Sequence-structure relations of pseudoknot RNA

Fenix W.D. Huang , Linda Y.M. Li and Christian M. Reidys*[1]

[1]Center for Combinatorics, LPMC-TJKLC, Nankai University, Tianjin 300071, PR China

Email: reidys@nankai.edu.cn;

*Corresponding author

Abstract

In this paper we study sequence structure relations of RNA. As structures we consider RNA pseudoknot structures with at most two mutually crossing bonds. These structures are folded by a novel, *ab initio* prediction algorithm cross. After giving some background on RNA pseudoknot structures we present various, statistical results on the mapping from RNA sequences of length 76 into 3-noncrossing RNA structures. We study properties, like the fraction of pseudoknotted structures, dominant shapes, neutral walks, neutral neighbors and local connectivity, which are of particular interest in the context of molecular evolution of RNA.

1 Background

Three decades ago Michael Waterman pioneered the combinatorics and ab initio prediction of the then rather exotic ribunucleic acid (RNA) secondary structures [1–5]. The motivation for this work was coming from a fundamental dichotomy represented by RNA. On the one hand RNA is described by its primary sequence, a linear string composed by 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 phenotypic executive. Therefore one molecule stands for both geno- and phenotype.

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 form of transfer RNA. The realization that RNA combines features of proteins and DNA led to the ”RNA world” hypothesis for the origin of life. The idea was that DNA and the much more versatile proteins took over RNA's functions in the transition from the ”RNA-world” to the “DNA/protein-world”.

Let us have a closer look at RNA phenotypes. RNA molecules form “helical” structures by pairing their nucleotides and thereby lowering their minimum free energy (mfe). Originally, these bonds were subject to strict combinatorial constraints, for instance “noncrossing” in RNA secondary structures. It is wellknown, however, that RNA structures are far more complex than secondary structures. One particularly prominent feature is the existence of cross-serial dependencies [6], that is crossing arcs or pseudoknots, see Fig. **??**. In fact, RNA pseudoknots are “everywhere”. They occur in functional RNA, like for instance RNAseP [7] as well as ribosomal RNA [8]. They are conserved in the catalytic core of group I introns, in plant viral RNAs pseudoknots mimic tRNA structure and in *in vitro* RNA evolution [9] experiments have produced families of RNA structures with pseudoknot motifs, when binding HIV-1 reverse transcriptase. Important mechanisms like ribosomal frame shifting [10] also involve pseudoknot interactions. For prediction algorithms the implications of cross-serial dependencies are severe–they imply a higher level of formal language: context-sensitive. In general, on this level of formal languages it is not clear whether polynomial time ab initio folding algorithms exist. Indeed, Lyngso *et.al.* [11] showed that "reasonable" classes of RNA pseudoknots require exponential time algorithms. There exist however, polynomial time folding algorithms, capable of the energy based prediction of certain pseudoknots: Rivas *et.al.* [12], Uemura *et.al.* [**?**], Akutsu [13] and Lyngso [11].

In analogy to RNA secondary structures, in order to analyze RNA structure with pseudoknots, key combinatorial properties have to be identified. Without such a specification one would arrive at an impossibly large configuration space. It turned out that the notion of k-noncrossing diagrams [14] is of importance in order to arrive at such an output-class. A diagram is a graph over the vertex set $[n] = \{1, \dots, n\}$ with vertex degrees less or equal to one, represented by drawing its vertices in a horizontal line and its arcs (i, j), where $i < j$, in the upper half-plane. The vertices and arcs correspond to nucleotides and Watson-Crick (**A-U**, **G-C**) and (**U-G**) base 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 Fig. **??**. 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 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 Fig. **??**. The noncrossing of arcs has far-reaching consequences. It implies that RNA secondary structures form a context free language and allow for the dynamic programming algorithms [15], predicting the loop-based mfe secondary structure in $O(n^3)$-time and $O(n^2)$-space.

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* [16], an implementation of the folding routine [17, 18], mentioned above. Two particularly prominent results of this line of work were the existence of neutral networks, i.e. vast extended networks composed by sequences folding into a given secondary structure [19] and the Intersection Theorem [19]. The latter guarentees 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 [20]. It became evident that the "statistical" properties of this mapping played a central role in the molecular evolution of RNA.

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. Large numbers 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 elegans and humans, and their particular relationship to certain intermediates in RNA interference (RNAi). This 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 and at more and more paralles between RNA and protein are currently revealed [21].

In this paper sequence structure relations of RNA pseudoknot structures will be studied. Let us briefly overview what we know about the combinatorics of our phenotypes, ultimatively allowing to enumerate biophysically relevant pseudoknot structures [22]. The key result comes from a seemingly unrelated field, the combinatorics of partitions. Chen *et al.* proved in a seminal paper [23] a bijection between walks in Weyl chambers and k-noncrossing partitions. This bijection has recently been generalized to tangled diagrams [24]. 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, via the reflection principle it was possible to compute the generating function of k-noncrossing and k-noncrossing canonical pseudoknot RNA [**?**, 14, 22]. Subsequent singularity analysis, [?, 22] showed that the exponential growth rates of canonical pseudoknot RNA are surprisingly small, see Tab. 4.12. For instance, the number of 3-noncrossing, 3-canonical RNA structures with arc-length greater of equal than four is asymptotically given by

$$c\,n^{-5}\,2.0348^n,$$

where c is some (exlicitly known) constant. This exponential growth rate is very close to Schuster *et al.*'s finding [25] for 2-canonical RNA secondary structures with arc-length greater of equal than four

$$1.4848\,n^{-3/2}\,1.8444^n. \qquad (1)$$

For the analysis presented here, we use the algorithm cross [26], which which produces a transparent output. This algorithm does not follow the dynamic programming paradigm, generating the mfe k-noncrossing τ-canonical structure via a combination of branch and bound, as well as dynamic programming techniques. cross inductively constructs k-noncrossing, τ-canonical RNA structures via motifs. Currently full loop-based energy models are derived and implememted 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 3. While cross is an exponential time algorithm it allows to fold sequences of length 100 within a few minutes.

2 Some basic facts

While it is beyond the scope of this paper to present the algorithm cross in detail, we shall discuss some basic properties of RNA pseudoknot structures. The properties in question show that we can indeed assign a unique, loop-based energy to an RNA pseudoknot structure. In addition, we show that an RNA pseudoknot structure can be constructed via simpler substructures. The latter are in fact the building blocks via which cross constructs the mfe RNA pseudoknot structure.

2.1 Loops

We shall begin by introducing loops of 3-noncrossing RNA structures. Loops are not only the basic building block for the mfe-evaluation but also of importance for the coarse grained notion of pseudoknot-shapes, discussed in Subsection 3.2. Let $\prec$ denote the following partial order over the arcs (written as (i,j), $i<j$) of a k-noncrossing diagram

$$(i_1,j_2) \prec (i_2,j_2) \quad \Longleftrightarrow \quad i_2 < i_1 \ \wedge\ j_1 < j_2\ . \quad (2)$$

Let α be an arc in the 3-noncrossing RNA structure, S and denote by $A_S(\alpha)$ 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 $\alpha' \prec \alpha$.

Let $[i,j]$ denotes the sequence $(i, i+1, \dots, j-1, j)$. It is shown in [27] 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 Fig. **??**, 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 $h \geq 2$.

(4) a *pseudoknot*-loop, see Fig. **??**, consisting of the following data:
(P1) a set of arcs

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

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.

2.2 Decomposition

In this section we give a result of [27] which shows that each 3-noncrossing RNA structure can uniquely be constructed by simpler substructures. Furthermore we show that each 3-noncrossing RNA structure has a unique loop decomposition–the basis of our energy evaluation.

The building blocks of RNA pseudoknot structures are obtained in two steps. First one considers motifs and then one builds their shadows. In order to understand motifs we recall the notion of a core [?]. A k-noncrossing core is a k-noncrossing diagram in which all stacks have size one. The core of a structure S, denoted by $c(S)$, is obtained by identifying all S-stacks with a single arc, keeping the unpaired nucleotides and finally relabeling the diagram, see Fig. **??**. A $\langle k, \tau \rangle$-motif, m, is a $\langle k, \tau \rangle$-diagram over $[n]$, having the following properties

(M1) m has a nonnesting core

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

A m-shadow, denoted by $\overline{m}$, is a k-noncrossing diagram obtained by successively increasing the stacks of m from top to bottom.

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

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

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

In Fig. **??** we show how these decompositions work.

3 Minimum free energy RNA pseudoknot structures

In this section we give some statistics on pseudoknotted RNA structures as a function of the sequence length. In order to put our findings into context we consider the following two variants of cross: first, cross_3, which generates the 3-noncrossing, 3-canonical mfe structure and second, cross_4, which produces the 3-noncrossing, 4-canonical mfe structure.

3.1 The fraction of pseudoknots

In this section we compute the fraction of RNA structures with pseudoknot-loops within all structures for cross_3 and cross_4. Fig. **??** displays the fraction of structures with pseudoknots as a function of sequence length.

3.2 Pseudoknot-shapes

Next we study which are the dominant pseudoknot "shapes" as the sequence length n increases. for this purpose we introduce some suitable notion of shape based on the notion of k-noncrossing cores [?]. The shape of a structure S, is a subset of $c(S)$-arcs, induced by all arcs either contained in pseudoknot-loops or arcs contained in multi-loops which contain nested pseudoknot-loops. In other words, the pseudoknot-shape contains all pseudoknot arcs and all arcs affecting the energy of pseudoknot-loops, see Fig. **??**. In Fig. **??** we display for cross_3 and cross_4 the dominant types for increasing n.

3.3 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. In order to assure generic findings we consider the variants of cross_3 and cross_4.

Let us first discuss the distribution of stacks in random pseudoknot structures. The naive approach would be to 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, whence we have to employ a different strategy. The key idea [28] is to consider the bivariate generating function

$$\mathbf{T}_{k,\tau}(x,u) = \sum_{n \geq 0} \sum_{0 \leq t \leq \frac{n}{2}} \mathsf{T}_{k,\tau}(n,t) u^t x^n \qquad (3)$$

where $\mathsf{T}_{k,\tau}(n,t)$ denotes the number of k-noncrossing, τ-canonical pseudoknot structures having exactly t stacks. Interestingly $\mathbf{T}_{k,\tau}(x,u)$ can be computed using the cores introduced in Section 3.2. The stack-distribution is now given by

$$\mathsf{P}(X^n_{k,\tau} = t) = \mathsf{T}_{k,\tau}(n,t)/\mathsf{T}_{k,\tau}(n) \qquad (4)$$

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

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

$$\sigma^2_{k,\tau} \;=\; \left(\frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)}\right)^2 - \frac{\gamma''_{k,\tau}(0)}{\gamma_{k,\tau}(0)}. \qquad (6)$$

where $\gamma_{k,\tau}(u)$ is the unique dominant singularity parameterized by $u = e^s$. In Tab. 4.12 we display the values $\mu_{k,\tau}$ and $\sigma^2_{k,\tau}$ for $k = 2, 3, 4$ and $\tau = 3, \dots, 7$. In Fig. ?? we present the stack distributions of cross_k and cross_k^{+5}, respectively, obtained by folding 3000 structures by randomly selecting sequences. In addition we display also in Fig. ?? the stack distribution of random 3-noncrossing and 4-noncrossing structures obtained from Tab. 4.12

4 Neutrality and local connectivity

Tab. 4.12 contains nontrivial information about the mapping from RNA sequences into their k-noncrossing, canonical structures. To be precise, Tab. 4.12, in combination with central limit theorems for the number of arcs in k-noncrossing RNA structures [29, 30] allow us to conclude that there exist exponentially many k-noncrossing canonical structures with exponentially large preimages. Indeed, according to Tab. 4.12 the exponential growth rate of the number of k-noncrossing canonical structures, $3 \le k \le 9$ 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 show a mean of $0.4\,n$ and a variance of XXX. 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 [?, 31] and can be viewed as a basic protocol of local statistics of a gentotype-phenotype map. The only exception is Subsection 4.3, which elaborates on the novel concept of local connectivity [32]. Exhaustive computations of the set of all sequences over the natural alphabet with fixed pseudoknot structure for $n > 40$ is at present time impossible. In order to put the genericity of our results into context, we perform the analysis for cross_k where $k = 3, 4$ as well as cross_k^{+} (5% increased pseudoknot loop-penalty).

4.1 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

$$(\mathbf{A},\mathbf{U}), (\mathbf{U},\mathbf{A}), (\mathbf{G},\mathbf{U}), (\mathbf{U},\mathbf{G}), (\mathbf{G},\mathbf{C}), (\mathbf{C},\mathbf{G}).$$

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

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

where the u_j denote the unpaired nucleotides and the $p_j = (x_i, x_k)$ all base pairs, respectively, see Figure ??. 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 $C[S]$.

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 Figures ??. Note however, that a p-neighbor has either Hamming distance one $((\mathbf{G},\mathbf{C}) \mapsto (\mathbf{G},\mathbf{U}))$ or Hamming distance two $((\mathbf{G},\mathbf{C}) \mapsto (\mathbf{C},\mathbf{G}))$. We call a u- or a p-neighbor, y, a compatible neighbor. If y is contained in the neutral network we refer to y 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 Fig. ?? we give several data on neutral walks for $n = 76$:

4.2 Neutral neighbors

The neutral walk data of Subsection 4.1 are in accordance with the findings for RNA secondary structures. One can easily neutrally traverse sequence space, suggesting the picture of a vast connected network of neutral sequences. We can furthermore conclude that our findings are robust since they hold for all version of cross. The next question is to obtain data on the actual number of neutral neighbors during these walks, which we display in Fig. **??** for our reference structures.

4.3 Local connectivity

The connectivity of a network or subgraph, Γ_n, of an n-cube does not imply that a small Hamming distance of two of its vertices guarantees 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. Intuitively speaking, if Γ_n is locally connected then the small Hamming distance does imply a Γ_n-distance scaled by at most a factor of $\Delta > 0$. Local connectivity does naturally arise for random induced subgraphs of n-cubes, i.e. where we select sequences with independent probability λ_n. Then Γ_n is locally connected if and only if almost surely (a.s.)

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

provided v, v' are in Γ_n. We immediately observe that, trivially, for any *finite* n such a Δ exists. However, the key point is that (†) employs the notion "almost surely", i.e. it holds for arbitrary n. Random graph theory [32] shows that on the one hand, for λ_n smaller than $n^\delta/\sqrt{n}$, where $\delta > 0$ is arbitrarily small, there exists a.s. no finite Δ satisfying (†). On the other hand, for λ_n larger or equal than $n^\delta/\sqrt{n}$, there exists a.s. some finite Δ satisfying (†). Accordingly, there exists a threshold value for local connectivity.

Suppose we are given a structure S and sequence v, contained in its neutral network. Observing that local connectivity refers to the two n-cubes $Q_4^{n_u}$ and $Q_6^{n_p}$ induced by S, see Fig. **??**, we consider the set of sequences in compatible distance two, $C_2 = |\{v' \mid C(v,v') = 2\}|$. We the proceed setting

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

and call $D_S(v)$ the degree of local connectivity of S at v. 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.

It is apparent that local connectivity is vital for molecular evolution and any type of evolutionary optimization. It has been shown in [32] that local connectivity is a prerequisite for preserving any type of sequence specific information. We perform the following experiment: along a neutral walk, see Subsection 4.1, we compute $D_S(v_i)$ and in Fig. **??** we display the distribuation of D_S of XXX neutral walks.

Conclusions

Acknowledgements

The thank absolutely noone.

References

1. Penner RC, Waterman MS: **Spaces of RNA secondary structures**. *Adv Math* 1993, **101**:31–49.
2. Waterman MS: **Combinatorics of RNA hairpins and cloverleaves**. *Stud Appl Math* 1979, **60**:91–96.
3. Smith TF, Waterman MS: **RNA secondary structure**. *Math Biol* 1978, **42**:31–49.
4. Schmitt WR, Waterman MS: **Linear trees and RNA secondary structure**. *Discr Appl Math* 1994, **51**:317–323.
5. Howell JA, Smith TF, Waterman MS: **Computation of generating functions for biological molecules**. *J Appl Math* 1980, **39**:119–133.
6. Searls DB: **The language of genes**. *Nature* 2002, **420**:211–217.
7. Loria A, Pan T: **Domain Structure of the ribozyme from eubacterial ribonuclease**. *RNA* 1996, **2**:551–563.
8. Konings DAM, Gutell RR: **A comparison of thermodynamic foldings with comparatively derived structures of 16s and 16s-like rRNAs**. *RNA* 1995, **1**:559–574.
9. 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.
10. 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.
11. Lyngsø RB, Pedersen CNS: **RNA Pseudoknot Prediction in Energy-Based Models**. *J Comp Biol* 2000, **7**:409–427.
12. Rivas S E Eddy: **A dynamic programming algorithm for RNA structure prediction including pseudoknots**. *J Mol Biol* 1999, **285(5)**:2053–2068.

13. Akutsu T: **Dynamic programming algorithms for RNA secondary prediction with pseudoknots**. *Discr Appl Math* 2000, **104**:45–62.

14. Jin EY, Qin J, Reidys CM: **Combinatorics of RNA structures with Pseudoknots**. *Bull Math Biol* 2008, **70(1)**:45–67.

15. Waterman MS, Smith TF: **Rapid dynamic programming methods for RNA secondary structure**. *Adv Appl Math* 1986, **7**:455–464.

16. 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.

17. Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information**. *Nucl Acids Res* 1981, **9**:133–148.

18. Nussinov R, Jacobson AB: **Fast Algorithm for Predicting the Secondary Structure of Single-Stranded RNA**. *Proc Natl Acad Sci, USA* 1980, **77**:6309–6313.

19. Reidys CM, Stadler PF, Schuster P: **Generic properties of combinatory maps: neutral networks of RNA secondary structures**. *Bull Math Biol* 1997, **59(2)**:339–397.

20. Schultes EA, Bartel DP: **One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds**. *Science* 2000, **289. no. 5478**:448 – 452.

21. Jolly A (Ed): **Mapping RNA form and function**. In *Science* 2005, **309**:1441–1632.

22. Ma G, Reidys CM: **Canonical RNA Pseudoknot Structures**. *J Comp Biol*, in press.

23. 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.

24. Chen WYC, Qin J, Reidys CM: **Combinatorics of k-noncrossing Tangled-diagram**. *Elec J Comb*, in press.

25. Hofacker IL, Schuster P, Stadler PF: **Combinatorics of RNA Secondary Structures**. *Discr Appl Math* 1998, **88**:207–237.

26. Chen WYC, Qin J, Reidys CM: **Crossing and Nesting of Tangled-diagrams**. *Elec J Comb* 2008, **15**.

27. Huang FWD, Peng WWP, Reidys CM: **Folding RNA pseudoknot structures**. [In preparing].

28. Han HSW, Reidys CM: **Stacks in canonical RNA pseudoknot structures**. *Comp Appl Math*, in press.

29. Jin EY, Reidys CM: **Central and Local Limit Theorems for RNA Structures**. *J Theo Biol* 2008, **250(3)**:547–559.

30. Huang FWD, Reidys CM: **Statistics of canonical RNA pseudoknot structures**. *J Theor Biol*, in press.

31. Fontana W, Schuster P: **Shaping Space: the Possible and the Attainable in RNA Genotype-Phenotype Mapping**. *J Theo Biol* 1998, **194**(4):491–515.

32. Reidys CM: **Local Connectivity of Neutral Networks**. *Bull Math Biol*, in press.

Figures

4.4 Figure 3–a 3-noncrossing pseudoknot structure

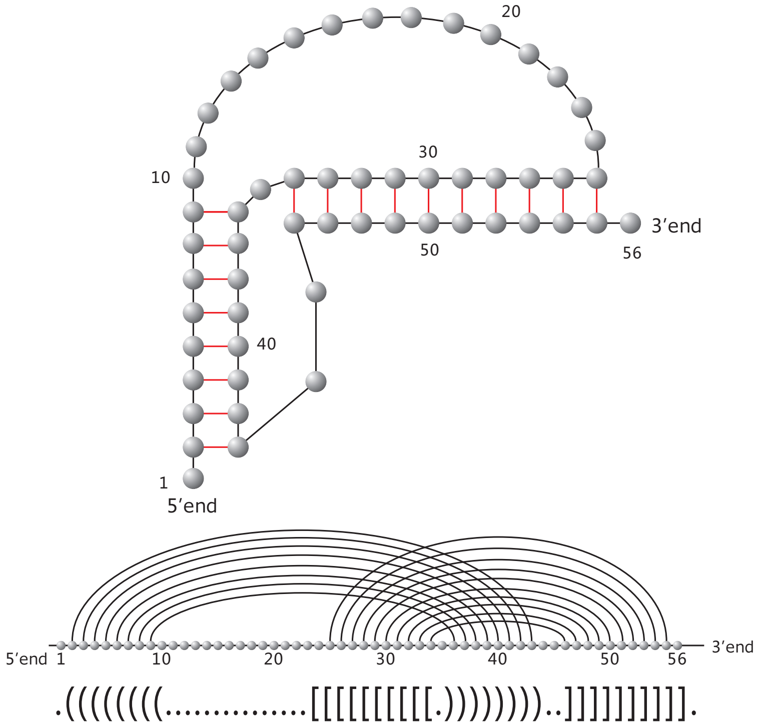


Figure 1: Pseudoknot structure.

Figure 1–k-noncrossing diagrams

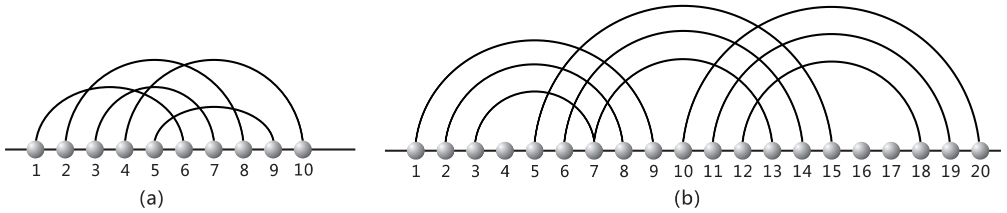


Figure 2: k-noncrossing diagrams: we display a 4-noncrossing, arc-length $\lambda \geq 4$ and $\sigma \geq 1$ (upper) and 3-noncrossing, $\lambda \geq 4$ and $\sigma \geq 2$ (lower) diagram.

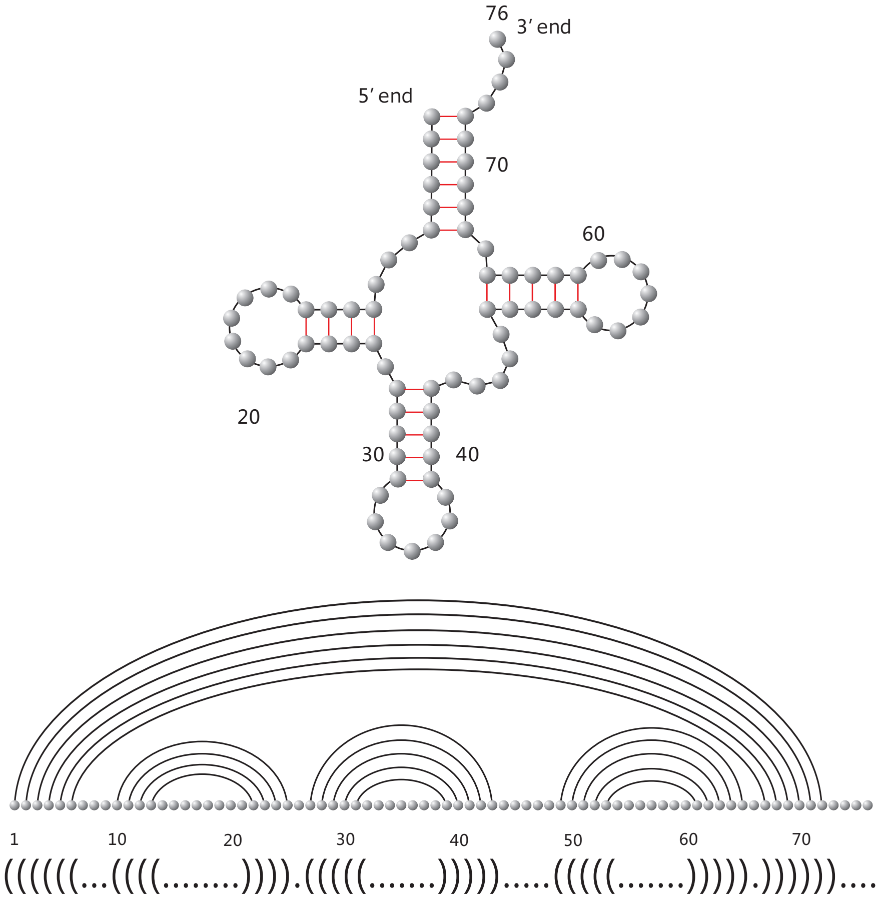


Figure 3: Secondary structure

4.5 Figure 2–a secondary structure

4.6 Figure 4–standard loops

4.7 Figure 5–pseudoknot loops

4.8 Figure 6–Cores

4.9 Figure 7–decomposition

4.10 Figure 7–PK-shapes

4.11 Figure 8–stack distribution

4.12 Figure 11–neutral neighbors

Tables

Table 1 - Exponential growth rates of k-noncrossing, τ-canonical RNA structures

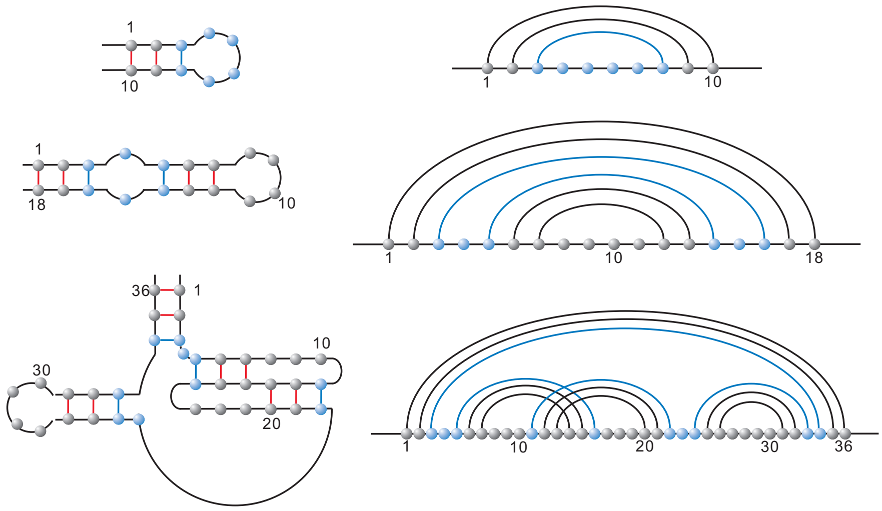


Figure 4: standard loop

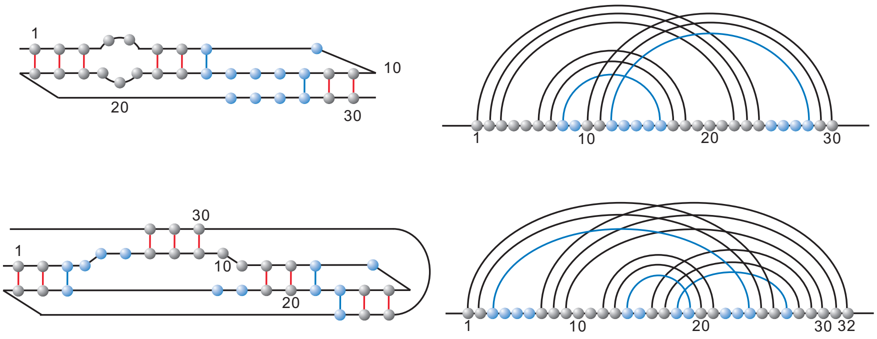


Figure 5: pseudoknot loop

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

Exponential growth rates of $\langle k, 4, \sigma\rangle$-structures where $\sigma \geq 3$.

Table 2 - Mean and variance of the number of stacks in pseudoknot RNA

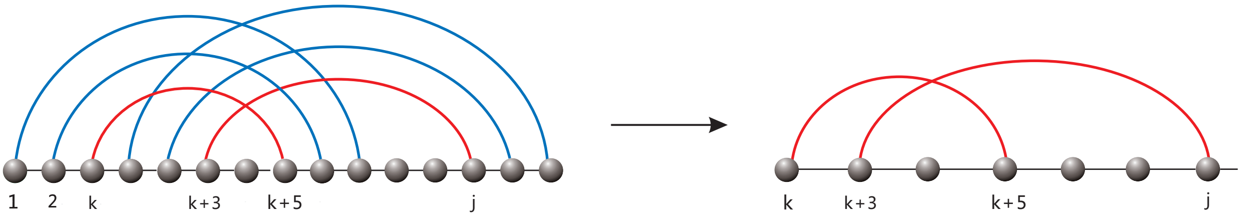


Figure 6: Cores

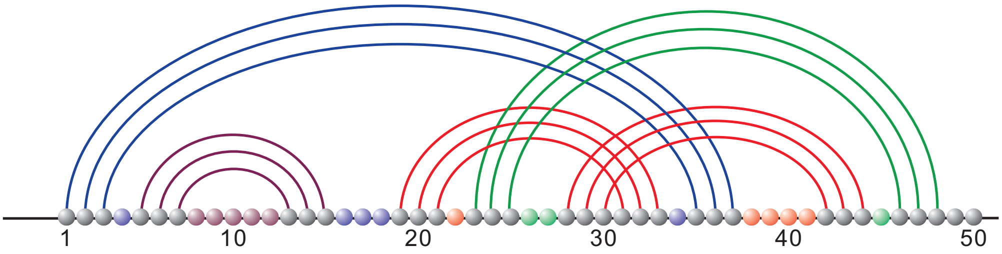


Figure 7: decomposiiton

	$k=2$		$k=3$		$k=4$	
	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

Normal limit distributions of the random variable $X^n_{k,\tau}$, for different k and τ. We list mean (μ) and variance (σ^2).

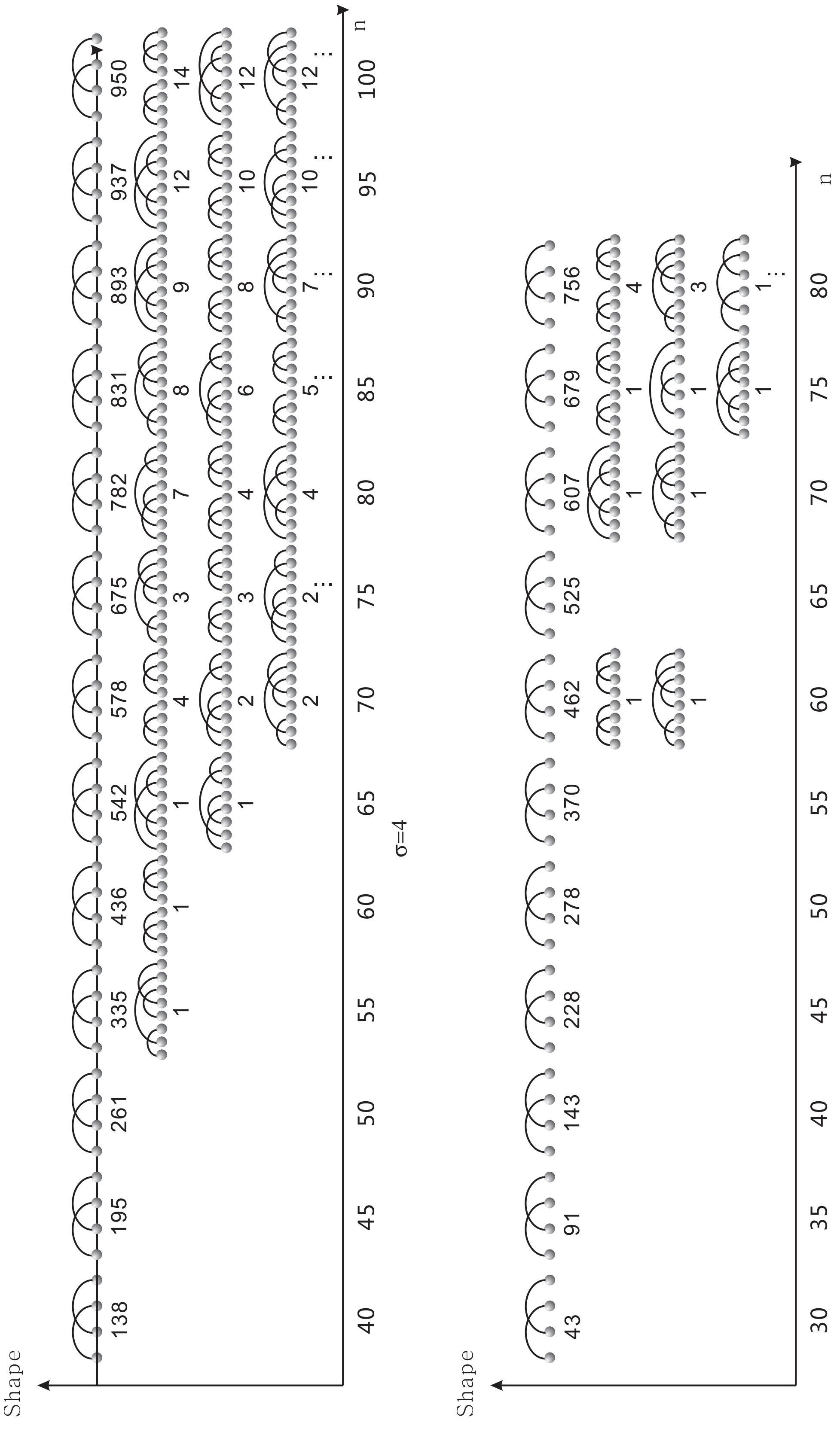
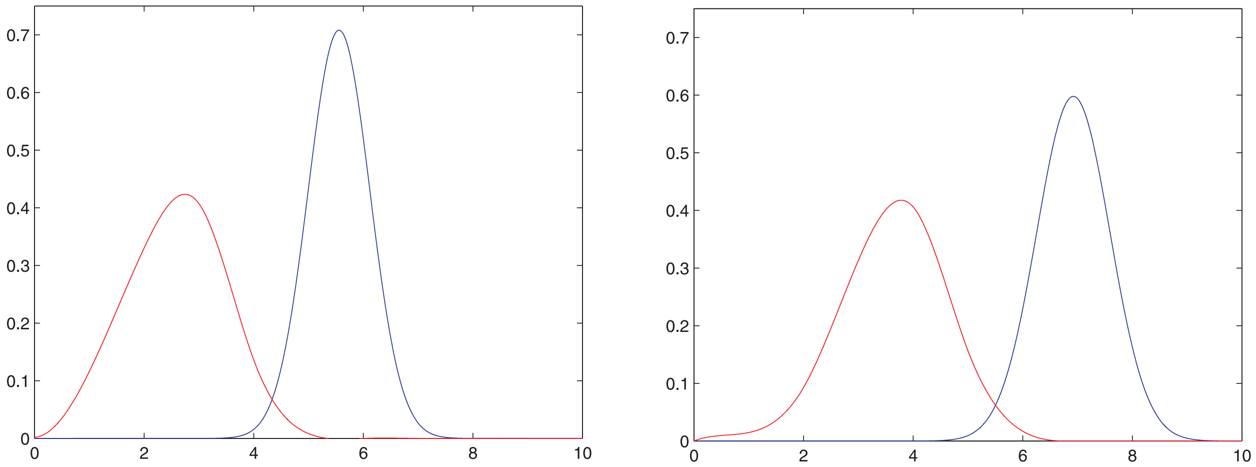


Figure 8: shape



$\sigma = 4$

$\sigma = 5$

Figure 9: The red curve is experiment, and blue curve is from theory.

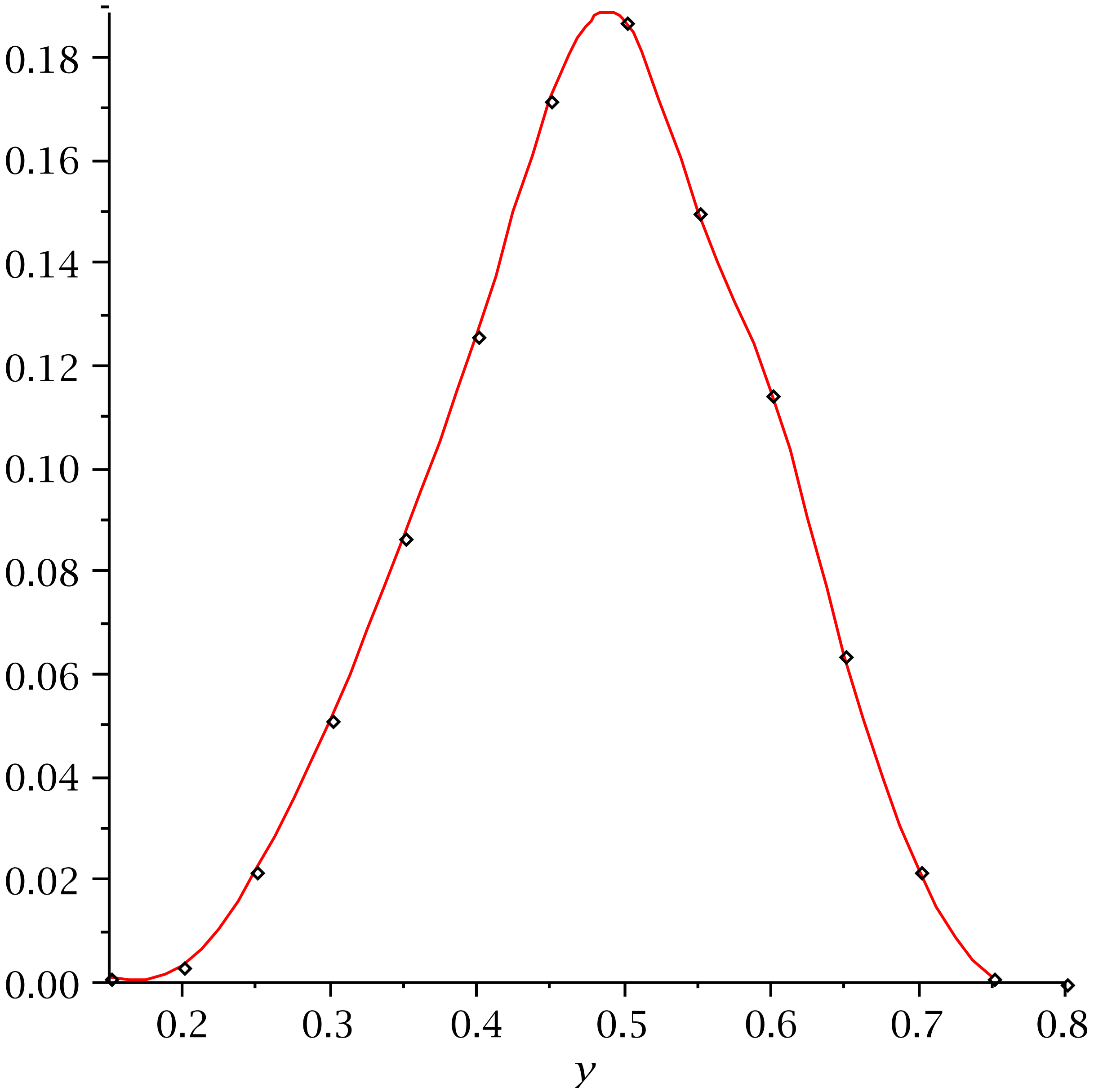


Figure 10: Neutral fraction distribution of sequences random sequences

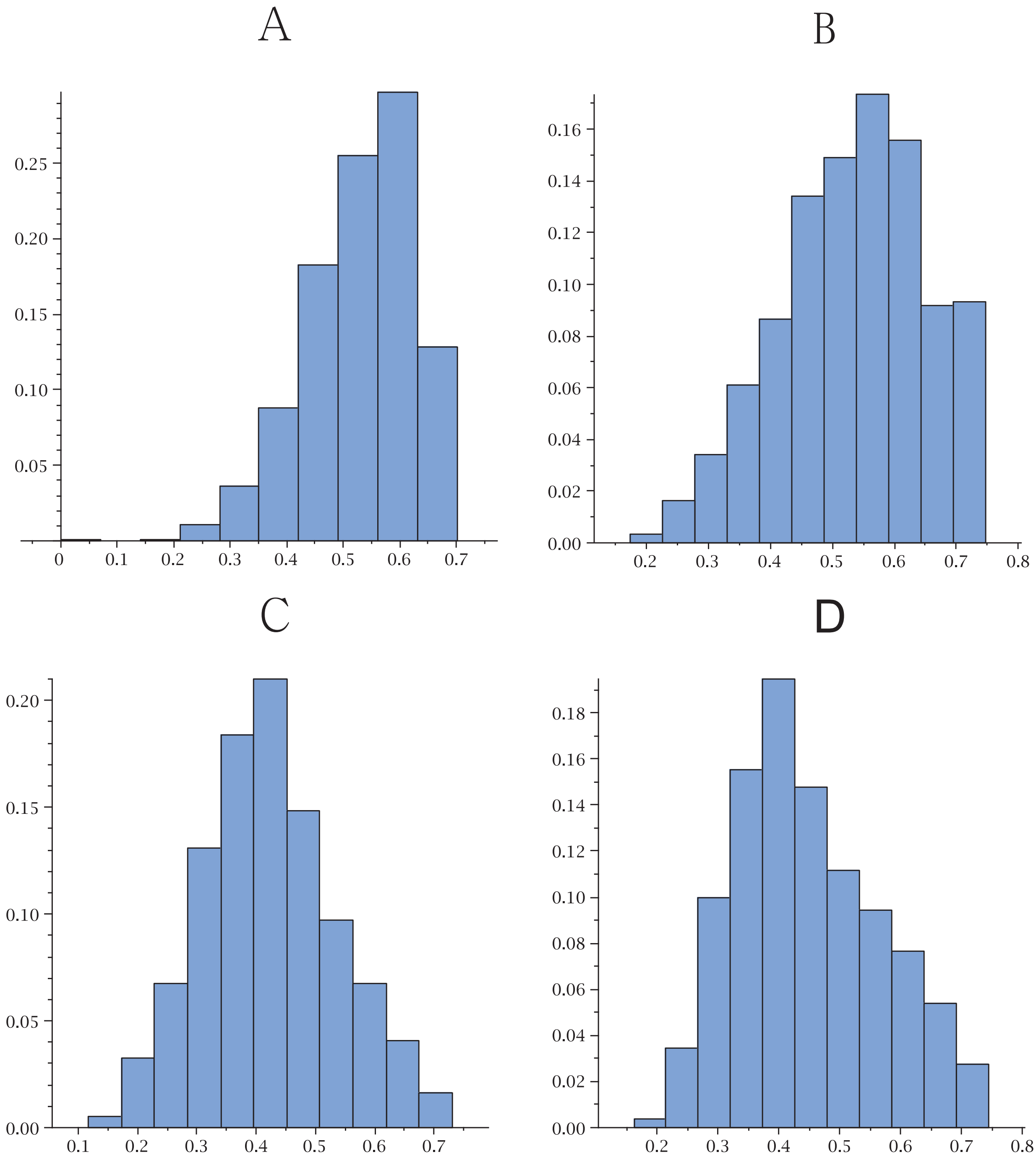


Figure 11: A. Neutral fraction distribution of sequences in tRNA path:B. Neutral fraction distribution of sequences in a hairpin sequence path: C. Neutral fraction distribution of sequences in a interior sequence path: D. Neutral fraction distribution of sequences in a hloop sequence path

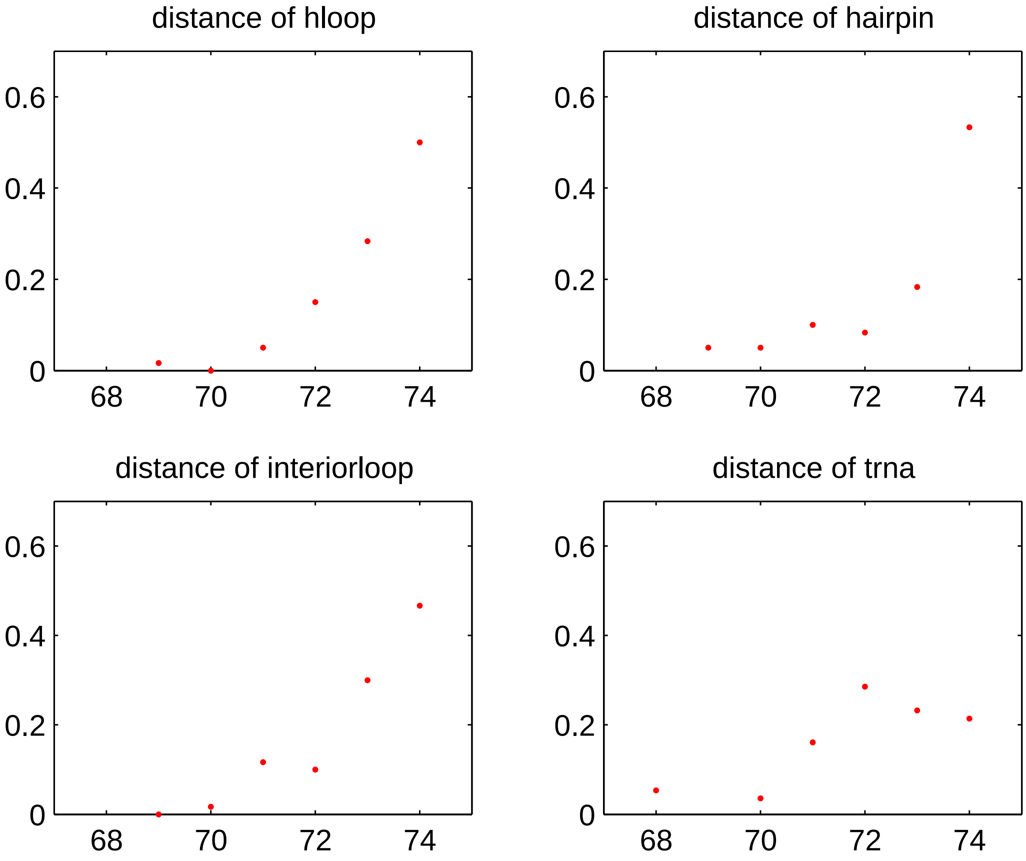


Figure 12: Distance distribution of

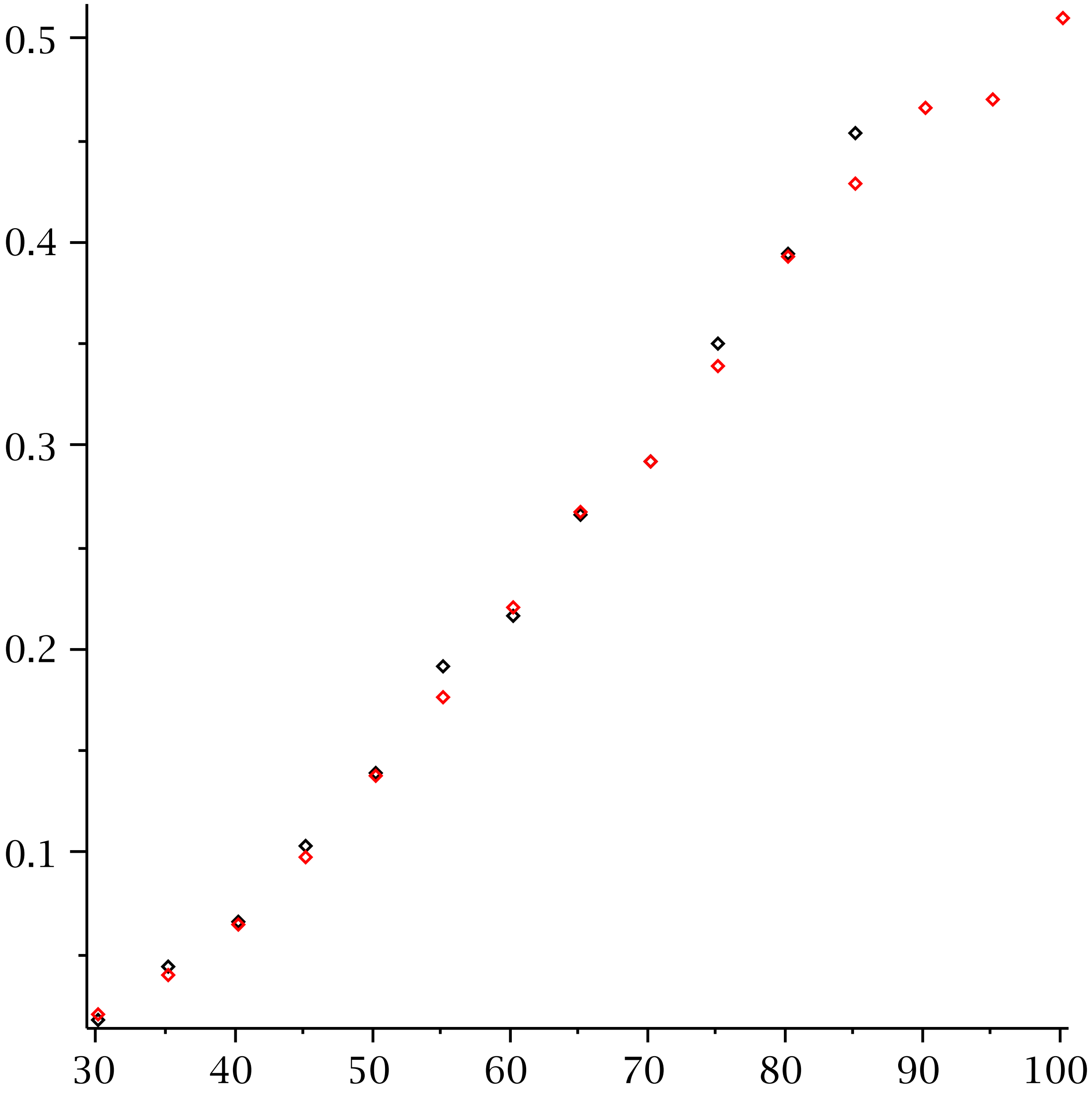


Figure 13: Black.Peudoknot fraction distribution of $\sigma = 3$ Red.Peudoknot fraction distribution of $\sigma = 4$

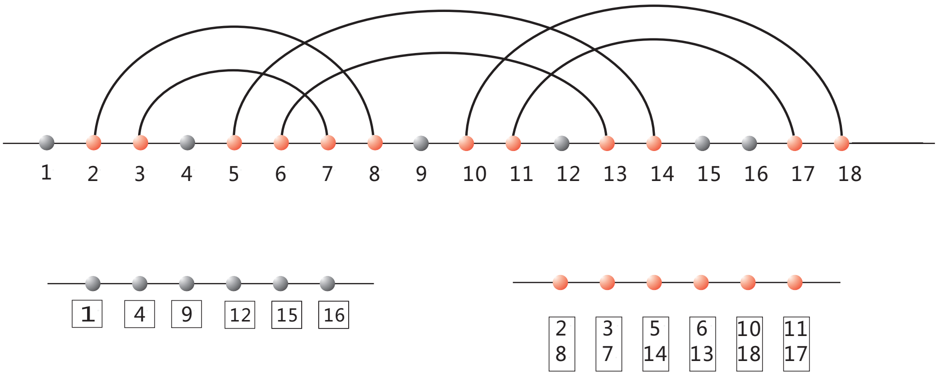


Figure 14: Deriving the two subcubes $Q_4^{n_u}$ and $Q_6^{n_p}$: a structure gives rise to rearrange its compatible sequences into unpaired and paired segment. The former is a sequence over the original alphabet $\mathbf{A}$, $\mathbf{U}$, $\mathbf{G}$, $\mathbf{C}$ and for the latter we derive a sequence over the alphabet of base pairs, $(\mathbf{A},\mathbf{U})$, $(\mathbf{U},\mathbf{A})$, $(\mathbf{G},\mathbf{U})$, $(\mathbf{U},\mathbf{G})$, $(\mathbf{G},\mathbf{C})$, $(\mathbf{C},\mathbf{G})$.

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 Hammimg distance one.