontal steps, correspond to secondary structures. Our findings suggest, that while secondary structures, decompose nontrivially, higher and higher crossing numbers change the picture. This complicates the computation of mfe-pseudoknot RNA due to their imminent irreducibility.

Both versions of cross produce analogous findings, confirming the generality of our results. The vast majority of pseudoknot-shapes is of a single type. As expected, cross₃ exhibits more structural variety due to the fact that its minimum stack-length is only three. The ratio of pseudoknot structures shifts significantly from $n = 80$ to $n = 100$ to approximately 50%. We can conclude from this that pseudoknots cannot be ignored, they evidently become the dominant structure class for n greater than or equal to 100. Figure 13 shows that the fraction of sequences for which cross₃ and cross₄ coincide, decreases linearly as a function of sequence length. This indicates that larger and larger sequences will exhibit more subtle structural elements whose emergence is facilitated by stabilizing large stacks.

Furthermore, the mfe-pseudoknot structures generated by cross are far from being random. The central limit theorems for random k -noncrossing canonical RNA structures,

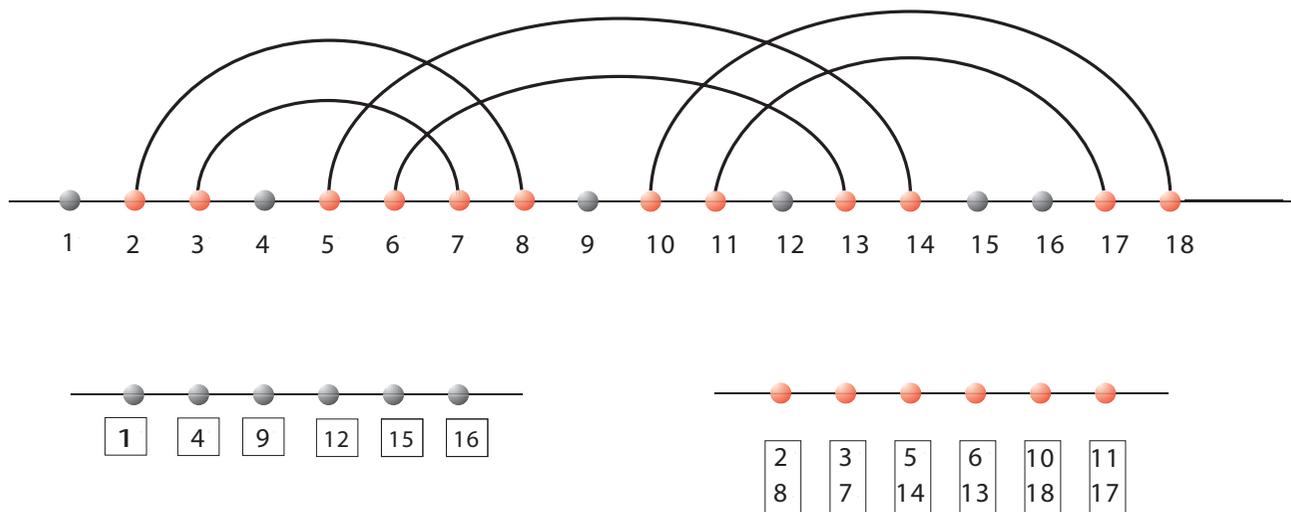


Figure 20

Deriving the two subcubes $Q_4^{n_u}$ and $Q_6^{n_p}$. A structure gives rise to rearrange its compatible sequences into unpaired and paired segment. The former is a sequence over the original alphabet **A, U, G, C** and for the latter we derive a sequence over the alphabet of base pairs, **(A, U), (U, A), (G, U), (U, G), (G, C), (C, G)**

given in Table 2 imply, that stacks and consequently loops

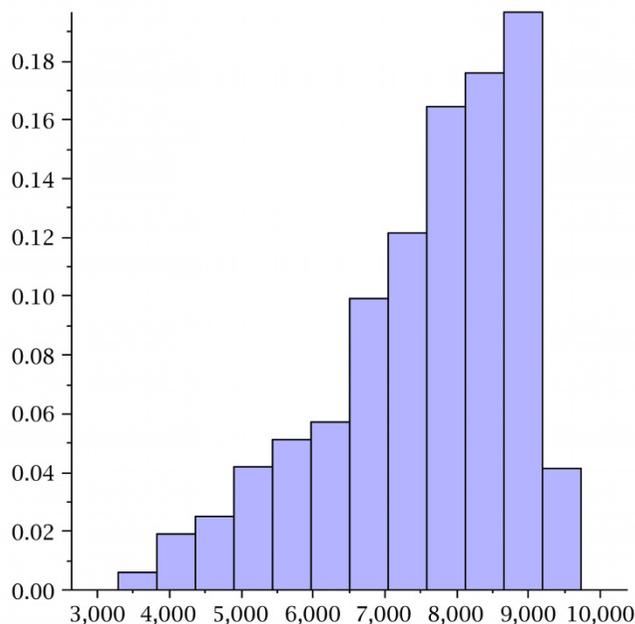


Figure 21
Local connectivity of the UTR-pseudoknot. Presented are the numbers of locally connected sequences during various neutral walks. The degree of local connectivity, $D_5(v)$, is one for all steps.

scale linearly with the sequence length. Figure 14 clearly shows that the mfe-structures, generated by $cross_4$ and $cross_5$, have for $n = 76$ two stacks less than random 3-noncrossing structures with minimum stack-length greater than four and five, respectively. This deviation is significant and indicates that mfe-pseudoknot structures are far from "typical" random structures. We remark that, while it is straightforward to generate random RNA secondary structures, it is nontrivial to obtain random pseudoknot structures. In particular, at present time, no polynomial time algorithm is known which generates a random 3-noncrossing RNA structure with uniform probability.

The organization of the sequences contained in neutral networks of RNA pseudoknot structures seems to be very analogous to the neutral networks of RNA secondary structures [23]. Figure 16 shows that neutral walks can effectively traverse sequence space and the fractions of neutral neighbors, presented in Figure 18 and Figure 19 suggest a high degree of neutrality.

We discussed in Subsection local connectivity, a property of neutral networks which implies the existence of short, neutral paths. It is apparent that local connectivity is of central importance for molecular evolution and any type of evolutionary optimization. It has been shown in [48] that local connectivity is a prerequisite for preserving any type of sequence specific information. Having a random

graph threshold value localized at $1/\sqrt{n}$, local connectivity appears much later than connectivity, being localized at $1/n$. However, the high neutrality degrees of RNA pseudoknot structures of Figure 18 and Figure 19 imply locally connected neutral networks. Our findings for the UTR-pseudoknot structure of the mouse hepatitis virus of length 56, given in Figure 21, confirm the local connectivity of neutral networks of particular pseudoknot RNA structures. At all steps of the neutral walks almost all sequences are locally connected.

List of abbreviations used

UTR: Untranslated Region; HDV: Hepatitis Delta Virus; DP: dynamic program; lhs: left hand side; rhs: right hand side.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed equally to this paper.

Acknowledgements

We are grateful to J.Z.M. Gao, H.S.W. Han and W.W.J. Peng for helpful discussions. This work was supported by the 973 Project, the PCSIRT Project of the Ministry of Education, the Ministry of Science and Technology, and the National Science Foundation of China.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 1, 2009: Proceedings of The Seventh Asia Pacific Bioinformatics Conference (APBC) 2009. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/10?issue=S1>

References

- Penner RC, Waterman MS: **Spaces of RNA secondary structures.** *Adv Math* 1993, **101**:31-49.
- Waterman MS: **Combinatorics of RNA hairpins and cloverleaves.** *Stud Appl Math* 1979, **60**:91-96.
- Smith TF, Waterman MS: **RNA secondary structure.** *Math Biol* 1978, **42**:31-49.
- Schmitt WR, Waterman MS: **Linear trees and RNA secondary structure.** *Discr Appl Math* 1994, **51**:317-323.
- Howell JA, Smith TF, Waterman MS: **Computation of generating functions for biological molecules.** *J Appl Math* 1980, **39**:119-133.
- Nussinov R, Jacobson AB: **Fast Algorithm for Predicting the Secondary Structure of Single-Stranded RNA.** *Proc Natl Acad Sci, USA* 1980, **77**:6309-6313.
- Searls DB: **The language of genes.** *Nature* 2002, **420**:211-217.
- Webpage of HDV-pseudoknot structure in natural** [<http://www.ekevanbatenburg.nl/PKBASE/PKB00075.HTML>]
- Loria A, Pan T: **Domain Structure of the ribozyme from eubacterial ribonuclease.** *RNA* 1996, **2**:551-563.
- Konings DAM, Gutell RR: **A comparison of thermodynamic foldings with comparatively derived structures of 16s and 16s-like rRNAs.** *RNA* 1995, **1**:559-574.
- Schneider D, Tuerk C, Gold L: **Selection of high affinity RNA ligands to the bacteriophage R17 coat protein.** *J Mol Biol* 1992, **228**:862-869.
- Chamorro M, Parkin N, Varmus HE: **An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA.** *Proc Natl Acad Sci, USA* 1992, **89**(2):713-7. 1309954
- Lyngsø RB, Pedersen CNS: **RNA Pseudoknot Prediction in Energy-Based Models.** *J Comp Biol* 2000, **7**:409-427.
- Rivas E, Eddy S: **A dynamic programming algorithm for RNA structure prediction including pseudoknots.** *J Mol Biol* 1999, **285**(5):2053-2068.
- Uemura Y, Hasegawa A, Kobayashi S, Yokomori T: **Tree adjoining grammars for RNA structure prediction.** *Theor Comp Sci* 1999, **210**:277-303.
- Akutsu T: **Dynamic programming algorithms for RNA secondary prediction with pseudoknots.** *Discr Appl Math* 2000, **104**:45-62.
- Tacker M, Stadler PF, Bornberg-Bauer EG, Schuster P, Hofacker IL, Schuster P: **Algorithm independent properties of RNA secondary structure predictions.** *Europ Biophys* 1996, **25**:115-130.
- Jin EY, Reidys CM: **Asymptotic enumeration of RNA structures with pseudoknots.** *Bull Math Biol* in press.
- Jin EY, Qin J, Reidys CM: **Combinatorics of RNA structures with Pseudoknots.** *Bull Math Biol* 2008, **70**(1):45-67.
- Waterman MS, Smith TF: **Rapid dynamic programming methods for RNA secondary structure.** *Adv Appl Math* 1986, **7**:455-464.
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures.** *Monatsh Chem* 1994, **125**:167-188.
- Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.** *Nucl Acids Res* 1981, **9**:133-148.
- Reidys CM, Stadler PF, Schuster P: **Generic properties of combinatorial maps: neutral networks of RNA secondary structures.** *Bull Math Biol* 1997, **59**(2):339-397.
- Schultes EA, Bartel DP: **Implications for the Emergence of New Ribozyme Folds.** *Science* 2000, **289**(5478):448-452.
- Jolly A, (Ed): **Mapping RNA form and function.** *Science* 2005, **309**:1441-1632.
- Ma G, Reidys CM: **Canonical RNA Pseudoknot Structures.** *J Comp Biol* in press.
- Chen WYC, Deng EYP, Du RRX, Stanley RP, Yan CH: **Crossings and nestings of matchings and partitions.** *Trans Am Math Soc* 2007, **359**:1555-1575.
- Chen WYC, Qin J, Reidys CM: **Crossing and Nesting in Tangled diagrams.** *Elec J Comb* 2008, **15**:
- Jin EY, Reidys CM: **RNA-LEGO: Combinatorial Design of Pseudoknot RNA.** *Adv Appl Math* in press.
- Hofacker IL, Schuster P, Stadler PF: **Combinatorics of RNA Secondary Structures.** *Discr Appl Math* 1998, **88**:207-237.
- McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure.** *Biopolymers* 1990, **29**:1105-1119.
- Fresco JR, Alberts BM, Doty P: **Some Molecular Details of the Secondary Structure of Ribonucleic Acid.** *Nature* 1960, **188**:98-101.
- Jun IT, Uhlenbeck OC, Levine MD: **Estimation of Secondary Structure in Ribonucleic Acids.** *Nature* 1971, **230**:362-367.
- DeLisi C, Crothers DM: **Prediction of RNA secondary structure.** *Proc Natl Acad Sci USA* 1971, **68**:2682-2685.
- Huynen M, Stadler PF, Fontana W: **Smoothness within ruggedness: the role of neutrality in adaptation.** *Proc Natl Acad Sci USA* 1996, **93**:397-401.
- Babajide A, Hofacker IL, J SM, Stadler PF: **Neutral Networks in Protein Space A Computational Study Based on Knowledge-Based Potentials of Mean Force.** *Folding Design* 1997, **93**:261-269.
- Schuster P: **Genotypes with phenotypes: Adventures in an RNA Toy World.** *Biophys Chem* 1997, **6**:75-110.
- Fontana W, Schuster P: **Shaping Space: The Possible and the Attainable in RNA Genotype-Phenotype Mapping.** *J Theor Biol* 1998, **194**:491-515.
- Stadler PF: **Fitness Landscapes Arising from the Sequence-Structure Maps of Biopolymers.** *J Mol Struct (THEOCHEM)* 1999, **463**:7-19.
- Schuster P, Fontana W: **Chance and Necessity in Evolution.** *Lessons from RNA Physica* 1999, **133**:427-452.
- Reidys CM, Stadler PF: **Combinatorial Landscapes.** *SIAM Review* 2002, **44**:3-54.
- Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, Stadler PF: **Automatic Detection of Conserved RNA Struc-**

- ture Elements in Complete RNA Virus Genomes.** *Nucl Acids Res* 1998, **26**:3825-2836.
43. Schuster P, Fontana W, Stadler PF, Hofacker IL: **From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures.** *Proc Roy Soc Lond B* 1994, **255**:279-284.
 44. Gruener W, Giegerich R, Strothmann D, Reidys CM, J W, Hofacker IL, Stadler PF, Schuster P: **Analysis of RNA sequence structure maps by exhaustive enumeration I. Neutral networks.** *Monatsh Chem* 1996, **127**:375-389.
 45. Gruener W, Giegerich R, Strothmann D, Reidys CM, J W, Hofacker IL, Stadler PF, Schuster P: **Analysis of RNA sequence structure maps by exhaustive enumeration. II.** *Monatsh Chem* 1996, **127**:355-374.
 46. Huang FWD, Peng WWP, Reidys CM: **Folding RNA pseudoknot structures.** . [In preparation].
 47. Han HSW, Reidys CM: **Stacks in canonical RNA pseudoknot structures.** *Comp Appl Math* in press.
 48. Reidys CM: **Local Connectivity of Neutral Networks.** *Bull Math Biol* in press.
 49. Jin EY, Reidys CM: **Central and Local Limit Theorems for RNA Structures.** *J Theor Biol* 2008, **250**(3):547-559.
 50. Huang FWD, Reidys CM: **Statistics of canonical RNA pseudoknot structures.** *J Theor Biol* in press.
 51. Fontana W, Schuster P: **Shaping Space: the Possible and the Attainable in RNA Genotype-Phenotype Mapping.** *J Theor Biol* 1998, **194**(4):491-515.
 52. Reidys CM: **Large components in random induced subgraphs of N-cubes.** *Discr Math* in press.
 53. Stadler PF, Haslinger C: **RNA Structures with Pseudo-Knots.** *Bull Math Biol* 1999, **61**:437-467.
 54. Haslinger C: **RNA Structures with Pseudoknots.** In *PhD thesis University of Vienna*; 1997.
 55. Shapiro L: **A survey of the Riordan Group.** *Proc Amer Math Soc* 1994.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

