Evolutionary Optimization of Biopolymers and Sequence Structure Maps

Christian Reidys[a,b,c], Stephan Kopp[a], Peter Schuster[a,b]

[a]Institut für Molekulare Biotechnologie
Beutenbergstraße 11, D-07708 Jena, Germany

[b]Santa Fe Institute
1399 Hyde Park Rd., Santa Fe, NM 87501, USA

[c]Los Alamos National Laboratory
Theoretical Division, Los Alamos, New Mexico, USA

E-Mail: duck@santafe.edu

Abstract

Searching for biopolymers having a predefined function is a core problem of biotechnology, biochemistry and pharmacy. On the level of RNA sequences and their corresponding secondary structures we show that this problem can be analyzed mathematically. The strategy will be to study the properties of the RNA sequence to secondary structure mapping that is essential for the understanding of the search process. We show that to each secondary structure s there exists a neutral network consisting of all sequences folding into s. This network can be modeled as a random graph and has the following generic properties: it is dense and has a giant component within the graph of compatible sequences. The neutral network percolates sequence space and any two neutral nets come close in terms of Hamming distance. We investigate the distribution of the orders of neutral nets and show [10] that above a certain threshold the topology of neutral nets allows to find practically all frequent secondary structures.

1. Introduction

An RNA structure with a given shape or function is assumed to be formed by many RNA sequences. Their distribution in sequence space is of particular importance for the hardness of the corresponding search problem. The structure of a biopolymer is defined only in the context of some physical conditions. Minimum free energy structures for example fulfill the thermodynamic condition of a molecular ground state, or kinetic structures that are understood as the well defined outcome of a controlled process of biopolymer formation. In an abstract sense this means that one is interested in a (local) point to point assignment of sequence space and shape space. In general such a mapping will not be one-to-one: many sequences may be mapped into the same structure. The degree of this redundancy will strongly depend on the notion of structure applied. Structure in X-ray crystallography is tantamount to a set of atomic coordinates and at sufficiently high resolution structures are unique in the sense that structures from different sequences will never coincide. Molecular biologists, however, commonly apply another, a coarse-grained notion of structure when, for example, they say intuitively that two proteins have the same structure. An appropriate coarse grained notion of structure apparently is context dependent and thus anything but trivial.

In this paper we are dealing with RNA molecules. Secondary structures are used as appropriate examples for structural coarse-graining. They are sufficiently simple to allow statistical analysis by means of conventional combinatorics [9]. The relation between RNA sequences and secondary structures is understood as a (non invertible) mapping from sequence space into shape space [3, 5, 6]. RNA secondary structures distinguish only paired and unpaired regions irrespective of the particular bases at the individual positions (**G**, **C**, **A**, or **U**). Therefore many different sequences the so called compatible sequences can meet the base pairing conditions as determined by a given secondary structure. For the biophysical alphabet the number of compatible sequences is readily computed for any given secondary structure s, with u unpaired bases and p base pairs to be $4^u \cdot 6^p$. The number of compatible sequences is certainly substantially larger than the number of sequences that actually form the given structure as their minimum free energy conformation or, for example, as their kinetically determined structure [14]. Here we pursue to model the sequence structure mapping making the assumption that apart from biophysical pairing rules the mapping is essentially random. Connecting neighboring neutral sequences in sequence space, i.e., sequences that are mapped into the same point in shape space and which are inter-converted by a single move consisting of a base or a base pair exchange, yields neutral networks whose properties are studied by the analytical techniques of random graph theory [2]. The topology of these networks plays a fundamental role for the optimization problem.

The paper is organized as follows: first we summarize some theory of random induced subgraphs of generalized hypercubes following [12] (where all proofs can be found). The corresponding results should make the reader familiar with the generic properties of those random induced subgraphs and the main ideas of the proofs. Second we apply the random graph theory to construct neutral networks and we obtain complete sequence structure mappings by constructing neutral networks with respect to corresponding secondary structures iteratively. We then analyze the so called SOC (sequence of components) of single preimages of these mappings and compute the preimage size distribution depending on the underlying λ parameter. Finally, we discuss the implications for the optimization process of biopolymers.

2. Random Induced Subgraphs of Generalized Hypercubes

A *graph* G consists of a tuple $(\mathrm{v}[G], \mathrm{e}[G])$ and two maps

$$\begin{aligned} \mathrm{e}[G] &\longrightarrow \mathrm{v}[G] \times \mathrm{v}[G] \qquad e \to (\iota(e), \tau(e)) \\ \mathrm{inv} : \mathrm{e}[G] &\longrightarrow \mathrm{e}[G] \qquad e \to \mathrm{inv}(e) \end{aligned}$$

satisfying the following conditions: for each $e \in \mathrm{e}[G]$ we have $\mathrm{inv}^2(e) = e$, $\mathrm{inv}(e) \neq e$ and $\tau(e) = \iota(\mathrm{inv}(e))$. The set $\{e, \mathrm{inv}(e)\}$ is called *geometric edge*. $\mathrm{v}[G]$ is called *vertex set* and $\mathrm{e}[G]$ *edge set* respectively.
Next we recall some terminology of graph theory.

- G' is a *subgraph* of G, $G' < G$, if $\mathrm{v}[G'] \subset \mathrm{v}[G]$ and $\mathrm{e}[G'] \subset \mathrm{e}[G]$.
- Let $H \subset \mathrm{v}[G]$. The *induced subgraph* or *spanned subgraph* of H in G, $G[H]$, has the vertex set $\mathrm{v}[G[H]] = H$ and the edge set $\mathrm{e}[G[H]]$ is the subset of all edges in $\mathrm{e}[G]$ where both incident vertices belong to H.
- The *(out)degree* δ_v of a vertex v is the number of edges $e \in \mathrm{e}[G]$ such that $\iota(e) = v$.
- G is γ*-regular* if for each vertex $v \in \mathrm{v}[G]$ holds $\delta_v = \gamma$.
- The *order* of a graph G, $|\,G\,|$ is the cardinality of its vertex set, i.e.,$|\,\mathrm{v}[G]\,|$.
- A *path* π *in* G is a tuple of the form
$$\pi = (v = v_1, e_1, v_2, \ldots, e_{m-1}, v_m = v')$$
where $\iota(e_k) = v_k, \tau(e_k) = v_{k+1}$ for $1 \leq k < m$. Since π is already characterized by its vertices we use the notation $\pi = (v_i)_{1 \leq i \leq m}$.
We further say that the v_i and e_i *occur* in π. The path π *connects* the vertices v and v', if both occur in π.
- The *support* of a path π is the set
$$\mathrm{Supp}(\pi) := \{v \in \mathrm{v}[G] |\, v \text{ occurs in } \pi\,\}\,.$$
- The *length* of a path $\pi = (v_1, e_1, v_2, \ldots, e_{m-1}, v_m)$ is $\ell(\pi) := m - 1$,i.e. the number of edges that occur in π.
The set of all paths in G shall be denoted by $\Pi(G)$.
- Two vertices $v, v' \in \mathrm{v}[G]$ are called connected if there exists a path in G in which both vertices occur. A graph G is *connected* if for any two vertices $v, v' \in \mathrm{v}[G]$ are connected.
- The *distance* $d_G(v, v')$ of two vertices in G is the minimum length of a path connecting v and v'. If there is no path connecting v and v' we set $d_G(v, v') = \infty$. We shall drop the index G when no confusion is possible.
- The *diameter* of a graph G is the maximum of all distances of pairs of vertices $v, v' \in \mathrm{v}[G]$.
- The *ball* centered at $v \in \mathrm{v}[G]$ with radius r is the set
$$B_r(v) := \{v' \in \mathrm{v}[G]\,| d_G(v, v') = r\}\ .$$
- The *boundary* $\partial_G V$ in G of a set $V \subset \mathrm{v}[G]$ is
$$\partial_G V := \{\, v' \in \mathrm{v}[G] \setminus V \,|\, \exists v \in V : \, d_G(v, v') = 1\,\}.$$
The *closure* in G of $V \subset \mathrm{v}[G]$, $\overline{V}$, is given by $\overline{V} := V \dot\cup \partial_G V$.

In the sequel we write ∂ instead of ∂_G. We further introduce the r–th *factorial moment* of a positive, integer valued random variable $\hat{Y}$:

$$\mathbf{E}[\hat{Y}]_{\,r} := \sum_{j=r}^{\infty} [j]_r \, \boldsymbol{\mu}\{\, \hat{Y} = j\,\}$$

where $[j]_r = j \cdot (j-1) \cdot \ldots \cdot (j - (r-1))$.

Basic Model: *Let H be a finite graph and $\mathcal{G}(H)$ be the set of all induced subgraphs. We assume $0 \leq \lambda \leq 1$ to be given and set for $\Gamma \in \mathcal{G}(H)$*

$$\boldsymbol{\mu}_\lambda\{\Gamma\} \;:=\; \lambda^{|\,\mathrm{v}[\Gamma]\,|}\,(1-\lambda)^{|\,\mathrm{v}[H]\,| - |\,\mathrm{v}[\Gamma]\,|}\,.$$

Obviously $\boldsymbol{\mu}_\lambda$ fulfills $\sum_\Gamma \boldsymbol{\mu}_\lambda\{\Gamma\} = 1$ and we obtain choosing the set of all induced subgraphs as σ-field a probability space Ω.

2.1 Density

We shall discuss in this section the *density property* of random graphs $\Gamma_n < \mathcal{Q}_\alpha^n$, where $\mathcal{Q}_\alpha^n$ is a generalized hypercube, i.e., the graph formed by all n-tuples of coordinates x_i contained in a finite set $\mathcal{A}$ (of cardinality α) where each two tuples are neighbors when they differ in exactly one coordinate.
Let H be a finite graph. A subgraph $G < H$ is called *dense* in H if and only if $\overline{\mathrm{v}[G]} = \mathrm{v}[H]$.

We will establish the existence of a "critical" λ-value, λ^* that has the following property: for $\lambda < \lambda^*$ a.a.s. (asymptotically almost surely) no random graph Γ_n is dense and for $\lambda > \lambda^*$ a.a.s. every random graph Γ_n is dense. We will call λ^* the *threshold value for the density property.* For this purpose we consider the random variable

$$\hat{Z}_n(\Gamma_n) := |\{v \in \mathrm{v}[\mathcal{Q}_\alpha^n] \,|\, v \notin \overline{\mathrm{v}[\Gamma_n]}\}| \qquad (1)$$

that is defined on Ω_n and counts the number of vertices having no adjacent vertex $v \in \mathrm{v}[\Gamma_n]$. We first compute the asymptotic distribution of the following sequence of random variables $(\hat{Z}_n)$ associated to the sequence of probability spaces (Ω_n). For this purpose we make use of the *sieve formula* [2] (p.17) that implies a number of results about the convergence in distribution for a sequence of integer valued random variables $(\hat{X}_n)$.
Theorem 1. *Let $(\hat{X}_i)_{i\in\mathbb{N}}$ be a sequence of non-negative integer valued random variables such that*

$$\forall r \in \mathbb{N}: \lim_{n\to\infty} \mathbf{E}[\hat{X}_n]_r = \mathbf{E}[\hat{X}]_r$$

and

$$\forall m \in \mathbb{N}: \lim_{r\to\infty} \mathbf{E}[\hat{X}]_r r^m / r! = 0$$

Then we have the following convergence in distribution: $\hat{X}_n \longrightarrow \hat{X}$.
Proof. [2, p.23] ∎
Corollary 1. *Let $\mu = \mu(n)$ be a bounded, non–negative function on $\mathbb{N}$ and assume a sequence of non–negative integer valued random variables $(\hat{X}_i)_{i\in\mathbb{N}}$ to be given. Suppose for an arbitrary natural number r we have*

$$\lim_{n\to\infty} \mathbf{E}[\hat{X}_n]_r - \mu^r = 0\,.$$

Then the following convergence holds in distribution:

$$d(\hat{X}_n, P_\mu) \to 0\,,$$

where P_μ is the Poisson measure.
Lemma 1. *Suppose*

$$\mu := \lim_{n\to\infty} (|\mathcal{Q}_\alpha^n|\,(1-\lambda)^{\gamma_n+1}) \in \mathbb{R}_+ \cup \{0\} \cup \{\infty\}$$

exists. Then for $\mu < \infty$ the random variables $\hat{Z}_n$ converge in distribution to a Poisson distributed random variable, i.e.,

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n = \ell\} = \frac{\mu^\ell}{\ell!}\, e^{-\mu}\ . \qquad (2)$$

In particular we have

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n = 0\} = e^{-\mu},$$

and

$$\lim_{n\to\infty} \mathbf{E}[\hat{Z}_n] = \alpha^n\,(1-\lambda)^{\gamma_n+1}\,.$$

Finally, for $\mu = \infty$ and $\ell \in \mathbb{N}$ holds

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n \geq \ell\} = 1\,.$$

The following theorem shows that $\lambda^* := 1 - \sqrt[\alpha-1]{\alpha^{-1}}$ is a *threshold value* for the density property of random induced subgraphs as introduced in the basic model. Above λ^* a.a.s. all subgraphs are dense and below λ^* a.a.s. none of them.
Theorem 2. *Let $\lambda^* := 1 - \sqrt[\alpha-1]{\alpha^{-1}}$ then for $\lambda > \lambda^*$ holds*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \textit{ is dense in } \mathcal{Q}_\alpha^n\} = 1$$

and for $\lambda < \lambda^*$ we have

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \textit{ is dense in } \mathcal{Q}_\alpha^n\} = 0\,.$$

2.2 Connectivity and Giant Components

Let G be a finite graph. Being connected is an equivalence relation on $\mathrm{v}[G]$ and there exist maximal subsets $V \subset \mathrm{v}[G]$ consisting of connected vertices. A *component* of G is then the induced subgraph $G' = G[V]$ of such a maximal connected subset of vertices. If $V = \emptyset$, $G[\emptyset]$ is called a *trivial component.* If G is disconnected we shall investigate the so called *sequence of components*, i.e., the list of orders of the maximal connected subgraphs of G into which G can be decomposed.
Given a graph G, the *sequence of components* of G is the ordered tuple $(|\mathcal{X}_i|)_{1\leq i\leq |G|}$, where each $\mathcal{X}_i$ is a component of G and $|\mathcal{X}_i| \geq |\mathcal{X}_{i+1}|$. We call a component $\mathcal{X} < G$ a *giant component* if and only if $|\mathcal{X}| \geq \frac{2}{3}|G|$.
The key idea in the proof of the connectivity theorem bases on the following observation (formulated as lemma 2 below). A.a.s. *each* pair of vertices $v, v' \in \mathrm{v}[\Gamma_n]$ with $d(v,v') = k$, for fixed natural number k, is connected by a path in Γ_n.
For this purpose we refer to a certain family of independent paths in $\mathcal{Q}_\alpha^n$. I.e. for $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ with $d(v,v') = k$ we write v, v' as $v = (x_1, ..., x_k, x_{k+1}, ..., x_n)$ and $v' = (x_1', ..., x_k', x_{k+1}, ..., x_n)$. Then for $v_1 \in \partial\{v\} \cap B_{1+k}(v')$ we set

$$g_j(v_1) := (x_1, ..., x_j, x_{j+1}', ..., x_k', x_{k+1}, .., \hat{x}_r, ..., x_n)$$
$$0 \leq j \leq k \quad \hat{x}_r \neq x_r \qquad (3)$$

and inspect $g_k(v_1) = v_1$, $g_0(v_1) \in B_1(v') \cap B_{1+k}(v)$. We introduce the random variable $\hat{Y}_{n,k}^{v,v'}$ that counts the (independent) paths in the random graph Γ_n connecting the vertices v, v' having distance k.

Lemma 2. *Let k be a natural number, $\mathcal{Q}_\alpha^n$ a generalized hypercube and $\Gamma_n < \mathcal{Q}_\alpha^n$ a random graph with $\lambda > 1 - \sqrt[\alpha-1]{\alpha^{-1}}$. Then $\lim_{n\to\infty} \boldsymbol{\mu}_n\{T\} = 1$ where*

$$T := \{\Gamma_n \,|\, \forall v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n],\, d(v,v') = k : \exists\, v_1 \in \partial\{v\},\, v_1' \in \partial\{v'\} : \hat{Y}^{v_1,v_1'}_{n,d(v_1,v_1')} > 0\}\,.$$

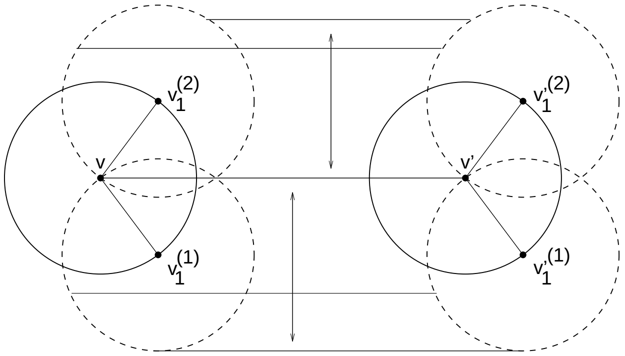

Fig. 1: An illustration for the proof of lemma 2. For given $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ each pair of vertices $(v_1^{(i)}, v_1'^{(i)})$ leads to "sufficiently many" independent pairwise disjoint paths in $\Pi(\mathcal{Q}_\alpha^n)$.

Now we are prepared to state the connectivity theorem:

Theorem 3. *Let $\mathcal{Q}_\alpha^n$ a generalized hypercube and $\Gamma_n < \mathcal{Q}_\alpha^n$ a random induced subgraph. Then*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is connected}\} = \begin{cases} 1 & \text{for} \quad \lambda > 1 - \sqrt[\alpha-1]{\alpha^{-1}} \\ 0 & \text{for} \quad \lambda < 1 - \sqrt[\alpha-1]{\alpha^{-1}}\,. \end{cases} \tag{4}$$

Remark. A related result in the special case of the Boolean hypercube can be found in [2]. The corresponding subgraphs A_p are constructed as follows: We set $\mathrm{v}[A_p] := \mathrm{v}[\mathcal{Q}_2^n]$ as vertex set and the edge set $\mathrm{e}[A_p]$ is obtained by independent random choices with probability p in the edge set $\mathrm{e}[\mathcal{Q}_\alpha^n]$. Then the idea of the proof is to establish an *edge boundary* of possible components using an *isoperimetric inequality* due to Harper, Bernstein, and Row [7, 2]. For Boolean hypercubes Ajtai, Komlós and Szemerédi 1982 proved the following related result: for random subgraphs A_p of $\mathcal{Q}_2^n$ obtained by edge selections, there exists a component of order $g\,2^n$ with constant $g \in \mathbb{R}_+$ if $p = c/n$ and $c > 1$ [1].

Finally, for any positive λ a.a.s. there exists a giant component in random induced subgraphs $\Gamma_n < \mathcal{Q}_\alpha^n$.

Theorem 4. *Let $0 < \lambda \le 1$. Then we have*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ has a giant component}\} = 1\,.$$

3. Sequence Structure Maps Via Random Graphs

3.1 Neutral Networks of Secondary Structures

A *pairing rule* Π is a symmetric relation in $\mathcal{A}\times\mathcal{A}$. Following [15] a *secondary structure* s w.r.t. Π is a vertex-labeled graph on n vertices $(x_1,\dots,x_n)$ with an adjacency matrix A fulfilling

(1) $a_{i,i+1} = 1$ for $1 \le i \le n-1$;
(2) For each i there is at most a single $k \neq i-1, i+1$ such that $a_{i,k} = 1$ and $[x_i, x_k] \in \Pi$;
(3) If $a_{i,j} = a_{k,l} = 1$ and $i < k < j$ then $i < l < j$.

We call an edge (x_i, x_k), $|i-k| \neq 1$ a *bond* or *base pair*. A vertex x_i connected only to x_{i-1} and x_{i+1} shall be called *unpaired*. The number of base pairs and the number of unpaired bases in a secondary structure s are $p(s)$ and $u(s)$, respectively. The size of the alphabet is α and the number of distinct base pairs is given by β.

We proceed by constructing the preimage of a fixed secondary structure as a random induced subgraph of the graph of compatible sequences. Let s be a secondary structure and

$$\Pi(s) := \{[i,k] \,|\, a_{i,k} = 1,\, k \neq i \pm 1\}$$

its *set of contacts*. Then a vertex $x \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ is said to be *compatible* to s if and only if $\forall [i,j] \in \Pi(s) : [x_i, x_j] \in \Pi$ i.e. the coordinates x_i and x_j are in Π for all pairs $[i,j] \in \Pi(s)$. We denote *the set of all compatible sequences* by $\mathbf{C}[s]$. Finally the *graph* of compatible sequences w.r.t. the secondary structure s, is given by

$$\mathcal{C}[s] := \mathcal{Q}_\alpha^{u(s)} \times \mathcal{Q}_\beta^{p(s)}\,.$$

Neutral Network: *Let $\Gamma_u < \mathcal{Q}_\alpha^u$ and $\Gamma_p < \mathcal{Q}_\beta^p$ be random subgraphs with underlying parameters λ_u and λ_p as introduced in the basic model. Then we set $\Gamma_n[s] = \Gamma_u \times \Gamma_p$,*

$$\boldsymbol{\mu}_{\lambda_u,\lambda_p}(\Gamma_n[s]) := \boldsymbol{\mu}_{u,\lambda_u}(\Gamma_u) \times \boldsymbol{\mu}_{p,\lambda_p}(\Gamma_p)\,,$$

and $\boldsymbol{\mu}_{\lambda_u,\lambda_p}$ is a probability measure and $\Gamma_n[s] < \mathcal{C}[s]$.

We can think of the neutral network, $\Gamma_n[s]$, to be obtained by selecting the coordinates v_1, v_2 of the vertex $(v_1, v_2) \in \mathrm{v}[\mathcal{C}[s]]$ with the probabilities λ_u and λ_p. This process leads to the vertex set $V_{\lambda_u,\lambda_p} \subset \mathbf{C}[s]$ whose induced subgraph $\mathcal{C}[s][V_{\lambda_u,\lambda_p}]$ is

$$\Gamma_n[s] = \Gamma_u \times \Gamma_p\,.$$

The theory presented in the previous chapter implies for neutral networks of secondary structures density and connectivity if both factors Γ_u and Γ_p are dense and connected.

3.2 Complete Mappings

Once we know how to construct a neutral network $\Gamma_n[s]$ we order the set of secondary structures $\mathcal{S}_n$ and define a complete mapping by iterating the construction process of the corresponding neutral network w.r.t. the ordering. Thereby we obtain w.r.t. a given λ parameter a complete sequence to structure mapping.
Let $\mathbf{C}^* : \mathcal{S}_n \to \mathcal{P}(\mathrm{v}[\mathcal{Q}^n_\alpha])$ and $r : \mathcal{S}_n \to \mathbb{N}$ be two mappings such that $j \le i \implies r(s_j) \ge r(s_i)$. A mapping $f : \mathcal{Q}^n_\alpha \to \mathcal{S}_n$ is called $\mathbf{C}^*$-map if and only if

$$(*) : \quad f(v) = s \quad \implies \quad v \in \mathbf{C}^*[s]\,.$$

A mapping $f_r : \mathcal{Q}^n_\alpha \to \mathcal{S}_n$ is called $\mathbf{C}^*$*-random-map* if and only if f_r is given by

$$f_r^{-1}(s_0) := \Gamma_n[s_0]$$
$$f_r^{-1}(s_i) := \Gamma_n[s_i] \setminus \bigcup_{j<i} [\Gamma_n[s_i] \cap \Gamma_n[s_j]]\ .$$

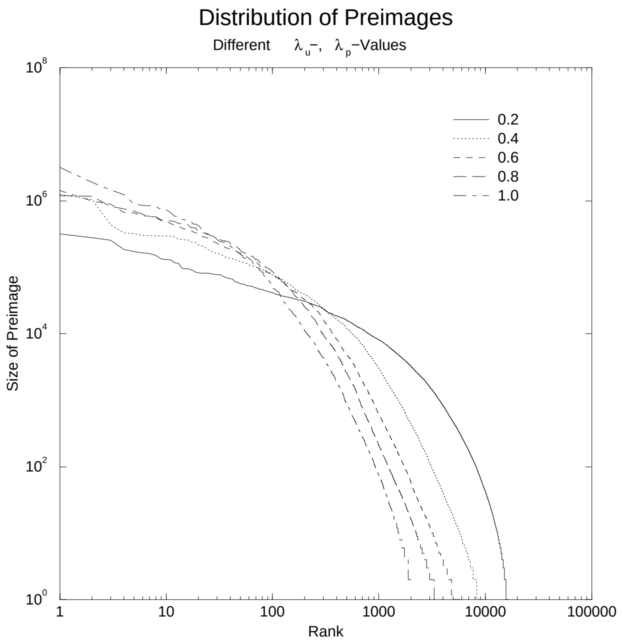

Fig. 2: We report here the logarithm of the sizes of the neutral networks $f^{-1}(s)$ obtained by a $\mathbf{C}^*$-random map and λ parameters $\lambda = 0.2, 0.4, 0.6, 0.8, 1.0$. The corresponding neutral networks are ordered on the x-axis by the logarithm of their orders. Note that the rank of the secondary structure s does not necessarily coincidence with the size of the corresponding preimage $|\, f^{-1}(s)\,|$.

In particular any RNA folding map is a $\mathbf{C}^*$-map setting $\mathbf{C}^*[s] := \mathbf{C}[s]$ since the neutral networks are constructed *a priori* in the graph of compatible sequences. We will assume that $\mathbf{C}^* = \mathbf{C}$ and then the recursion reads in the particular case of $\Gamma_n[s] = \mathbf{C}[s]$:

$$f^{-1}(s_{r_0}) := \mathbf{C}[s_0]$$
$$f^{-1}(s_{r_i}) := \mathbf{C}[s_i] \setminus \bigcup_{j<i} [\mathbf{C}[s_i] \cap \mathbf{C}[s_j]]$$

It turns out that the random maps in secondary structures as defined above have a few large preimages and many small ones as reported in figure 2. This observation fits in the computational results about RNA sequence structure maps that exhibit a characteristic rank order function known as a *generalized Zipf's Law*:

$$\psi(\imath) = a(1 + \imath/b)^{-c}\,,$$

as shown by extensive numerical calculations [13, 14]. Here a is a normalization constant, b is the number of frequent structures and c describes the power-law decay for rare structures.

4. Searching in Shape Space

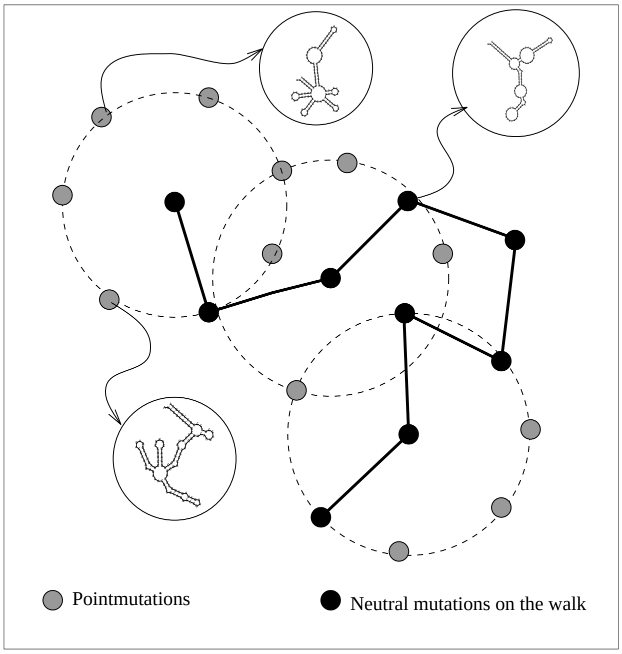

Fig. 3: A random walk on a neutral network. From each step (black) all point mutant sequences (grey) are formed and mapped into their corresponding secondary structures. Thereby, step by step, shape space is searched by w.r.t. one neutral network.

The time evolution of a population of asexually replicating molecules on a flat landscape can be described by a random walk. On more complex fitness landscapes a selective pressure drives the population towards sites of higher fitness. In the simplest case one may think of one or few fit structures. Then the populations searches "along one neutral network" [4] until sequences of another network corresponding to a fitter structure are found (see fig. 3).

The key question [10] is: to what extend is the shape space searched by a population diffusing [11] on a fixed neutral network? Here the following result [12, 11] gives some insight, implying that *any* two neutral networks that are dense and have a giant component come close in sequence space and therefore allow for transitions from one network to the other [4]. Practically *all* other structures are found forming one error mutants with respect to a fixed neutral net. Explicitly the result reads

Theorem 5. *[Intersection–Theorem] Let* Π *be a non-empty pairing rule on* $\mathcal{A}$ *and* s *and* s' *be arbitrary (non-empty) secondary structures. Then we have w.r.t.* Π

$$\mathbf{C}[s] \cap \mathbf{C}[s'] \neq \emptyset\,.$$

The topology of neutral networks (following the predictions of random graph theory) plays therefore a crucial role in the optimization process. Only the existence of a giant component in the net and its density guarantee that any other net can be reached. In the tables shown in appendix A, we report the existence of those giant components.

Finally we discuss the so called *search capacity* of our mappings. This means to a fixed λ parameter we compute the size of the preimage of those secondary structures that are found by a random walk of the particular (fixed) neutral network. Passing the critical parameter $\lambda = 0.5$ for density and connectivity, the shape space is searched effectively by a random walk on the corresponding neutral network. For the small sequence length considered in the computer experiments, it turns out that below $\lambda = 0.5$ it is difficult to perform a random walk on the network lasting sufficiently many steps.

Fractions of neutral neighbors can be determined numerically by RNA folding based on sequences over an **AUGC** alphabets [3, 9, 8]. These fractions turn out to be characteristically above the critical values derived from random graph theory, and therefore indicate how well suited nature is for optimization on secondary structures.

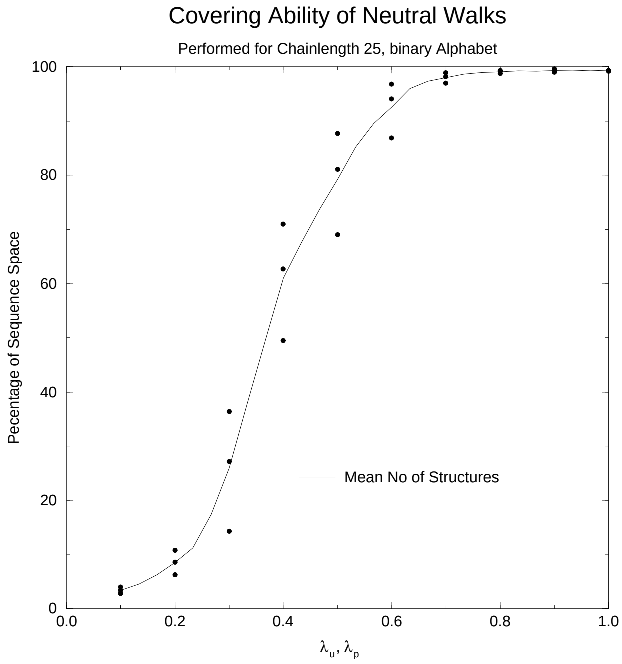

Fig. 4: We compute the percentage of the preimage of those secondary structures that are found by a random walk of one particular (fixed) neutral network. "Found" means that from each sequence realized in the walk we map all sequences of Hamming distance one and thereby obtain a stepwise increasing family of structures realized w.r.t. the random walk on the chosen network.

Acknowledgments

We want to thank Mr. F. Haubensak for the hints and support he gave us to create the two columns per page.

References

[1] Ajtai, Komlós, and Szemerédi. Largest random component of a k-cube. *Combinatorica*, 2:1 – 7, 1982.

[2] B. Bollobás. *Random Graphs.* ACADEMIC PRESS, 1985.

[3] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47(3):2083 – 2099, March 1993.

[4] C. V. Forst, C. Reidys and J. Weber. Evolutionary Dynamics and Optimization: Neutral Networks as Model-Landscapes for RNA Secondary-Stucture Folding-Landscapes. *Advances in Artificial Life*, 1995.

[5] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and

P. Schuster. Analysis of rna sequence structure maps by exhaustive enumeration I. neutral networks. *Monatshefte fuer Chemie*, 127:*–*, 1996.

[6] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of rna sequence structure maps by exhaustive enumeration II. structures of neutral networks and shape space covering. *Monatshefte fuer Chemie*, 127:*–*, 1996.

[7] L. Harper. Minimal numberings and isoperimetric problems on cubes. *Theory of Graphs, International Symposium, Rome*, 1966.

[8] I. L. Hofacker. *A Statistical Characterisation of the Sequence to Structure Mapping in RNA*. PhD thesis, University of Vienna, 1994.

[9] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, 125(2):167–188, 1994.

[10] M. Huynen, P. F. Stadler, and W. Fontana. Evolutionary dynamics of RNA and the neutral theory. *PNAS*, 1994. accepted.

[11] C. Reidys. *Neutral Networks of RNA Secondary Structures*. PhD thesis, Friedrich Schiller Universität, Jena, May 1995.

[12] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps and neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 1995. submitted.

[13] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc.Roy.Soc.(London)B*, 255:279–284, 1994.

[14] M. Tacker, W. Fontana, P. Stadler, and P. Schuster. Statistics of RNA melting kinetics. *Eur. Biophys. J.*, 23(1):29 – 38, 1994.

[15] M. S. Waterman. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics sup-*

Appendix A

Rank	$\lvert\Gamma\rvert$	Structure	μ_u	μ_p	Sequence of Components
1	1494359	.......((....((.....)).))	0.823	0.854	1494359
2	1471814	.((.((.....))...)).......	0.795	0.858	1471810, 4×1
3	1180817	(((......))).............	0.599	0.487	1180665, 2, 150×1
4	1131781	..((..((.........)).))...	0.736	0.723	1131775, 6×1
5	895743	....((......))....((...))	0.726	0.499	829839, 65886, 2, 16×1
6	811332	..(((((.........)))..))..	0.862	0.871	811332
7	769357	..(((.......)))((.....)).	0.852	0.825	769357
8	764731	..((.....(((...)))..))...	0.880	0.826	764731
9	676878	....((...)).((.......))..	0.608	0.501	676783, 95×1
10	577625	((.......)).(((......))).	0.740	0.723	577616, 9×1

Tab. 1: Mapping parameters $\lambda_u = \lambda_p = 0.9$, μ_u, μ_p are the percentages of unpaired and paired segments of sequences contained in the neutral net.

Rank	$\lvert\Gamma\rvert$	Structure	μ_u	μ_p	Sequence of Components
1	3466927	.......((......))........	0.567	0.618	3466927
2	1339085	...........((...)).......	0.409	0.306	1337547, 2× 3, 30× 2, 1472×1
3	718788	....((..((......))..))...	0.514	0.576	718747, 2, 39×1
4	650290	....(((....)))...........	0.357	0.305	646820, 6, 5, 14× 3, 165× 2, 3087×1
5	642094	((....((.........)).))...	0.500	0.513	642025, 2, 67×1
6	606699	.((.((......))..)).......	0.456	0.553	606464, 6× 2, 223×1
7	596554	..((.....))..............	0.230	0.237	570539, 10, 3× 7, 7× 6, 20× 5,
					56× 4, 277× 3, 1618× 2, 21551×1
8	575245	.........(((........)))..	0.313	0.336	569428, 6× 4, 25× 3, 280× 2, 5158×1
9	500107	.((...((...))......))....	0.456	0.393	499841, 5× 2, 256×1
10	447051	..........((.....))......	0.333	0.000	443329, 4, 15× 3, 122× 2, 3429×1

Tab. 2: Mapping parameters $\lambda_u = \lambda_p = 0.6$ (see also caption of tab. 1)

Rank	$\|\Gamma\|$	Structure	μ_u	μ_p	Sequence of Components
1	1604053	.......((......))........	0.387	0.481	1604009, 18$\times$ 2, 8$\times$1
2	1011867	..........((.....))......	0.279	0.349	1008832, 4, 6$\times$ 3, 169$\times$ 2, 2675$\times$1
3	649324	..((.....))..............	0.209	0.000	327019, 301123, 2$\times$ 8, 2$\times$ 7, 4$\times$ 6,
					9$\times$ 5, 53$\times$ 4, 215$\times$ 3, 1299$\times$ 2,
					17628$\times$1
4	511011	.....(((...........)))...	0.244	0.311	506282, 2$\times$ 6, 5, 8$\times$ 4, 41$\times$ 3,
					274$\times$ 2, 4009$\times$1
5	484477	.((...((...))......))....	0.351	0.555	484319, 3$\times$ 11, 2$\times$ 10, 8, 2$\times$ 7,
					6, 2$\times$ 2, 73$\times$1
6	471135	.........(((........)))..	0.270	0.219	359724, 107579, 9, 5, 8$\times$ 4,
					29$\times$ 3, 206$\times$ 2, 3287$\times$1
7	440221	.........((.....)).......	0.170	0.193	413333, 9, 8, 4$\times$ 7, 9$\times$ 6,
					29$\times$ 5, 143$\times$ 4, 411$\times$ 3, 2060$\times$ 2,
					20719$\times$1
8	375080	....((..((......))..))...	0.368	0.467	330454, 44442, 7$\times$ 7, 4$\times$ 6, 111$\times$1
9	363837	......(((......)))).......	0.272	0.181	274624, 85356, 6, 5, 8$\times$ 4,
					41$\times$ 3, 212$\times$ 2, 3267$\times$1
10	315083	((.((........)).)).......	0.248	0.443	313124, 2$\times$ 7, 3$\times$ 5, 4, 13$\times$ 3,
					97$\times$ 2, 1693$\times$1

Tab. 3: Mapping parameters $\lambda_u = \lambda_p = 0.4$ (see also caption of tab. 1)