5+ model YJTBI : 4848 Prod.Type:FTP pp.1–14(col.fig.:NIL) ED:AnupamaDeviY PAGN:Bhaskara SCAN:

ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Journal of Theoretical Biology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

Neutral networks of sequence to shape maps

Emma Y. Jin, Jing Qin, Christian M. Reidys*

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

Received 6 June 2007; received in revised form 10 September 2007; accepted 10 September 2007

Abstract

In this paper we present a combinatorial model of sequence to shape maps. Our particular construction arises in the context of representing nucleotide interactions beyond Watson–Crick base pairs and its key feature is to replace biophysical steric by combinatorial constraints. We show that these combinatory maps produce exponentially many shapes and induce sets of sequences which contain extended connected subgraphs of diameter n, where n denotes the length of the sequence. Our main result is to prove the existence of exponentially many shapes that have neutral networks.

Keywords: Combinatory map; Component; Diameter; Neutral network; Shape; Bipartite

1. Introduction

1.1. Background

Q1 Arguably one of the greatest challenges in present day biophysics is the understanding of sequence structure relations of biopolymers. For one particular class of biopolymers, the ribonucleic acid (RNA) secondary structures, (Fig. 1) molecular folding maps have been systematically analyzed by Schuster et al. (Fontana and Schuster, 1998; Schuster et al., 1994; Schuster, 2002). Folding maps play a central role in understanding the evolution of molecular sequences. Specific properties like, for instance *shape space covering* (Schuster et al., 1996) and *neutral networks* (Fig. 2) (Reidys et al., 1997) are critical for what may be paraphrased as "molecular computation by white noise". For instance, neutral networks played a central role in the *Science* publication authored by E. Schultes and P. Bartels *One sequence, two ribozymes*: *implications for the emergence of new ribozyme folds*, (v289, n5478, 448–452) where the authors designed experimentally a single RNA sequence (whose existence is implied by the intersection theorem in Reidys et al., 1997) that folds into two different, non-related, RNA secondary structures (Clote et al., 2005). Exhaustive enumeration of sequence spaces and subsequent detailed analysis of the mappings for **G**,**C**-sequences of length 30 were undertaken in Grüner et al. (1996a, b). In addition detailed analysis of neutral networks as well as exhaustive enumeration of **G**,**C**,**A**,**U**-sequences can be found in Göbel (2000). The findings were intriguing. Folding maps into RNA secondary structures exhibit a collection of distinct properties which makes them ideally suited for evolutionary optimization.

(a) Many structures have preimages of sequences (neutral networks) which have large components and large diameter, i.e. which percolate sequence space.
(b) Many structures have the property that any two of them have neutral networks that come close in sequence space.

*Corresponding author. Tel.: +1 86 22 2350 5133 8013; fax: +1 86 22 2350 9272.
E-mail address: reidys@nankai.edu.cn (C.M. Reidys).

0022-5193/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2007.09.012

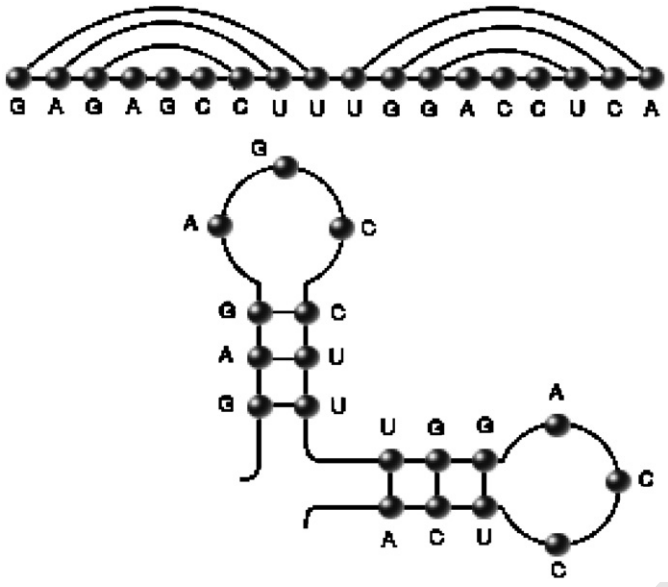


Fig. 1. RNA secondary structures. Diagram representation (top): the primary sequence, **GAGAGCCUUUGGACCUCA**, is drawn horizontally and its backbone bonds are ignored. All bonds are drawn in the upper half plane and secondary structures have the property that no two arcs intersect and all arcs have minimum length 2. Outer planar graph representation (bottom).

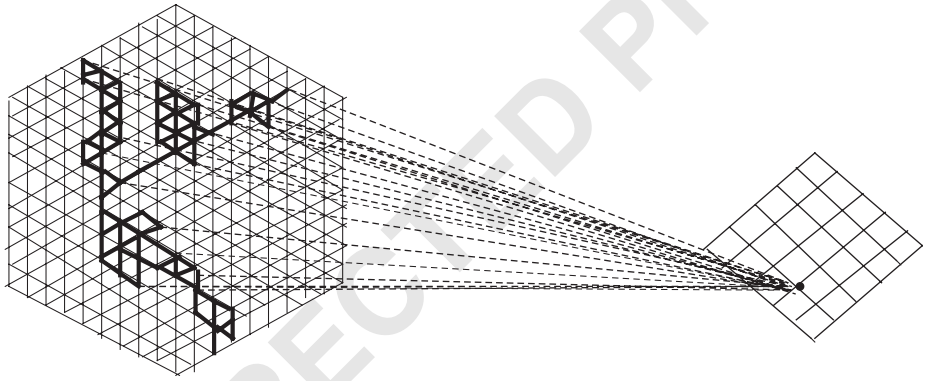


Fig. 2. The neutral network of a structure. Sequence space (right) and shape space (left) represented as lattices, respectively. We draw the edges between two sequences bold if they map into the one particular structure on the left. The two key properties of neutral nets are their connectivity and percolation. They allow sequences to move while maintaining a shape through sequence space.

Obviously, (a) is of central importance in the context of neutral evolution. Since replication is ~~erroneous~~ and only few if not single nucleotides can be exchanged the preimages of structures must contain large connected components. (b) showed that (many) new structures can easily be found during a random walk on a neutral network using steps only in which a single nucleotide is altered (point mutations).

Folding maps, however, are not obtained analytically. They are results of a computer algorithm, based on the combinatorial analysis of RNA secondary structures pioneered by Waterman et al. (Schmitt and Waterman, 1994; Waterman, 1978, 1979). It has to be remarked in this context that comparative sequence analysis (Woese and Pace, 1993; Puglisi and Williamson, 1999) provides more reliable means for determining the secondary structure of biological RNA (Batey et al., 1999), i.e. folding maps represent already an abstraction. In order to step beyond the secondary structure paradigm two main approaches with distinct goals are: (1) to study more advanced nucleotide interactions in RNA, like for instance pseudoknots, base triples or (2) consider genuine abstractions of molecular structures not aiming to model a biophysical folding map. In Jin et al. (2007) we pursue the first by developing the combinatorics of RNA structures with pseudoknots and in this contribution the second by studying combinatory maps. While (1) eventually produces the mathematical framework enabling us to derive more advanced representations (which eventually result in folding algorithms capable of producing structures like phenylalanine tRNA) (2) provides insights on the core question of which principles produce sequence to structure maps suitable for evolution. A type (2) abstraction inevitably evokes skepticism

Please cite this article as: Jin, E.Y., et al., Neutral networks of sequence to shape maps. J. Theor. Biol. (2007), doi:10.1016/j.jtbi.2007.09.012

since what can possibly be gained if no attempt is made to mimic the biological reality? However, we argue that sometimes it is exactly the right strategy to fundamentally understand the object under investigation.

1.2. Structures and correlations

A well-studied class of maps over sequence spaces are the *NK*-landscapes introduced by Kauffman (1993), where each index (locus) of a binary n-tuple viewed as the genotype composed by n loci is randomly linked to K other indices. The idea is that a locus i makes a contribution to the total fitness of the genotype which depends on the value of the allele (0 or 1) at i and the values at each of the epistatically linked loci. To each of those 2^{K+1} combinations there is a value (fitness) assigned uniformly at random. The apparent lack of neutrality led Barnett (1998) to refine *NK*-landscapes by *NKp*-landscapes, introducing a probability p with which an arbitrarily chosen allelic combination makes no contribution to the fitness. Our approach is connected to Kauffmann's intuition in that we consider a molecular structure as a combinatorial representation of nucleotide-correlations. As for nucleotide-correlations observations (a) and (b) are not bound to the particular concept of RNA secondary structures. For instance Stadler and Schuster (1999) as well as Bastolla et al. (2003) have shown that neutral networks exist for proteins, where nucleotide interactions are much more involved (Reidys, 2000). Therefore it is certainly not the uniqueness of Watson–Crick base pairings implying the existence of neutral networks. Our particular approach comes from this correlation perspective and observations from molecular interaction in RNA molecules. First there are secondary and tertiary interactions (Batey et al., 1999), the latter typically involving secondary structural elements. Furthermore interaction within RNA molecules can be categorized into three classes, helix–helix interaction, loop/bulge–helix and loop–loop interaction (Westhof and Jaeger, 1992; Batey et al., 1999). The structure of phenylalanine tRNA, and the hammerhead ribozyme (Wedekind and McKay, 1998) have served as paradigms in this context. Base triples and tetra-loops, as well as pseudoknots (Westhof and Jaeger, 1992; Konings and Gutell, 1995; Chamorro et al., 1991; Mapping RNA form and function, 2005), representing loop–loop interactions have led to generalizations of the secondary structure concept. These interactions are subject to steric constraints arising from the biochemistry of the interactions involved. These observations give rise to two different combinatorial abstractions: the consideration of k-noncrossing (where k is a natural number greater than 1) chemical bonds, which means that there exists no k-set of mutually intersecting bonds and of 2-diagrams (Chen et al., 2007) i.e. graphs whose vertices are drawn as a horizontal line having degree $\leqslant 2$ (and the combination of them, k-noncrossing 2-diagrams). The notion of k-noncrossing arises naturally in the context of pseudoknots leading to the concepts of k-noncrossing RNA structures (Jin et al., 2007) and to Stadler's bi-secondary structures (Haslinger and Stadler, 1999) (which are exactly the planar 3-noncrossing RNA structures). The notion of 2-diagrams allows to cast multiple nucleotide interactions and the expression or interactions of secondary structure elements (Fig. 3).

2. The basic construction

The notion of 2-diagrams discussed in the introduction is exactly the motivation of our particular approach. In the following we detail how to derive molecular shapes in which each nucleotide has at most two interactions but which,

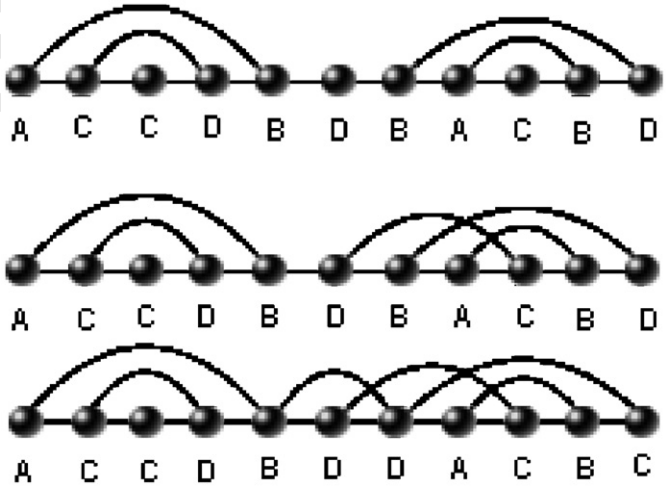


Fig. 3. Beyond secondary structures. Suppose we are given an abtract alphabet {**A**, **B**, **C**, **D**} with base pairs {{**A**, **B**}, {**D**, **C**}, {**D**, **B**}}. We present diagram representations of a secondary structure (top), 3-noncrossing structure (middle) and a 2-diagram structure (bottom). The difference between the first two structures is the crossing of bonds and the difference between the second two is the number of interactions for a nucleotide.

Please cite this article as: Jin, E.Y., et al., Neutral networks of sequence to shape maps. J. Theor. Biol. (2007), doi:10.1016/j.jtbi.2007.09.012

difference to biophysical structures, have combinatorial constraints imposed on their nucleotide interactions. This idea is to the best of our knowledge new. For a given alphabet base pairing rules specify which nucleotides can pair. However, not any two nucleotides are able to establish a bond. For instance, they may be restricted by conditions like no two edges can cross each other when representing a shape as a diagram (Haslinger and Stadler, 1999). The noncrossing condition and uniqueness of base pairs are two key properties of RNA secondary structures and allow for Motzkin-path enumeration and tree bijections (Schmitt and Waterman, 1994; Waterman, 1979; Zuker and Sankoff, 1984; Waterman, 1978; Hofacker et al., 1998). We replace these restrictions on nucleotide interactions by stipulating that (a) there exists some base graph H whose sole function is to restrict all possible correlations and (b) we are given a symmetric relation $\mathscr{R}$, tantamount to a base pairing rule. In order to avoid any confusion we work over the abstract alphabet $\{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}\}$.

In this framework a shape $\mathscr{S}$ of a sequence is then the unique maximal H-subgraph subject to the property that for any $\mathscr{S}$-edge the incident nucleotides satisfy $\mathscr{R}$. We denote a combinatory map assigning to a sequence a H-subgraph by ϑ_H. It is remarkable that this simple definition already produces a well-defined sequence to structure map. Moreover this definition is in line with the biological point of view: mapping sequences into shapes rather than fixing some shape and then to consider its sequences. It now can be asked what the right choice of H should be and how generic the respective conclusions are. As for dependency on H the answer is that it a.s. (almost surely in the sense of random graph theory, i.e. in the limit of long sequences) depends on the number of edges, only. Therefore, the choice of $H=\mathscr{H}$ is not critical for the validity of the main results. To understand why, we consider a generalization of the concept of combinatory maps, i.e. combinatory maps induced the random graph $G_{n,p}$ (the random graph over n vertices in which each edge is selected with independent probability p). In the sub-critical phase these random combinatory maps a.s. produce, modulo constants, all properties of the maps induced by $\mathscr{H}$ (Theorem 2) (Fig. 4).

Theorem (*Reidys, 2007, Neutral networks*). *Let $p_n=(1-\varepsilon)/n$, $\beta<\sqrt{2}$ and suppose ω_n is an unbounded function of n that grows arbitrarily slowly and $\vartheta_{G_{n,p}}$ is a random combinatory map. Then there exists asymptotically surely at least β^n shapes $\mathscr{S}$ with the following two properties*:

(I) *the set of all sequences mapping into $\mathscr{S}$ has a connected component of size at least $(\sqrt{2})^n$;*
(II) *the set of all sequences mapping into $\mathscr{S}$ percolates, i.e. has diameter $n-\omega_n$.*

The great advantage of choosing $H=\mathscr{H}$ is the simplicity and the resulting algorithmic nature of all proofs. We can explicitly construct all paths involved by diagram chasing. In contrast, the proof of the above result is based on a nontrivial analysis of tree components in the random graph $G_{n,p}$.

Having justified our approach let us introduce our framework: let H be a graph over $\{1,2,\ldots,n\}$, $\mathscr{A}=\{\mathbf{A},\mathbf{B},\mathbf{D},\mathbf{C}\}$ and Q_4^n be the generalized n-cube, Q_4^n, i.e. the graph over the sequences $(x_1,\ldots,x_n)$, where $x_i\in\mathscr{A}$ and in which two sequences are adjacent if they differ in exactly one nucleotide. Let $d(v,v')$ be the number of nucleotides by which v and v' differ. A component of a graph H is a maximal connected subgraph. We consider relations $\mathscr{R}$ over the abstract alphabet $\mathscr{A}=\{\mathbf{A},\mathbf{B},\mathbf{D},\mathbf{C}\}$, i.e. $\mathscr{R}\subset\mathscr{A}\times\mathscr{A}$ satisfying the following three conditions:

$$(x,y)\in\mathscr{R}\ \Leftrightarrow\ (y,x)\in\mathscr{R}, \tag{2.1}$$

$$(x,y)\in\mathscr{R}\Rightarrow x\neq y, \tag{2.2}$$

$$\forall x\neq z\quad (x,y)\in\mathscr{R}\wedge(y,z)\in\mathscr{R}\Rightarrow(x,z)\notin\mathscr{R}. \tag{2.3}$$

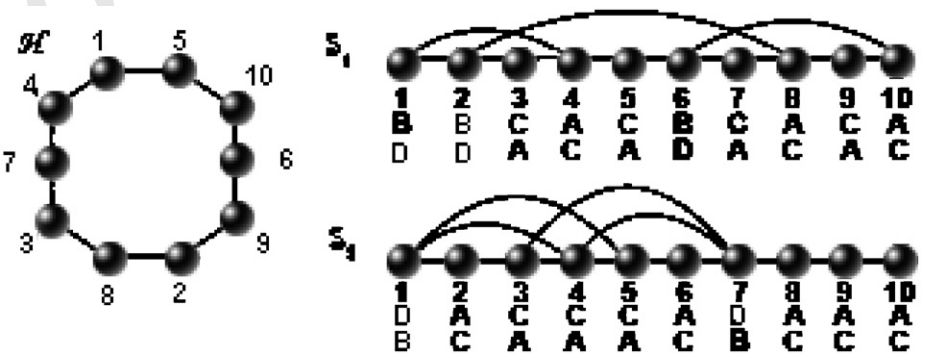


Fig. 4. Combinatory maps: the base graph $\mathscr{H}$ is displayed on the l.h.s.. Here $\mathscr{H}$ is an even length cycle (with an additional isolated point) when n is even (odd). The r.h.s. shows two shapes $\mathscr{S}_1$ and $\mathscr{S}_2$ with two particular sequences that are contained in their respective preimages. For both sequences the shapes are maximal, i.e. not a single $\mathscr{H}$-edge can be drawn without violating base pairing rules, here $\{\{\mathbf{A},\mathbf{B}\},\{\mathbf{D},\mathbf{C}\},\{\mathbf{D},\mathbf{B}\}\}$.

Please cite this article as: Jin, E.Y., et al., Neutral networks of sequence to shape maps. J. Theor. Biol. (2007), doi:10.1016/j.jtbi.2007.09.012

These conditions are motivated from abstracting form 2D and 3D interactions of the phenylalanine tRNA and the hammerhead ribozyme (Batey et al., 1999). In both molecules mutual interactions of 3-nucleotides are absent but multiple pair interactions are responsible for the tertiary structure. In view of Eqs. (2.1) and (2.2) each relation can be viewed as a graph over $\{\mathbf{A},\mathbf{B},\mathbf{D},\mathbf{C}\}$ and obviously, Eq. (2.3) is equivalent to this graph being bipartite.[1] A bipartite graph is a graph whose vertex set can be decomposed into two disjoint sets A, B such that no two vertices within A and B are connected, respectively. We will be particularly interested in the base pairing rule $\mathscr{R}^{\dagger}$ represented as the graph A —— B —— D —— C i.e. we allow for the following interactions: $\{\{\mathbf{A},\mathbf{B}\},\{\mathbf{D},\mathbf{C}\},\{\mathbf{D},\mathbf{B}\}\}$. In this sense our nucleotide interactions are more general than those of RNA secondary structures since, for instance, we can express coaxial stacking of helical regions and the formation of isosteric $\mathbf{C}\cdot\mathbf{G}-\mathbf{C}$ triples (Batey et al., 1999). We introduce the H-subgraph $H_{\mathscr{R}}(v)$ having vertex and edge set given by

$$V_{H_{\mathscr{R}}(v)}=\{1,\ldots,n\}$$

and

$$E_{H_{\mathscr{R}}(v)}=\{\{i,k\}\mid\{i,k\}\text{ is an }H\text{-edge and }(x_i,x_k)\in\mathscr{R}\} \tag{2.4}$$

and call $H_{\mathscr{R}}(v)$ a shape $\mathscr{S}$ and the mapping $\vartheta_H: Q_4^n\longrightarrow\{\mathscr{S}\mid\mathscr{S}=H_{\mathscr{R}}(v)\}$ a combinatory map. We will write sometimes $H(v)$ instead of $H_{\mathscr{R}}(v)$. Note that the above construction entails an implicit notion of maximality, i.e. a shape of a sequence $(x_1,\ldots,x_n)$ is the maximal H-subgraph which satisfies $\mathscr{R}^{\dagger}$ for all 2-sets of coordinates $\{x_i,x_j\}$, $\{i,j\}$ being a H-edge. In this sense a shape represents a saturated structure. As for $\mathscr{H}$, suppose first n is even. We set $C_n(1)$ to be the graph over $\{1,\ldots,n\}$ with edge set $\{i,i+1\}$ where the vertices are labeled modulo n. Let σ_n be some permutation of n-letters, we then set $C_n(\sigma_n)$ with edges $\{\sigma_n(i),\sigma_n(i+1)\}$ and $\mathscr{H}=C_n(\sigma_n)$. Next assume n is odd. Then we select an arbitrary element of $\{1,\ldots,n\}$, say u and define $\mathscr{H}=C_{n-1}(\sigma_{n-1})\cup\{u\}$ i.e. the graph with edges $\{\sigma_{n-1}(i),\sigma_{n-1}(i+1)\}$ for $i\neq u$ and $i+1\neq u$, where σ_{n-1} is an arbitrary permutation of$\{1,\ldots,n\}\backslash\{u\}$. To summarize we have

$$\mathscr{H}=\begin{cases}C_n(\sigma_n) & \text{for } n \text{ even},\\ C_{n-1}(\sigma_{n-1})\cup\{u\} & \text{for } n \text{ odd}.\end{cases} \tag{2.5}$$

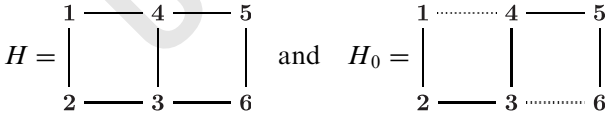
3. Shapes

In this section we answer the following basic questions:

(1) What is the relation between base pairing rules and the resulting molecular shapes?
(2) How many shapes does a combinatory map have?
(3) Are there "many" shapes with large sets of sequences folding into them?

All of the above properties are central for RNA secondary structures and none of them can be answered analytically, despite the fact that we have generating functions for RNA secondary structures. For instance, it is impossible to assess *a priori* how many secondary structures have an actual sequence folding into them. The number of RNA structures that actually occur as minimum free energy structures can be much smaller than the total number. For $n=16$, due to finite size effects for the RNA folding, only 63% of the possible RNA structures are realized as minimum free energy structures (Göbel, 2000).

Let us begin by providing some more background: graph H' is called an induced subgraph of H iff there exists some set $M\subset\{1,\ldots,n\}$ such that $E_{H'}=\{\{i,j\}\mid\{i,j\}\in E_H\wedge i,j\in M\}$. Intuitively, induced subgraphs come from vertex sets and are far more restricted that arbitrary subgraphs. We now give a simple example of the fact that not every bipartite subgraph of a shape is a shape. For instance, consider $\vartheta_H: Q_4^6\longrightarrow\{H'<H\}$ where



(3.1)

where the dotted lines represent missing edges. Clearly, H is bipartite and it is easy to check that indeed $H=H(\mathbf{D},\mathbf{C},\mathbf{D},\mathbf{C},\mathbf{D},\mathbf{C})$ holds. Therefore H is a shape but H_0 is not. Every sequence realizing H_0 has necessarily either **A** at **1**, and **C** at **4** or vice versa. In the first case **D** is necessarily at **3** and **5**, which leaves no valid choice for **6**. The second case follows analogously.

[1]For instance, it is easy to check that the relation implied by all Watson–Crick base pairs (i.e. {(**A**,**U**),(**U**,**A**),(**G**,**C**),(**C**,**G**)}) and {(**G**,**U**),(**U**,**G**)}, satisfy conditions Eqs. (2.1)–(2.3).

The above observation is insofar remarkable since making the graph H (being responsible for all interactions) more complex can simply imply that not all of its subgraphs can be folded by sequences. This is due, as the example indicates, to the nature of the base pairing rule and shows clearly that both: the graph H and the relation $\mathscr{R}$ determine what is a shape and what is not. For simple base graphs, like for instance $\mathscr{H}$, the lemma below shows that *any* subgraph of $\mathscr{H}$ (Eq. (2.5)) is a shape. What we can deduce from this is (a) there exist many shapes and (b) $\mathscr{H}$ is so simple that it is indeed only $\mathscr{R}^{\dagger}$ that is relevant for the shapes. The result is:

Lemma 1. *Suppose H is an arbitrary combinatorial graph over $\{1,\ldots,n\}$.*

(a) *For any relation $\mathscr{R}$ any shape $\mathscr{S}$ is bipartite.*
(b) *For the relation $\mathscr{R}^{\dagger}$ and arbitrary base graph H, any induced, bipartite subgraph of H is a shape.*
(c) *For the relation $\mathscr{R}^{\dagger}$ and the base graph $\mathscr{H}$ any $\mathscr{H}$-subgraph H' is a shape.*

Since any $\mathscr{H}$-subgraph is a shape we have for instance for sequences of length 16 exactly $2^{16} = 65536$ different shapes in difference to only 274 RNA secondary structures realized by the minimum free energy folding analyzed in Göbel (2000). This seems to indicate a vast difference between combinatory maps and RNA secondary structure folding, however, closer inspection reveals that in fact most of these structures are very "rare", i.e. only a few have large preimage sizes. To understand what is happening we present in Fig. 5 the data on the complete mapping from sequences of length 16 into subgraphs of the cycle $\mathscr{H}_{16}$. We plot the logarithm of the preimage sizes of a combinatory map over the logarithm of the rank. We can deduce from Fig. 5 that there are 393 shapes with a preimage of size greater than 0.5×10^6. The data on RNA secondary structures in Göbel (2000) show that there are 132 RNA minimum free energy structures with this property. Fig. 5 shows that combinatory maps exhibit 393 shapes with a preimage of size greater than 0.5×10^6. As for RNA secondary structures the data in Göbel (2000) show that there are 132 RNA minimum free energy structures with this property. But what happens for larger sequence length? The asymptotics of RNA secondary structures (Hofacker et al., 1998; Jin and Reidys, 2007) shows that the number of RNA secondary structures, $\mathsf{S}_2(n)$, satisfies $\mathsf{S}_2(n) \sim \kappa n^{-3/2}\alpha^n$ where $1.8488 \leqslant \alpha \leqslant 2.64$, depending on what one considers a "realistic" secondary structure. In comparison a combinatory map produces (Lemma 1) 2^n shapes. Therefore combinatory maps produce a total number of structures which is, for large n, in a comparable size-range. The above observations motivate the question about the number of shapes with large preimages (Flamm et al., 2001). For notational convenience let

$$\mu_+ = \left(\frac{1+\sqrt{5}}{2}\right) \quad \text{and} \quad \mu_- = \left(\frac{1-\sqrt{5}}{2}\right). \tag{3.2}$$

We next prove that there are many shapes with large preimages.

Lemma 2. *Suppose the relation $\mathscr{R}^{\dagger}$ and the base graph $\mathscr{H}$ are given, then there exist at least $(\sqrt{2})^{n-1}$ shapes with the property that there are at least $2(\mu_+^n + \mu_-^n)$ sequences folding into them.*

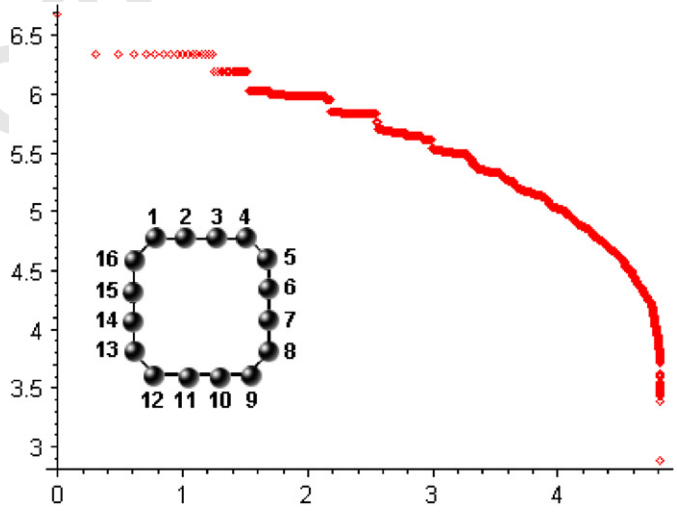


Fig. 5. A double logarithmic plot (base 10) of the preimage sizes of a combinatory map for $n = 16$ as a function of the rank. The underlying graph $\mathscr{H}_{16}$ is displayed in the lower right. The plot shows that there are a few shapes with large and many shapes with very small preimages. This observation is in complete analogy with RNA secondary structure folding maps.

Lemma 2 sets the stage for the further investigation of how this set of sequences is organized. Now, knowing that there are exponentially large sets of sequences realizing particular shapes what can be said about their organization? Are they randomly distributed or clustered in sequence space? What is their graph-structure considered as induced subgraphs of sequence space?

4. Neutral networks of combinatory maps

One difficulty in the context of neutral networks is that it is practically impossible to prove they exist. Exhaustive enumeration of sequence spaces is limited to small sequence length $n \leqslant 20$ for four letter alphabets (Grüner et al., 1996a) and the results are of limited value since finite size effects distort the picture. In case of **A**, **U**, **G**, **C**-sequences about 60% of all sequences fold into the open structure (Göbel, 2000). Several attempts have been made to derive somewhat local criteria whether neutral networks exist (Goebel and Forst, 2002), where the key idea is the probing for paths adopted from the actual random graph proof in Reidys et al. (1997), Reidys (2003). In this context local parameters are the only quantities that give some clue about the existence and properties of neutral networks. In case of neutral networks modeled as random graphs, it is the number of neutral neighbors that controls global properties like connectivity and density of the corresponding neutral network. A neutral neighbor is a neighboring sequence which folds into the same structure and the fraction (Reidys, 1997)

$$\lambda^* = 1 - \sqrt[\alpha-1]{\alpha^{-1}} \tag{4.1}$$

is actually the threshold value for connectivity and density. In the following we can derive for combinatory maps the entire distribution of neutral neighbors of particular shapes. The result is actually not "local" at all and entails detailed information about the *entire* preimage of these shapes. To be precise we can actually derive the underlying rational generating function using the transfer matrix method of enumerative combinatorics. Let $M \subset \{1,\dots,k\}$; we study the quantity $\lambda_{\mathscr{S}_M}(m)$ being the number of sequences folding into the shape $\mathscr{S}_M$ having exactly m neutral neighbors. Here $\mathscr{S}_M$ denotes the shape obtained from $\mathscr{H}$ by removing all edges incident to vertices $j \in M$. Our result reads:

Theorem 1. *For arbitrary shape $\mathscr{S}_M$, which is obtained from $\mathscr{H}$ by deleting the two edges incident to a vertex $i \in M$, we have*

$$\forall m \in \mathbb{N}: \lambda_{\mathscr{S}_M}(m) \geqslant \lambda_{C_{2k}}(m) \tag{4.2}$$

and the generating function of $\lambda_{C_{2k}}(m)$, $F(x,y) = \sum_{k\geqslant 2}\sum_m \lambda_{C_{2k}}(m)x^m y^{2k}$ is given by

$$F(x,y) = \frac{2(-4x^3y^6 + 2x^2y^6 + 3x^2y^4 - 5 + 4x^2y^2 + 8xy^2 - 6x^3y^4 + 2x^4y^6)}{-2x^3y^6 + x^2y^6 + x^2y^4 - 1 + 2xy^2 + x^2y^2 - 2x^3y^4 + x^4y^6}. \tag{4.3}$$

The bi-variate function $F(x,y)$ provides detailed information about neutral neighbors, of the entire preimages of shapes $\mathscr{S}_M$. For instance, Taylor expansion of Eq. (4.3) yields

$$F(x,y) = 10 + (2x^2 + 4x)y^2 + (12x^2 + 2x^4)y^4 + (6x^2 + 16x^3 + 12x^4 + 2x^6)y^6 + O(y^8)$$

and the term $(12x^2 + 2x^4)y^4$ shows that for $n = 4$ there are at least 12 vertices with 2 and 2 vertices with 4 neutral neighbors. Likewise, for $n = 6$, there are at least 6 with 2, 16 with 3, 12 with 4 and 2 vertices with 6 neutral neighbors. In addition Eq. (4.2) guarantees that $\mathscr{H}$ itself provides a lower bound on the numbers of neutral neighbors. I.e. we can pinpoint a specific reference shape providing key information about the neutrality of the entire combinatory map (Fig. 6).

In the previous section we have shown that there are many shapes with large preimages. However, it is not obvious what the graph structure of these preimages is. In this section we will study this structure in detail and prove two remarkable properties. First there are many shapes with sets of sequences having diameter n i.e. there exist two sequences which differ in *all* nucleotides both of which map into the particular shape. This finding is tantamount to percolation and indicates that the preimages are indeed extended and not confined in some "local" region of sequence space. Secondly we prove that the preimages of exponentially many shapes contain large connected components. In other words we can actually prove the existence of neutral networks for sequence to shape maps, i.e. many shapes have sets of sequences in which there exists a component of size $\geqslant(\sqrt{2})^n$ and of diameter n (Fig. 7).

Theorem 2 (*Neutral networks*). *Suppose the relation $\mathscr{R}^\dagger$ and the base graph $\mathscr{H}$ are given. Then there exist at least $(\sqrt{2})^{n-1}$ many shapes $\mathscr{S}$ with the properties*:

(I) *the set of all sequences mapping into $\mathscr{S}$ has a connected component of size at least $\mu_+^n + \mu_-^n$;*
(II) *the set of all sequences mapping into $\mathscr{S}$ percolates, i.e. has diameter n.*

Please cite this article as: Jin, E.Y., et al., Neutral networks of sequence to shape maps. J. Theor. Biol. (2007), doi:10.1016/j.jtbi.2007.09.012

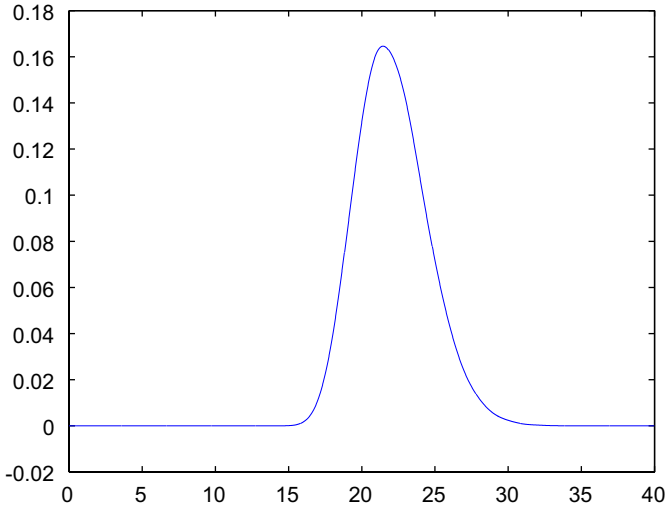


Fig. 6. The distribution of neutral neighbors for the entire preimage of the "reference" shape $\mathscr{S} = \mathscr{H}_{40}$, where $n = 40$ denotes the sequence length. We plot the frequency (y-axis) of numbers of neutral neighbors (x-axis) obtained from Theorem 1. Note that the degree of a vertex in Q_4^{40} is 120, showing that the lower bounds on the fractions of neutral neighbors range between 13% and 24%.

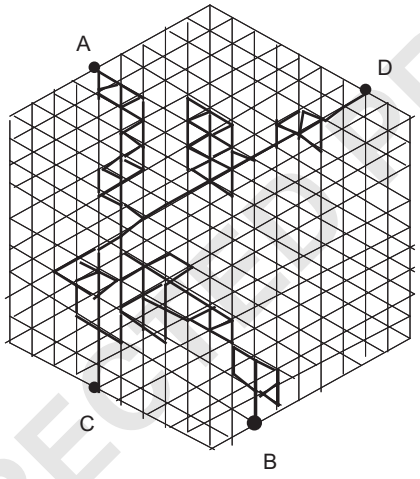
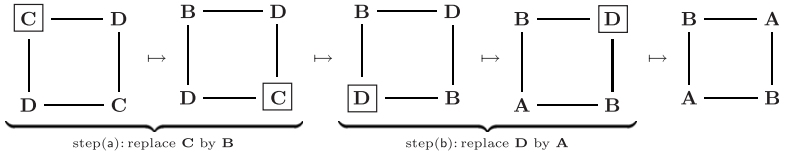


Fig. 7. Neutral network. Sequence space is represented as lattice and the neutral net is an induced subgraph (bold edges). We label the pairs of sequences representing antipodal pairs by (A, B) and (C, D). The two key properties of neutral nets are their connectivity and percolation.

In comparison with the corresponding result for random graphs we observe that the neutral networks are slightly bigger and the diameter *equals* n. This is a result from the fact that the simpler graph $\mathscr{H}$ allows for a different proof strategy. The proof indicates how to explicitly obtain these paths of diameter n, while the random graph analogue can only produce their existence. In this sense both constructions complement each other. To illustrate the idea of Theorem 2 let us consider the cycle C_4 and the shape $\mathscr{S} = C_4$. Then we have (using the notation of the proof of Theorem 2)

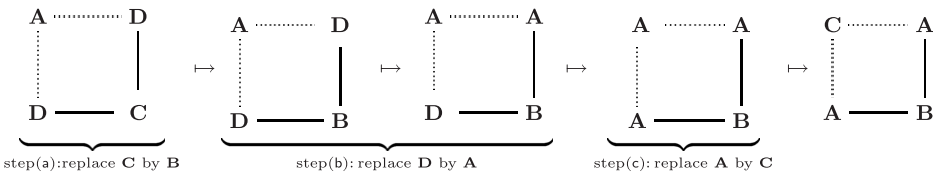
$$a^{\emptyset} = (\mathbf{C}, \mathbf{D}, \mathbf{C}, \mathbf{D}) \quad \text{and} \quad \mathscr{H}_{\mathscr{R}^\dagger}((\mathbf{C}, \mathbf{D}, \mathbf{C}, \mathbf{D})) = C_4.$$

Theorem 2 guarantees the existence of the antipodal sequence $\tilde{a}^{\emptyset} = (\mathbf{B}, \mathbf{A}, \mathbf{B}, \mathbf{A})$ and a path connecting $a^{\emptyset}$ and $\tilde{a}^{\emptyset}$ obtained via the steps (a), (b) and (c). Explicitly this path for $\mathscr{S}_{\emptyset}$ from $a^{\emptyset}$ to $\tilde{a}^{\emptyset}$ is given by



Theorem 2 holds for many shapes. For instance the neutral path for $\mathscr{S}_{\{1\}}$, which has length $\mathsf{diam}(Q_4^4) = 4$ and which connects the sequences $a^{\{1\}}, \tilde{a}^{\{1\}}$ is given by

Please cite this article as: Jin, E.Y., et al., Neutral networks of sequence to shape maps. J. Theor. Biol. (2007), doi:10.1016/j.jtbi.2007.09.012



Acknowledgments

We thank F.W.D. Huang and L.C. Zuo for helpful suggestions. 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.

Appendix A

Proof of Lemma 1. To show (a) we first prove that for any relation satisfying Eqs. (2.1), (2.2) and (2.3) a shape $\mathscr{S}$ is bipartite.

Claim. *Any closed walk in $\mathscr{S}$ has even length. Since $\mathscr{S}$ is a shape we have $\mathscr{S} = H(v)$, whence for any closed walk $w = (w_1, w_2 \ldots, w_r, w_1)$ in $\mathscr{S}$ there exists at least one sequence $x = (x_{w_1}, x_{w_2}, \ldots, x_{w_r}, x_{w_1})$, where $x_h \in \{\mathbf{A}, \mathbf{D}, \mathbf{B}, \mathbf{C}\}$. Therefore there exists an injection*

$$\{(x_{w_1}, x_{w_2}, \ldots, x_{w_r}, x_{w_1}) \mid w \text{ is a closed walk in } \mathscr{S}\} \longrightarrow \{\gamma \mid \gamma \text{ is a closed walk in } G(\mathscr{R})\}.$$

In view of Eqs. (2.1) and (2.2) each relation can be represented as the combinatorial graph $G(\mathscr{R}) = (V_G(\mathscr{R}), E_G(\mathscr{R}))$, where $V_G(\mathscr{R}) = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}, E_G(\mathscr{R}) = \{\{x, y\} \mid x, y \in V_G(\mathscr{R}), (x, y) \in \mathscr{R}\}$. The idea is to show that

$$\{\gamma \mid \gamma \text{ is a closed walk in } G(\mathscr{R}) \text{ of odd length}\} = \emptyset.$$

Suppose γ is a closed walk of minimal, odd length in $G(\mathscr{R})$. Obviously, there are only four vertices in $G(\mathscr{R})$. We can conclude from this that γ contains a cycle of length 3 which is in view of Eq. (2.3) impossible, whence the claim.

We next select an arbitrary vertex, $i \in \{1 \ldots n\}$ and color all vertices in even distance to i blue and all vertices in odd distance red. Suppose this procedure leads to two monochromatic adjacent vertices j, r. Then we obtain a closed walk containing i, j and r of odd length. By induction we can conclude that this walk contains a cycle of odd length, which is impossible, whence $\mathscr{S}$ is bipartite and assertion (a) follows.

Next we show (b) by constructing a vertex $v = (x_1, \ldots, x_n) \in Q_4^n$ with the property $H_{\mathscr{R}^\dagger}(v) = H'$, where H' is an arbitrary induced, bipartite subgraph of H. Since H' is induced in H there exists some set $M \subset \{1, \ldots, n\}$ such that $E_{H'} = \{\{i, j\} \mid \{i, j\} \in E_H \wedge i, j \in M\}$. First, for all coordinates x_j where $j \notin M$ we set $x_j = \mathbf{A}$. Then by definition of $\mathscr{R}^\dagger$ for $i, i' \notin M$, $\{x_i, x_{i'}\} \notin \mathscr{R}^\dagger$ holds. Since H' is bipartite there exists for the vertices $j \in M$ a bi-coloring (red/blue) such that no two H'-adjacent vertices are monochromatic. Suppose x_j, x_k are coordinates where $j, k \in M$. We choose a bi-coloring (red/blue) and set $x_j = \mathbf{D}$ for j being colored red and $x_k = \mathbf{C}$ for k being colored blue, respectively. In view of $(\mathbf{D}, \mathbf{C}), (\mathbf{C}, \mathbf{D}) \in \mathscr{R}^\dagger$, we can conclude that for $j, k \in M$ and $\{j, k\} \in H$ we have $\{x_j, x_k\} \in \mathscr{R}^\dagger$. Since $(\mathbf{A}, \mathbf{C}), (\mathbf{A}, \mathbf{D}) \notin \mathscr{R}^\dagger$ we derive that for $i \notin M$ and $j \in M$, $\{x_i, x_j\} \notin \mathscr{R}^\dagger$ holds. Therefore $H_{\mathscr{R}^\dagger}((x_1, \ldots, x_n)) = H'$ i.e. any induced bipartite subgraph of H is a shape.

Next we show (c), i.e. for $\mathscr{H}$ (Eq. (2.5)) any $H' < \mathscr{H}$ is a shape. We proceed by explicitly constructing a vertex $v = (x_1, \ldots, x_n) \in Q_4^n$ with the property $\mathscr{H}_{\mathscr{R}^\dagger}(v) = H'$. W.l.o.g. we can assume that n is even since the isolated point u does not contribute to the $\mathscr{H}$-shapes. Then we have $\mathscr{H} = C_{2k}$ and $V_{C_{2k}} = \{1, \ldots, 2k\}$. We label the H'-vertices $\{1, \ldots, 2k\}$ clockwise such that the (clockwise) first vertex in one largest H'-component is 1. Then H' corresponds to a unique sequence of components. We assume now $x_i \in \{\mathbf{A}, \mathbf{B}\}$ and label all H'-vertices except of those contained in the component proceeding vertex 1. We set inductively

$$x_i = \begin{cases} \mathbf{A} & \text{iff } i = 1, \\ x_{i-1} & \text{iff } \{i-1, i\} \text{ is not an edge in } H', \\ \overline{x_{i-1}} & \text{iff } \{i-1, i\} \text{ is an edge in } H', \end{cases} \tag{A.1}$$

where $\overline{\mathbf{B}} = \mathbf{A}$ and $\overline{\mathbf{A}} = \mathbf{B}$. As for the labeling of the component preceding the component containing vertex 1, we start with $x_j = \mathbf{C}$ and continue inductively $x_{j+1} = \mathbf{D}, x_{j+2} = \mathbf{C}, \ldots$. This procedure results in a labeling compatible with H' since for

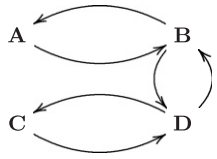
$\{i-1,i\} \in H'$ we have either $\{\mathbf{C},\mathbf{D}\}$ or $\{\mathbf{A},\mathbf{B}\}$ and for $\{i-1,i\} \notin H'$ we have $\{\mathbf{A},\mathbf{A}\}$, $\{\mathbf{B},\mathbf{B}\}$ and $\{\mathbf{A},\mathbf{C}\}$ or $\{\mathbf{B},\mathbf{C}\}$ (at the beginning of the last component) and $\{\mathbf{D},\mathbf{A}\}$ or $\{\mathbf{C},\mathbf{A}\}$ (at the end of the last component). Accordingly we obtain a sequence $\tilde{v}_{H'}$ with the property $\mathscr{H}(\tilde{v}_{H'}) = H'$. □

Proof of Lemma 2. By definition, there exists a unique component of $\mathscr{H}$ which is a cycle of even length, C_{2k}. C_{2k} contains for n even all and for n odd all but one $\mathscr{H}$-vertices. Suppose C_{2k} contains the vertices $\{i_1, j_1, \ldots, i_k, j_k\}$, where $i_1 < j_1 < i_2 < \cdots < i_k < j_k$.

Claim. *The number of 2k-tuples* $(x_{i_1}, x_{j_1}, \ldots, x_{i_k}, x_{j_k})$ *such that* $C_{2k}((x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k})) = C_{2k}$ *i.e.* $(x_{i_1}, x_{j_1}, \ldots, x_{i_k}, x_{j_k}) \in \vartheta_{C_{2k}}^{-1}(C_{2k})$ *is given by*

$$2(\mu_+^{2k} + \mu_-^{2k}). \tag{A.2}$$

To prove the claim we observe that $\mathscr{R}^\dagger$ induces the digraph $D_{\mathscr{R}^\dagger}$ defined as follows:

$D_{\mathscr{R}^\dagger} = D_{\mathscr{R}^\dagger} =$ 

and $A_{D_{\mathscr{R}^\dagger}} =$

	A	**B**	**D**	**C**
	0	1	0	0
	1	0	1	0
	0	1	0	1
	0	0	1	0

The number of $2k$-tuples $(x_{i_1}, x_{j_1}, \ldots, x_{i_k}, x_{j_k})$ with the property $C_{2k}((x_{i_1}, x_{j_1}, \ldots, x_{i_k}, x_{j_k})) = C_{2k}$ is equal to the number of closed walks of length $2k$ in $D_{\mathscr{R}^\dagger}$. Indeed, in order to obtain such an $2k$-tuple we fix an index, i_1, say. Then we start with successively **A**, **B**, **D** and **C** and form closed walks of length $2k$ in $D_{\mathscr{R}^\dagger}$ starting and ending at **A**, **B**, **D** and **C**. All these walks are counted, respectively, since we have labeled graphs. The number of closed walks of length ℓ in $D_{\mathscr{R}^\dagger}$ starting and ending at i is given by $(A^\ell_{D_{\mathscr{R}^\dagger}})_{i,i}$, whence the number of all closed walks of length ℓ is simply $\mathsf{Tr}(A^\ell_{D_{\mathscr{R}^\dagger}}) = \sum_i (A^\ell_{D_{\mathscr{R}^\dagger}})_{i,i}$. From the definition of the characteristic polynomial, i.e. $\mathsf{Tr}(A^\ell_{D_{\mathscr{R}^\dagger}}) = \omega_1^\ell + \cdots + \omega_r^\ell$, where $\omega_1, \ldots, \omega_r$ are the eigenvalues of $A_{D_{\mathscr{R}^\dagger}}$ (note $r = 4$). We obtain

$$\begin{aligned}\sum_{\ell \geqslant 0} \mathsf{Tr}(A^\ell_{D_{\mathscr{R}^\dagger}}) z^\ell &= \sum_{\ell \geqslant 0} [\omega_1^\ell + \cdots + \omega_r^\ell] z^\ell \\ &= \sum_{\ell \geqslant 0} [(1 + (-1)^\ell)(\mu_+^\ell + \mu_-^\ell)] z^\ell\end{aligned}$$

and the claim follows.

Suppose $(x_{i_1}, x_{j_1}, \ldots, x_{i_k}, x_{j_k}) \in \vartheta_{C_{2k}}^{-1}(C_{2k})$ and $M \subset \{1, \ldots, k\}$. We consider the involution $\tau: \mathscr{A} \to \mathscr{A}$, where $\tau(\mathbf{A}) = \mathbf{B}$ and $\tau(\mathbf{D}) = \mathbf{C}$ and set

$$I_M(x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k}) = (y_{i_1}, x_{j_1} \ldots, y_{i_k}, x_{j_k}), \text{where } y_{i_\ell} = \begin{cases} \tau(x_{i_\ell}) & \text{for } i_\ell \in M, \\ x_{i_\ell} & \text{for } i_\ell \notin M. \end{cases} \tag{A.3}$$

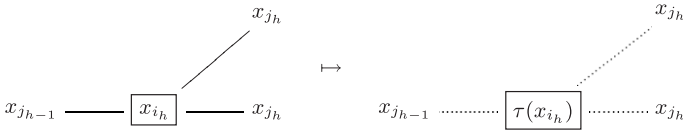
Claim. *There exists a bijection*

$\beta: \{M \subset \{1, 2, \ldots, k\}\} \to \{\mathscr{S}_M\}, \quad M \mapsto \mathscr{S}_M$

where $\mathscr{S}_M$ *is obtained by deleting any two* C_{2k}*-edges incident to the vertices* $i_h \in M$ *and*

$$\forall (x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k}) \in \vartheta_{C_{2k}}^{-1}(C_{2k}), \mathscr{S}_M = C_{2k}(I_M(x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k})). \tag{A.4}$$

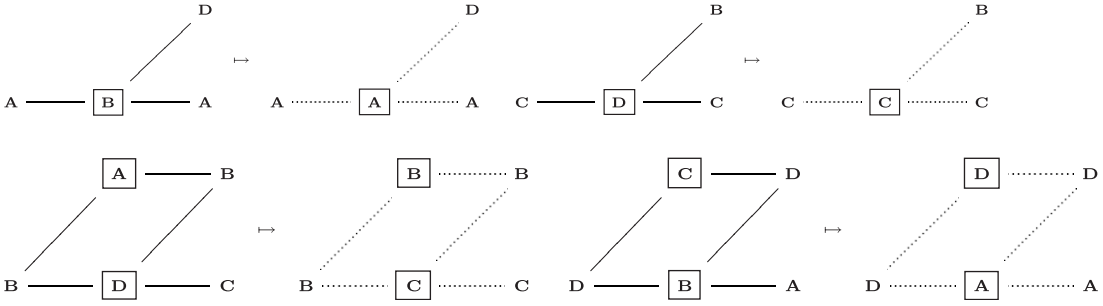
Suppose $M \neq M'$ then w.l.o.g. we can assume that there exists some index $i_h \in M \backslash M'$, i.e. i_h is isolated in $\mathscr{S}_M$ but not in $\mathscr{S}_{M'}$. Since j_{h-1} and j_h are both in $\mathscr{S}_M$ and $\mathscr{S}_{M'}$ we have $\{j_{h-1}, i_h\}, \{j_h, i_h\}, \in \mathscr{S}_{M'}$ but not in $\mathscr{S}_M$, whence $\mathscr{S}_M$ and $\mathscr{S}_{M'}$ are different shapes. Since $\mathscr{S}_M$ is an induced bipartite subgraph, Lemma 1 implies that any $\mathscr{S}_M$ is a shape. When $i_h \in M$ the following diagram:



shows that I_M has the property: for arbitrary

$(x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k}) \in \vartheta_{C_{2k}}^{-1}(C_{2k})$

the shape $C_{2k}(I_M(x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k}))$ differs from C_{2k} exactly by deleting the two C_{2k}-edges incident to all $i_\ell \in M$; explicitly



and the claim is proved. The claim implies that I_M induces the injection

$$I_M: \vartheta_{C_{2k}}^{-1}(C_{2k}) \longrightarrow \vartheta_{C_{2k}}^{-1}(\mathscr{S}_M), (x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k}) \mapsto I_M(x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k}). \tag{A.5}$$

This injection allows us to relate the sets $\vartheta_{C_{2k}}^{-1}(C_{2k})$ and $\vartheta_{C_{2k}}^{-1}(\mathscr{S}_M)$ and in particular

$$|\vartheta_{C_{2k}}^{-1}(C_{2k})| \leqslant |\vartheta_{C_{2k}}^{-1}(\mathscr{S}_M)|. \tag{A.6}$$

Since $M \subset \{1, \ldots, k\}$ was arbitrary we can conclude that there are 2^k subsets and hence 2^k distinct shapes $\mathscr{S}_M$. Hence there exist at least

$$2^k \geqslant (\sqrt{2})^{n-1}$$

shapes $\mathscr{S}$ with the property

$$|\vartheta_{\mathscr{H}}^{-1}(\mathscr{S})| \geqslant |\vartheta_{\mathscr{H}}^{-1}(\mathscr{H})| \geqslant 2(\mu_+^{2k} + \mu_-^{2k}).$$

In case of $n \not\equiv 0 \bmod 2$ we have exactly one more isolated point, i.e.

$$|\vartheta_{\mathscr{H}}^{-1}(\mathscr{S})| \geqslant 8(\mu_+^{n-1} + \mu_-^{n-1}) \tag{A.7}$$

and since $4 \geqslant (\mu_+ + \mu_-)$ the lemma follows. □

Proof of Theorem 2. We first prove that at least $(\sqrt{2})^{n-1}$ shapes $\mathscr{S}$ have a preimage $\vartheta_{\mathscr{H}}^{-1}(\mathscr{S})$ with diameter n. We will work with the particular set of shapes $\{\mathscr{S}_M \mid M \subset \{1, \ldots, k\}\}$, introduced in Lemma 2 and prove that all of them have a component of size $\geqslant \mu_+^n + \mu_-^n > (\sqrt{2})^n$ and $\mathsf{diam}(\vartheta_{\mathscr{H}}^{-1}(\mathscr{S})) = n$. Let C_{2k} be the $\mathscr{H}$-cycle, which contains all $\mathscr{H}$-vertices for n even and all but one $\mathscr{H}$-vertices, for n odd. Let $V_{C_{2k}} = \{i_1, j_1, \ldots, i_k, j_k\}$, where $i_1 < j_1 < i_2 < \cdots i_k < j_k$.

Claim 1. *Let $M \subset \{1, \ldots, k\}$, then there exist at least 2^k shapes S_M over Q_4^{2k} such that*

$$\mathsf{diam}(\vartheta_{\mathscr{H}}^{-1}(\mathscr{S}_M)) = \begin{cases} n & \text{for } n \equiv 0 \bmod 2, \\ n-1 & \text{for } n \not\equiv 0 \bmod 2. \end{cases} \tag{A.8}$$

We first show that for each M there exists a pair of antipodal sequences, i.e. $(a^M, \tilde{a}^M)$ with $d(a^M, \tilde{a}^M) = 2k$ and a path $(a^M, w_1^M, \ldots, w_{2k-1}^M, \tilde{a}^M)$ such that $\vartheta_{C_{2k}}(w_i^M) = \mathscr{S}_M$.

$$a^M = (a_{i_1}^M, a_{j_1} \ldots, a_{i_k}^M, a_{j_k}), \quad \text{where } a_{j_h} = \mathbf{D} \quad \text{and} \quad a_{i_h}^M = \begin{cases} \mathbf{A} & \text{for } i_h \in M, \\ \mathbf{C} & \text{otherwise.} \end{cases} \tag{A.9}$$

In particular we have $a^\emptyset = (\mathbf{C}, \mathbf{D}, \ldots, \mathbf{C}, \mathbf{D})$. Then $\mathscr{S}_M = C_{2k}(a^M)$, i.e. $\mathscr{S}_M$ is the shape obtained by removing for each $i_h \in M$ the two incident C_{2k}-edges. Next we define an antipode $\tilde{a}^M$, i.e. an element of Q_4^{2k} with the property $d(a^M, \tilde{a}^M) = 2k$ as follows:

$$\tilde{a}^M = (\tilde{a}_{i_1}^M, \tilde{a}_{j_1} \ldots, \tilde{a}_{i_k}^M, \tilde{a}_{j_k}) \quad \text{where } \tilde{a}_{j_h} = \mathbf{A} \quad \text{and} \quad \tilde{a}_{i_h}^M = \begin{cases} \mathbf{C} & \text{for } i_h \in M, \\ \mathbf{B} & \text{otherwise.} \end{cases} \tag{A.10}$$

We can transform a^M into $\tilde{a}^M$ by successively changing exactly one coordinate in three steps: (a) replace (in any order) for $i_h \notin M$ successively all $a_{i_h} = \mathbf{C}$ by $\mathbf{B}$, (b) replace (in any order) successively all $a_{j_h} = \mathbf{D}$ by $\mathbf{A}$ and finally (c) substitute (in any order) for all $i_h \in M$ $a_{i_h} = \mathbf{A}$ by $\mathbf{C}$.

This proves that there exists a Q_4^{2k}-path

$$(a^M, w_1^M, \ldots, w_{2k-1}^M, \tilde{a}^M) \tag{A.11}$$

connecting a^M and $\tilde{a}^M$, such that

$$\forall 1 \leqslant i \leqslant 2k-1, \quad C_{2k}(w_i^M) = \mathscr{S}_M. \tag{A.12}$$

I.e. all intermediate steps of the path are mapped by $\vartheta_{\mathscr{H}}$ into the shape $\mathscr{S}_M$. As shown in Lemma 2 there are 2^k different shapes $\mathscr{S}_M$ induced by the subsets $M \subset \{1, \ldots, k\}$, whence Claim 1.

In case of $n \equiv 0 \bmod 2$ we derive $2^k = (\sqrt{2})^n$. In case of $n \not\equiv 0 \bmod 2$ there exists exactly one vertex u which is isolated in $\mathscr{H}$. Then we simply add the isolated point u to each shape $\mathscr{S}_M$ and shall in the following identify these new shapes with $\mathscr{S}_M$. Then $|\vartheta_{\mathscr{H}}^{-1}(\mathscr{S}_M)| = 4|\vartheta_{C_{2k}}^{-1}(\mathscr{S}_M)|$. We can choose $a_u = \mathbf{A}$ and $\tilde{a}_u = \mathbf{B}$ and

$$a_u^M = (a_{i_1}^M, a_{j_1} \ldots, a_u, \ldots, a_{i_k}^M, a_{j_k}),$$

$$\tilde{a}_u^M = (\tilde{a}_{i_1}^M, \tilde{a}_{j_1} \ldots, \tilde{a}_u, \ldots, \tilde{a}_{i_k}^M, \tilde{a}_{j_k})$$

satisfy $d(a_u^M, \tilde{a}_u^M) = n$ and there exists a Q_4^n-path $(a_u^M, w_1^M, \ldots, w_{2k}^M, \tilde{a}_u^M)$ connecting a_u^M and $\tilde{a}_u^M$, with the property

$$\forall 1 \leqslant i \leqslant 2k, \quad C_{2k}(w_i^M) = \mathscr{S}_M. \tag{A.13}$$

Therefore we have proved that at least $(\sqrt{2})^{n-1}$ shapes $\mathscr{S}_M$ have a preimage $\vartheta_{\mathscr{H}}^{-1}(\mathscr{S}_M)$ with diameter n.

Claim 2.

$$|\{\mathscr{S}_M \mid |\mathscr{C}(\vartheta_{\mathscr{H}}^{-1}(\mathscr{S}))| \geqslant \mu_+^{2k} + \mu_-^{2k}\}| \geqslant 2^k. \tag{A.14}$$

To prove Claim 2 we first observe that $\vartheta_{\mathscr{H}}^{-1}(\mathscr{H})$ has exactly two components of equal size

$$\mu_+^{2k} + \mu_-^{2k}. \tag{A.15}$$

Indeed, any vertex $v \in \vartheta_{\mathscr{H}}^{-1}(\mathscr{H})$ can be transformed into either

$$a^{\emptyset} = (\mathbf{C}, \mathbf{D}, \mathbf{C}, \ldots, \mathbf{D}, \mathbf{C}) \quad \text{or} \quad b^{\emptyset} = (\mathbf{D}, \mathbf{C}, \ldots, \mathbf{D}, \mathbf{C}, \mathbf{D})$$

successively using the two steps (I) replace (in any order) all **A** by **D** and (II) replace all (in any order) **B** by **C**. Hence there exist exactly two components and the map

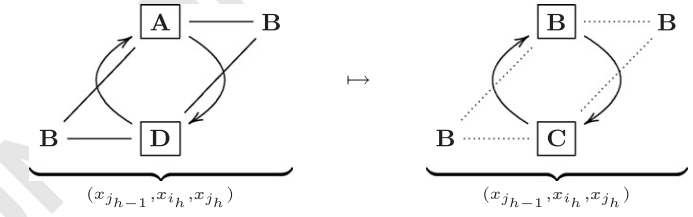
$$\sigma(x_{i_1}, x_{j_1}, \ldots, x_{i_k}, x_{j_k}) = (x_{j_k}, x_{i_1}, \ldots, x_{j_{k-1}}, x_{i_k})$$

is a bijection between them, whence they have equal size. Eq. (A.15) then follows from Eq. (A.2) in Lemma 2. We next claim that the mapping I_M of Eq. (A.3) is in fact an injective graph morphism

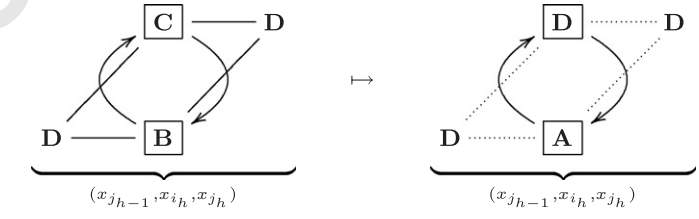
$$I_M : \vartheta_{C_{2k}}^{-1}(C_{2k}) \longrightarrow \vartheta_{C_{2k}}^{-1}(\mathscr{S}_M), (x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k}) \mapsto I_M(x_{i_1}, x_{j_1} \ldots, x_{i_k}, x_{j_k}). \tag{A.16}$$

I.e. for two adjacent vertices $v, v' \in \vartheta_{C_{2k}}^{-1}$, the vertices $I_M(v)$ and $I_M(v')$ are adjacent. To prove this we consider the diagrams

$x_{j_{h-1}} = x_{j_h} = \mathbf{B}$:



$x_{j_{h-1}} = x_{j_h} = \mathbf{D}$:



The above diagrams represent the two scenarios for two adjacent vertices $v, v' \in \vartheta_{C_{2k}}^{-1}(C_{2k})$. I.e. if v and v' are both contained in $\vartheta_{C_{2k}}^{-1}(C_{2k})$ and differ in x_{i_h} and x'_{i_h} then we have either $x_{j_{h-1}} = x_{j_h} = \mathbf{B}$ and $x_{i_h} = \mathbf{D}$ and $x'_{i_h} = \mathbf{A}$ or $x_{j_{h-1}} = x_{j_h} = \mathbf{D}$ and $x_{i_h} = \mathbf{B}$ and $x'_{i_h} = \mathbf{C}$. Suppose we apply I_M and $i_h \in M$, then the resulting vertices $I_M(v)$ and $I_M(v')$ are again

adjacent, whence I_M is an injective graph morphism. Accordingly, I_M maps components into components, from which we can conclude that for each $M \subset \{1,\dots,k\}$ the shape $\mathscr{S}_M$ has a component of size $\mu_+^{2k}+\mu_-^{2k}$ and Claim 2 is proved.

In case of $2k=n$ the assertion follows directly. For n odd we have to repeat the argument in Lemma 2, where we considered the isolated point u in Eq. (A.7). Since we used the same set of shapes $\{\mathscr{S}_M \mid M \subset \{1,\dots,k\}\}$ for both claims the theorem follows. □

Proof of Theorem 1. It is clear that we can restrict our analysis to the case $n \equiv 0 \bmod 2$, i.e. $\mathscr{H}=C_{2k}$, since the isolated point contributes always four neutral neighbors for any shape. Eq. (4.2) is a direct consequence of

$$I_M\colon \vartheta_{C_{2k}}^{-1}(C_{2k}) \longrightarrow \vartheta_{C_{2k}}^{-1}(\mathscr{S}_M),\ (x_{i_1},x_{j_1}\dots,x_{i_k},x_{j_k}) \mapsto I_M(x_{i_1},x_{j_1}\dots,x_{i_k},x_{j_k})$$

being an injective graph morphism. Thus it suffices to prove Eq. (4.3). We observe that for $v \in \vartheta_{C_{2k}}^{-1}(C_{2k})$

$$v=(x_{i_1},x_{j_1},\dots,x_{i_k},x_{j_k}) \mapsto (t_{i_1},t_{j_1},\dots,t_{i_k},t_{j_k}), \text{where } t_s=\begin{cases}(x_{j_{h-1}},x_{i_h},x_{j_h}) & \text{for } s=i_h,\\ (x_{i_h},x_{j_h},x_{i_{h+1}}) & \text{for } s=j_h\end{cases}$$

is a bijection, where h is considered modulo k. Hence every $v \in \vartheta_{C_{2k}}^{-1}(C_{2k})$ can be uniquely decomposed into a sequence of triples. Since $v \in \vartheta_{C_{2k}}^{-1}(C_{2k})$ there are exactly the following 10 triples

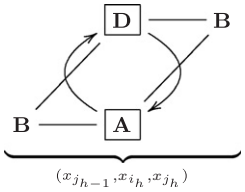
$$V_D=\{\mathbf{ABA},\mathbf{ABD},\mathbf{BAB},\mathbf{BDB},\mathbf{BDC},\mathbf{DBD},\mathbf{DBA},\mathbf{DCD},\mathbf{CDC},\mathbf{CDB}\}$$

and setting

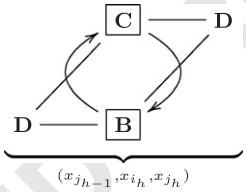
$$E_D=\{((x_{j_{h-1}},x_{i_h},x_{j_h}),(x_{i_h},x_{j_h},x_{i_{h+1}})) \mid (x_{j_{h-1}},x_{i_h},x_{j_h}) \in V_D\}$$

we obtain the digraph D. Suppose we are given $v,v' \in \vartheta_{C_{2k}}^{-1}(C_{2k})$ with $d(v,v')=1$ then we have the following alternative:

$x_{j_{h-1}}=x_{j_h}=\mathbf{B}$:



$x_{j_{h-1}}=x_{j_h}=\mathbf{D}$:



The idea is now to count all triples i.e. $(x_{j_{h-1}},x_{i_h},x_{j_h})$, $(x_{i_{h-1}},x_{j_{h-1}},x_{i_h})$ contained in $\Theta=\{\mathbf{BAB},\mathbf{BDB},\mathbf{DBD},\mathbf{DCD}\}$ in $\vartheta_{C_{2k}}^{-1}(C_{2k})$. Let next $R[x]$ be a polynomial ring and $w\colon E_D \longrightarrow R[x]$ a function given by $w(e)=x$ iff the arc e has terminus $\tau \in \Theta$, otherwise $w(e)=1$. If $\Gamma=e_1e_2\dots e_\ell$ is a walk of length ℓ in E_D, then the weight of Γ is defined by $w(\Gamma)=w(e_1)w(e_2)\dots w(e_\ell)$. Introducing the formal variable x in w allows us to count the triples in Θ within some $v \in \vartheta_{C_{2k}}^{-1}(C_{2k})$. The number of closed walks of length ℓ in D is $\sum_{v\in V_D}[A_D^\ell]_{v,v}=\mathsf{Tr}(A_D^\ell)$, where A_D is the adjacency matrix of D.

Suppose B is a $p\times p$ matrix and $\{\eta_i\}_{i=1}^p$ are all the eigenvalues of B, then we have $\mathsf{det}B=\prod_i \eta_i$. Let $\{\xi_i\}_{i=1}^p$ and $\{\omega_i\}_{i=1}^p$ be all the eigenvalues of $I-yA$ and A, respectively, then we have $\xi_i=1-y\omega_i$, where $1\leqslant i\leqslant p$. For the set of all the nonzero eigenvalues of A, $\{\omega_i\}_{i=1}^r$ we derive $\mathsf{det}(I-yA)=\prod_{i=1}^r(1-y\omega_i)$. We set $Q(y)=\mathsf{det}(I-yA)$ and have $p=10=|V_D|$, $A=A_D$ and $r=6$ for $x\neq 1$, whence

$$\sum_{\ell\geqslant 1}\mathsf{Tr}(A_D^\ell)y^\ell=\sum_{\ell\geqslant 1}(\omega_1^\ell+\dots+\omega_r^\ell)y^\ell=\sum_{i=1}^r\frac{\omega_i y}{1-\omega_i y}=\frac{-yQ'(y)}{Q(y)}. \tag{A.17}$$

After some computation we derive $Q(y)=1-2xy^2-x^2y^2+2x^3y^4-x^4y^6+2x^3y^6-x^2y^6-x^2y^4$ and the lemma follows from Eq. (A.17). □

References

Barnett, L., 1998. Ruggedness and neutrality—the NKp family of fitness landscapes. In: Adami, C., Belew, R.K., Kitano, H., Taylor, C. (Eds.), Artificial Life VI: Proceedings of the Sixth International Conference on Artificial Life. The MIT Press, Cambridge, MA, pp. 18–27.

Bastolla, U., Porto, M., Roman, H., Vendruscolo, M., 2003. Conectivity of neutral networks, overdispersion, and structural conservation in protein evolution. J. Mol. Evol. 56, 243–254.

Batey, R., Rambo, R., Doudna, J., 1999. Tertiary motifs in RNA structure and folding. Angew. Chem. Int. Ed. 38, 2326–2343.

Chamorro, M., Parkin, N., Varmus, H., 1991. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. J. Proc. Natl Acad. Sci. USA 89, 713–717.

Q2 Chen, W., Qin, J., Reidys, C., 2007. Crossings and nestings in 2-diagrams. Electron. J. Combin., Submitted for publication.

Clote, P., Gasieniec, L., Kolpakov, R., Kranakis, E., Krizanc, D., 2005. On realizing shapes in the theory of RNA neutral networks. J. Theoret. Biol. 236, 216–227.

Flamm, C., Hofacker, I., Maurer-Stroh, S., Stadler, P., Zehl, M., 2001. Design of multistable RNA molecules. RNA 7, 254–265.

Fontana, W., Schuster, P., 1998. Continuity in evolution. Science 289, 1451–1455.

Göbel, U., 2000. Neutral networks of minimum free energy rna secondary structures. Ph.D. thesis, University of Vienna.

Goebel, U., Forst, C., 2002. RNA pathfinder–global properties of neutral networks. Z. Phys. Chem. 216, 175.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I., Stadler, P., Schuster, P., 1996a. Analysis of RNA sequence structure maps by exhaustive enumeration I. neutral networks. Chemical Monthly 127, 355–374.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I., Stadler, P., Schuster, P., 1996b. Analysis of RNA sequence structure maps by exhaustive enumeration II. Structures of neutral networks and shape space covering. Chemical Monthly 127, 375–389.

Haslinger, C., Stadler, P., 1999. RNA structures with pseudo-knots. Bull. Math. Biol. 61, 437–467.

Hofacker, I., Schuster, P., Stadler, P., 1998. Combinatorics of RNA secondary structures. Discrete Appl. Math. 88, 207–237.

Q3 Jin, E.Y., Qin, J., Reidys, C.M., 2007. Combinatorics of RNA structures with Pseudoknots. Bull. Math. Bio., in press.

Q4 Jin, E.Y., Reidys, C.M., 2007. Asymptotic enumeration of RNA structures with pseudoknots. Bull. Math. Biol., accepted for publication.

Kauffman, S.A., 1993. The Origin of Order. Oxford University Press, New York, Oxford.

Konings, D.A.M., Gutell, R., 1995. A comparison of thermodynamic foldings with comparatively derived structures of 16s and 16s-like rRNAs. RNA 1, 559–574.

Mapping RNA form and function. Science 2, 2005.

Puglisi, J.D., Williamson, J.R., 1999. The RNA World, second ed. Cold Spring Harbor Laboratory Press, pp. 403–425.

Reidys, C.M., 1997. Random induced subgraphs of generalized *n*-cubes. Adv. Appl. Math. 19, 360–377.

Reidys, C.M., 2000. Random structures. Ann. Combin. 4, 375–382.

Reidys, C.M., 2003. Distances in random induced subgraphs of generalized *n*-cubes. Combin. Probab. Comput. 11, 599–605.

Reidys, C.M., 2007. Random Combinatory Maps, Private communication.

Reidys, C.M., Stadler, P.F., Schuster, P.K., 1997. Generic properties of combinatory maps and neutral networks of RNA secondary structures. Bull. Math. Biol. 59 (2), 339–397.

Schmitt, W., Waterman, M., 1994. Linear trees and RNA secondary structure. Discrete Appl. Math. 51, 317–323.

Schuster, P., 2002. A testable genotype-phenotype map: modeling evolution of RNA molecules. Michael Laessig and Angelo Valeriani, editors, Springer.

Schuster, P., Fontana, W., Stadler, P.F., Hofacker, I.L., 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proc. Roy. Soc. Lond. B 255, 279–284.

Schuster, P., Weber, J., Gruener, W., Reidys, C.M., 1996. Molecular evolutionary biology: from concepts to technology. In: Flyvbjerg, H., Hertz, J., Jensen, M.H., Mouritsen, O.G., Sneppen, K. (Eds.), Physics of Biological Systems: From Molecules to Species. Springer, Berlin, Heidelberg, New York.

Stadler, P.F., Schuster, P., 1999. Discrete models of biopolymers. In: Konopka, A. (Ed.), Handbook of Computational Chemistry and Biology. Marcel Dekker Inc, New York.

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

Waterman, M.S., 1979. Combinatorics of RNA hairpins and cloverleafs. Stud. Appl. Math. 60, 91–96.

Wedekind, J.E., McKay, D.B., 1998. Annu. Rev. Biophys. Biomol. Struct. 27, 475–502.

Westhof, E., Jaeger, L., 1992. RNA pseudoknots. Curr. Opin. Struct. Biol. 2, 327–333.

Woese, C.R., Pace, N.R., 1993. The RNA world. Cold Spring Harbor Laboratory Press, pp. 91–117.

Zuker, M., Sankoff, D., 1984. RNA secondary structures and their prediction. Bull. Math. Biol. 46 (4), 591–621.