Title: MAPPING IN RANDOM-STRUCTURES

Author(s): Christian M. Reidys

Submitted to: Artificial Life V
Nara, Japan
May 16-18, 1996
6.5.1996. Last Changes by Duck

Mapping in random-structures

submitted to SIAM journal of discrete mathematics

By

Los Alamos National Laboratory
TSA-DO/SA
NM 87545, USA

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

*Los Alamos National Laboratory
TSA-DO/SA
NM 87545, USA

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

Mailing Address:
Santa Fe Institute
1399 Hyde Park Rd., Santa Fe, NM 87501, USA
Phone: 001 505-984-8808 Fax: 001 505-982-0565
E-Mail: duck@santafe.edu

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
# Table of Contents

1. Introduction  
2. Random-structures  
3. Preimages of random-structures  
4. Applications  
Bibliography
Abstract: A mapping in random-structures is defined on the vertices of a generalized hypercube \(Q^n\). A random-structure will consist of (i) a random contact graph and (ii) a family of relations imposed on adjacent vertices. The vertex set of a random contact graph will be the set of all coordinates of a vertex \(P \in Q^n\). Its edge will be the union of the edge sets of two random graphs. The first is a random 1-regular graph on \(2m\) vertices (coordinates) and the second is a random graph \(G_p\) with \(p = \frac{2}{n}\) on all \(n\) vertices (coordinates). The structure of the random contact graphs will be investigated and it will be shown that for certain values of \(m, c_2\) the mapping in random-structures allows to search by the set of random-structures. The results will finally be applied to evolutionary optimization of biopolymers.

Key words. random-structure, sequence-structure mapping, random graph, connectivity, branching process, optimization

AMS subject classifications. 05C38, 05C40, 05C80, 05C90

1. Introduction

The folding of sequences into their (spatial) structures is of central interest in biophysics. Its properties are of essential importance for the understanding of evolutionary optimization of biopolymers and the theory of molecular evolution. Whereas in biophysics structure is defined in terms of some physical conditions, like minimum free energy, in this paper random-structures are introduced. A random-structure consists of (i) a random contact graph, \(\Gamma_{m,c_2}^n\), and (ii) a family of relations that specify the correlations of adjacent vertices. The contact graph is a graph on the \(n\) coordinates of a vertex \(V \in Q^n\). Its edges set is the union of the edges sets of two random graphs, a 1-regular graph on \(2m\) vertices, where \(2m \leq n\), and a random graph \(G_p\) on the \(n\) vertices with \(p = \frac{2}{n}\). The structure of the random contact graphs will be investigated and the implications for the corresponding mappings in random-structures will be discussed. Finally the results will be applied on evolutionary optimization of biopolymers.

2. Random-structures

A graph \(X\) consists of a tuple \((v[X], e[X])\) and two maps

\[
\begin{align*}
\sigma \times t : e[X] &\rightarrow v[X] \times v[X] \\
\text{inv} : e[X] &\rightarrow e[X]
\end{align*}
\]

\(e \rightarrow (\sigma(e), t(e))\)  \(e \rightarrow \text{inv}(e)\)
satisfying the following conditions: for each \( e \in e[X] \) we have \( \text{inv}^2(e) = e, \text{inv}(e) \neq e \) and \( t(e) = o(\text{inv}(e)) \). The set \( \{e, \text{inv}(e)\} \) is called geometric edge. \( v[X] \) is called vertex set and \( e[X] \) edge set. An element \( P \in v[X] \) is called vertex of \( X \); an element \( y \in e[X] \) is called an (oriented) edge, and \( \text{inv}(y) \) is called inverse edge. The vertex \( o(y) = t(\text{inv}(y)) \) is called origin of \( y \) and the vertex \( t(y) = o(\text{inv}(y)) \) is called the terminus of \( y \); \( o(y), t(y) \) are called the extremities of some edge \( y \).

There is an obvious notion of a graph-morphism and if \( Y \rightarrow X \) is an embedding we call \( Y \) an subgraph of \( X \). We call a subgraph \( Y \) induced if for any \( P, P' \in Y \) being extremities of an edge \( y \in X \) it follows \( y \in Y \). A graph \( X \) with the property \( e[X] + 1 = v[X] \) is called a tree, a tree in which each \( P \in X \) has at most two adjacent vertices is called a line-graph. A graph \( X \) in which each vertex \( P \in X \) has exactly two adjacent vertices is called cycle. A path in \( X \) is a morphism of a line graph into \( X \) and \( X \) is called connected if any two vertices occur in a path of \( X \). Being connected is an equivalence relation in \( X \) and the maximal connected subsets of vertices are called components of \( X \). Let \( Y \) be a subgraph of \( X \) then the closure of \( Y \) in \( X \), \( \overline{Y} \), is the induced subgraph of all extremities of the edges \( e[Y] \). A subgraph \( Y \rightarrow X \) is called dense if and only if \( \overline{Y} = X \). Finally, a vertex \( P \) is called isolated in \( X \) if it is not an extremity to an edge \( y \in e[X] \).

Let \( 1 \geq c_1, c_2 > 0 \) be positive constants and \( n \) a natural number. Suppose further that \( m(n) \) is a natural number such that \( 2m(n)/n \) is an isotone sequence with limit \( \lim_{n \rightarrow \infty} \frac{2m}{n} = c_1 \). Further let \( X_1 \) be a 1-regular graph on \( 2m \) vertices \( \{P_1, \ldots, P_{2m}\} \subset \{P_1, \ldots, P_n\} \) and \( X_2 \) a subgraph of the complete graph on \( \{P_1, \ldots, P_n\}, K_n \).

There exist \( L_1(2m, n) = \frac{1}{m!} \prod_{k=0}^{m-1} \left( \frac{n-2k}{2} \right) \) different 1-regular graphs. The graphs \( X_1 \) form a finite probability space by assigning to each 1-regular graph uniform probability. Analogously, the graphs \( X_2 \) form a finite probability space consisting of all subgraphs of \( K_n \) where each subgraph with \( k \) edges has probability \( p^k(1 - p)^{(\frac{n}{2}) - k} \) with \( p = c_2/n \) [4]. This probability space will be referred to as \( G_p \).

The graphs \( X_1, X_2 \) induce the graph \( X_1 \cup X_2 \) whose vertex and edge set are given by \( v[X_1 \cup X_2] := v[X_2], e[X_1 \cup X_2] := e[X_1] \cup e[X_2] \) having induced mappings

\[
\begin{align*}
\text{o} \times t : e[X_1 \cup X_2] & \rightarrow v[X_1 \cup X_2] \times v[X_1 \cup X_2] \quad y \rightarrow (o(y), t(y)) \\
\text{inv} : e[X_1 \cup X_2] & \rightarrow e[X_1 \cup X_2] \quad y \rightarrow \text{inv}(y)
\end{align*}
\]

Clearly there are different pairs \( X_1, X_2 \) leading to \( X_1 \cup X_2 \) i.e. \( (X_1, X_2) \mapsto X_1 \cup X_2 \) is not injective. I will call the probability space consisting of graphs \( X_1 \cup X_2 \) with underlying measure \( \mu \{X_1 \cup X_2\} = \sum_{(x_1, x_2) \rightarrow X_1 \cup X_2} \mu \{X_1\} \mu \{X_2\} \) the random contact graph \( \Gamma_{m, c_2} \).
Theorem 1. Suppose that $0 \leq c_2, c_1 < 1$ and $\frac{2m(n)}{n} \not< c_1$. Further let $\bar{T}$ be the r.v. counting the number of vertices of a random graph $\Gamma_{m, c_2}^n$ that are contained in tree-components. Then the following assertions hold

(i) For $[c_1 + c_2] < 1$ asymptotically almost all vertices of $\Gamma_{m, c_2}^n$ are contained in tree-components, i.e.

$$\lim_{n \to \infty} \left[ \frac{E[\bar{T}]}{n} \right] = 1$$

Moreover there exists a constant $C(c_1, c_2) > 0$ such that a.a.s. all paths in $\Gamma_{m, c_2}^n$ have length $\leq C \ln(n)$.

(ii) For $c_2 < \frac{1}{4}$ and arbitrary $c_1$ there exists a constant $C(c_2) > 0$ such that a.a.s. all tree-components in $\Gamma_{m, c_2}^n$ have the property $|T| \leq C \ln(n)$.

Proof. (i) Suppose $P_1$ is a vertex of $\Gamma_{m, c_2}^n$ that belongs to a component containing a cycle. This statement is equivalent to the existence of a path $P_1, \ldots, P_h$ in $\Gamma_{m, c_2}^n$ where $h \in \mathbb{N}$ such that there exists an edge $y$ in $\Gamma_{m, c_2}^n$ with the property $t(y) = P_h$ and $o(y) = P_j$ for $1 \leq j \leq h - 2$. These paths will be referred to as $*$-paths (as introduced by B.Pittel). The r.v.'s whose states are the number of $*$-paths of length $h$ and of length $h$ having $\ell$ $X_1$-edges will be referred to as $Z_h^n$ and $Z_{\ell, h}^n$, respectively.

Claim: $\lim_{n \to \infty} E[Z_h^n(\Gamma_{m, c_2}^n)] \leq h[c_1 + c_2]^h$.

We first consider $*$-paths $P_1, \ldots, P_h$ in $\Gamma_{m, c_2}^n$ having exactly $\ell$ edges from $X_1$. The expectation value, $E[Z_{\ell, h}^n]$, has the upper bound

$$\forall h \geq 3: \quad E[Z_{\ell, h}^n] \leq h \left( \frac{h}{\ell} \right) \left( \frac{2m}{n} \right)^\ell c_2^{1-\ell}.$$ 

To show this we first conclude by induction that the probability for having $\ell$ specific $X_1$-edges in $\Gamma_{m, c_2}^n$ fulfills

$$\frac{L_1(2(m - \ell), n - 2\ell)}{L_1(2m, n)} = \frac{m!}{(m - \ell)!} \prod_{k=0}^{m-\ell-1} \frac{(n-2\ell-2k)}{2} \leq \left[ \frac{2m}{n} \right]^{\ell} \frac{1}{n(n-2) \ldots (n-2(\ell-1))}.$$ 

Obviously there are at most $(n)_{2\ell}$ ordered $\ell$-tuples of $X_1$-edges and those are incident to exactly $2\ell$ vertices of the $*$-path. Hence the expectation value for having $\ell$ $X_1$-edges in a $*$-path of $\Gamma_{m, c_2}^n$ has the upper bound

$$(n)_{2\ell} \frac{L_1(2m - 2\ell, n - 2\ell)}{L_1(2m, n)} \leq \left[ \frac{2m}{n} \right]^\ell n^{\ell}.$$
The remaining edges belong to $X_2$ and are selected with probability $\left[ \frac{2m}{n} \right]^{h-\ell}$. The $X_1$-edges fix $2\ell$ vertices and there are at most $n^{h-2\ell}$ different choices for the $X_2$-edges. Finally there are $h-2$ possibilities for taking a vertex $P_j$, $1 \leq j \leq h-2$ and we obtain

$$E[Z_{\ell,n}] \leq \left[ \frac{2m}{n} \right]^{\ell} \binom{h}{\ell} n^\ell \left[ \frac{c_2}{n} \right]^{h-\ell} (h-2) n^{h-2\ell} \leq h \left( \frac{2m}{n} \right)^{\ell} \binom{h}{\ell} c_2^{h-\ell}.$$ 

Next, using for $0 < \alpha < 1$: $\sum_{3 \leq h} h x^h \leq \left[ \frac{\alpha}{(1-x)^2} \right]$ we take the sum over all $1 \leq \ell \leq h$, $h \geq 3$ and derive

$$\sum_{3 \leq h, 1 \leq \ell \leq h} E[Z_{\ell,h}] \leq \sum_{h \geq 3} h \left( \frac{2m}{n} + c_2 \right)^h \leq \frac{\left( \frac{2m}{n} + c_2 \right)}{(1 - \left( \frac{2m}{n} + c_2 \right))^2}.$$ 

Therefore taking the limit, the expected number of vertices of a random graph $\Gamma_{m,c_2}$ contained in components having cycles is finite i.e. $\lim_{n \to \infty} E[\bar{T}] / n = 1$.

In complete analogy to the above argument the expected number of paths of length $h$ in $\Gamma_{m,c_2}$ has the upper bound

$$\lim_{n \to \infty} E[P_h] \leq \lim_{n \to \infty} (n-h)[c_1 + c_2]^h.$$ 

For $h' := C \ln(n)$ we obtain $\lim_{n \to \infty} E[P_{h'}] \leq \lim_{n \to \infty} n^{1+C\ln(c_1+c_2)}$ and

$$h \geq h' \implies \lim_{n \to \infty} E[P_h] \leq \lim_{n \to \infty} E[P_{h'}],$$

whence $\sum_{h \geq h'} \lim_{n \to \infty} E[P_h] \leq \lim_{n \to \infty} n^{2+C\ln(c_1+c_2)}$. Clearly, we can choose $C(c_1, c_2)$ such that $2 + C \ln(c_1 + c_2) < 0$ proving that the expectation value of the number of paths of length $\geq C \ln(n)$ tends to 0 whence (i).

(ii) We consider a branching process in $\Gamma_{m,c_2}$ following Karp [14]. Let $N_t, L_t$ and $D_t$ be subsets of vertices being neutral, alive and dead vertices of corresponding sizes $n_t, l_t, d_t$. To initialize the process we select an initial vertex $P$, set $L_0 := \{P\}$, $N_0 := \{x_1, \ldots, x_n\} \setminus \{P\}$ and $D_0 := \emptyset$. The update from time $t-1$ to time $t$ is done as follows: we pick a vertex $P \in X_{t-1}$ and select all edges $y \in \Gamma_{m,c_2}$ such that $o(y) = P$ and $t(y) \in N_t$. The r.v. counting the number of vertices $t(y)$ at time $t$ will be called $z_t$. Then we set all vertices of the form $t(y)$ life and set $P$ dead. Obviously the process ends if $L_t = \emptyset$.

Suppose now the vertices $P' \in L_t$ and $P'' \in N_t$ are extremities of two distinct $X_1$-edges, $y', y''$. Under this assumption there are exactly 4 edges connecting $y', y''$ and those are checked in two
time steps of the process. The probability that at least one of them is selected as $X_2$-edge reads $\frac{4c_2}{n}$.

Thus it suffices to show that the branching process stops after logarithmic time on the random graph $G_{4c_2}$. In this situation $n_t$ is $B(n - 1, (1 - \frac{4c_2}{n})^t)$ distributed since each vertex different from $P$ has probability $(1 - \frac{4c_2}{n})^t$ to stay neutral $t$-times. By construction no edge can be selected twice. Clearly $n_t = n - l_t - t$ and we observe that $l_t > 0$ is equivalent to $n_t > n - t$, whence $\mu\{l_t > 0\} = \mu\{n_t > n - t\}$. Further we observe for $T$ being the stopping time of the process that $\mu\{T > t\} \leq \mu\{\sum_{t=0}^{t-1} z_t - t > 0\}$ and

$$\lim_{n \to \infty} \left[ \frac{B(n - (t - 1) - L_t - 1, \frac{4c_2}{n}, k)}{B(n, \frac{4c_2}{n}, k)} \right] = 1.$$ 

Hence for fixed $i$, $z_i$ is asymptotically Poisson i.e. $\lim_{n \to \infty} \mu\{z_i = h\} = (\frac{4c_2}{n})^h e^{-4c_2}$. Accordingly, $\sum_{i=0}^{t-1} z_i$ is Poisson with mean $\mu = t \cdot 4c_2$. Since $4c_2 < 1$ we have

$$\mu\{\sum_{i=0}^{t-1} z_i - t > 0\} = \mu\{\sum_{i=0}^{t-1} z_i - t \cdot 4c_2 > \frac{1 - 4c_2}{2c_2}[t \cdot 4c_2]\}$$

and for any sum of mutually independent indicator r.v.'s, $z$, with mean $E[z] = \mu$ holds [2], p.239 (see e.g. [5])

$$\text{For } \alpha > 0 : \quad \mu\{|z - \mu| \geq \alpha \mu\} \leq 2e^{-\alpha^2 \mu}.$$ 

where $c_\alpha > 0$. Consequently there exists a constant $K > 0$ such that $\mu\{T > t\} < e^{-Kt}$. In particular, if we choose $C = C(c_2)$ such that $CK > 1$, the inequality reads $\mu\{T > C \ln n\} < n^{-CK}$. There are at most $n$ choices for the initial vertex whence the expectation value of components of size $\geq C\ln(n)$ tends to zero and (ii) follows.

Referring to an induced subtree of $\Gamma_{n, c_2}$ as $T$, a random-structure $s_n$ consists of the following pieces of data:

(i) the graphs $X_1, X_2$

(ii) a family of symmetric relations $(R_x, R_y)_{y \in X_2 \setminus X_1}$, where $R_x, R_y \subset A \times A$ for $y \in e[T] < \Gamma_{n, c_2}$ and each $R_y$ is supposed to fulfill $(+)$ $\forall a \in A \exists b \in A : a R_y b$
Remark. The relation $R_y$ is motivated by Watson and Crick base-pairing rules observed in RNA secondary-structures and for $y \in T \cap X_2 \setminus X_1$ the relation $R_y$ is a specific (tertiary) interaction. Here I only assume that w.r.t. $R_y$ there exists to any vertex at least one in $R_y$.

A vertex $V \in Q_n^a$ is called compatible with respect to $s_n$ if and only if

$$\forall y \in e[X_1] \wedge y \in T : o(y)R_1t(y)$$

$$\forall y \in X_2 \setminus X_1 \wedge y \in T : o(y)R_yt(y).$$

Theorem 1 implies that a.a.s. there exists at least one vertex in $Q_n^a$ that is compatible to a given random-structure. By construction there are $n - 2m$ vertices not incident to an $X_1$-edge and we obtain asymptotically $[n - 2m]e^{-c_2}$ isolated vertices in $\Gamma_{m,c_2}^n$.

3. Preimages of random-structures

The preimage of a particular random-structure, $s_n$, is clearly contained in the set of compatible vertices $C(s_n)$. In fact it will be introduced as an random induced subgraph $\Gamma_n$ of $Q_n^a$ with underlying measure $\mu_{\lambda}(\Gamma_n) := \lambda|\Gamma_n|(1 - \lambda)|C(s_n)|^{-1}$. Hence the preimages form a finite probability space $({\{\Gamma_n\}, \mu_{\lambda}})$

Each vertex $V$ in $Q_n^a$ has asymptotically a mean of $[n - 2m]e^{-c_2}$ isolated vertices in the contact graph of the underlying random-structure $s_n(V)$. There is a projection on the $n$ isolated vertices $\{P_{j_1}, \ldots, P_{j_n}\}$ in the contact graph of $s_n(V)$, $\pi : Q_n^a \to Q_a^n$. Clearly $\pi(\Gamma_n)$ is a random induced subgraph of $Q_n^a$ and exhibits a threshold value concerning its density and its connectivity property c.p. [1, 4].

Definition 1. Let $\lambda^* := 1 - \frac{\alpha}{\sqrt[3]{n} - \frac{1}{3}}$ then for $\lambda > \lambda^*$ holds

$$\lim_{n \to \infty} \mu_n\{\Gamma_n \text{ is dense in } Q_n^a\} = 1$$

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

$$\lim_{n \to \infty} \mu_n\{\Gamma_n \text{ is dense in } Q_n^a\} = 0.$$

A proof can be found in [16, 17]. We next show that $\lambda^*$ is also a threshold value for the connectivity property of random induced subgraphs of $Q_a^n$. 
Definition 2. Let $Q^n_\alpha$ be a generalized hypercube and $\Gamma_n < Q^n_\alpha$ a random induced subgraph. Then

$$\lim_{n \to \infty} \mu_n\{\Gamma_n \text{ is connected}\} = \begin{cases} 1 & \text{for } \lambda > 1 - \frac{\alpha}{\sqrt{\alpha - 1}} \\ 0 & \text{for } \lambda < 1 - \frac{\alpha}{\sqrt{\alpha - 1}}. \end{cases}$$

(1)

A proof can be found in [16, 17] where it is shown that the number of isolated vertices in random graphs $\Gamma_n$ is Poisson with mean $\mu := \alpha^n(1 - \lambda)^{o(n)}$.

We obtain mappings $f : Q^n_\alpha \to \{s_n\}$ by constructing the corresponding preimages as random graphs iteratively: we fix a mapping $r : \{s_n\} \to \mathbb{N}$ having the property $j \leq i \implies r(s_j) \geq r(s_i)$ and set

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

In this section we will consider how preimages of random-structures are embedded in $Q^n_\alpha$. For this purpose we analyze the graph $\Gamma_{m,c_2}^n \cup \Gamma_{m,c_2}^n$.

In $\Gamma_{m,c_2}^n \cup \Gamma_{m,c_2}^n$ let now $C_1$ be the number of cycles composed by $X_1, X_1'$-edges and $\hat{C}_1$ the number of vertices contained in those. Suppose further that $S_n$ is the symmetric group over the set $\{P_1, \ldots, P_n\}$. We consider pairs of indices of incident vertices w.r.t. edges $y \in X_1, (i, k)(y)$, as transpositions $\tau(y) = (i, k)(y)$ and obtain the embedding

$$1 : \{X_1\} \to \quad S_n \quad X_1 \to \prod_{y \in X_1} (i, k)(y).$$

Theorem 2. Suppose $\Gamma_{m,c_2}^n$ and $\Gamma_{m,c_2}^n$ are two random contact-graphs such that $\lim_{n \to \infty} \frac{2m}{n-1} = c_1 > 0$ and $0 \leq c_2 \leq 1$. Then the following assertions hold:

(i) For $c_1 < 1$ and $c_2 = 0$ asymptotically almost all vertices of $\Gamma_{m,c_2}^n \cup \Gamma_{m,c_2}^n$ are contained in components that are line-graphs. There exists a constant $C > 0$ with the property that a.a.s.

all components in $\Gamma_{m,c_2}^n \cup \Gamma_{m,c_2}^n$ have lengths $\leq C \ln(n)$.

(ii) Suppose $8c_1[2 - c_1]c_2 > 1$ and $\xi \neq 1$ solves $(1 - x) = e^{8c_1(2-c_1)c_2 x}$. Then $\Gamma_{m,c_2}^n \cup \Gamma_{m,c_2}^n$ has a.a.s. components $C^n$ with the property $|C^n| \geq (1 - x)n^{\frac{4m}{n}} - (2m)^2n$. 


Proof. (i): We have already observed that any two 1-regular subgraphs $X_1, X'_1$ correspond uniquely to two involutions and the latter form a dihedral group. The way this group acts upon the vertices implies that all cycles composed of $X_1, X'_1$-edges have even order and $h \geq 2$.

Now we show that for $c_1 < 1$ and $c_2 = 0$, asymptotically, the expected number of vertices in a random graph $\Gamma_{m,c_2}^n \cup \Gamma_{m,c_2}^n$ contained in cycles composed by $X_1, X'_1$-edges is finite.

To see this we only have to compute the expected value of $s$-paths (cf. the proof of theorem 1) of length $2h$, composed exclusively by $X_1$- and $X'_1$-edges. The probability of selecting $h$ $X_1$-edges in a random contact graph $\Gamma_{m,c_2}^n$ reads $\prod_{k=0}^{h-1} \left( \frac{n-k}{n} \right)$. If a cycle is composed by $X_1$- and $X'_1$-edges the vertices incident to the above $X_1$-edges determine already the corresponding $h$ $X'_1$-edges completely. Analogously the probability to select these $X'_1$-edges reads $\prod_{k=0}^{h-1} \left( \frac{n-k}{n} \right)$. There are $(n)^{2h+1}$ different selections for the edges of a path of length $2h$ and each path can begin with an $X_1$- or an $X'_1$-edge, whence

$$E[C_1[2h]] \leq 2 \frac{(m)^{2h}}{(n)^{2h}} \Rightarrow \lim_{n \to \infty} E[C_1[2h]] \leq 2 c_1^{2h}.$$ 

We immediately observe that $\lim_{n \to \infty} \sum_{h \geq 1} E[C_1[2h]] < \infty$ and $\lim_{n \to \infty} \sum_{h \geq 1} 2h E[C_1[2h]] < \infty$. Therefore, asymptotically, only a finite number of vertices is contained in cycles composed by $X_1, X'_1$-edges.

Analogously we deduce that the expected value of paths composed by $X_1$- and $X'_1$-edges of length $d$ has asymptotically the upper bound

$$\lim_{n \to \infty} E[P_d] \leq \lim_{n \to \infty} 2 n c_1^d$$

and consequently for $d := C \ln(n)$ $\lim_{n \to \infty} E[P_d] \leq \lim_{n \to \infty} 2 n^{1+C \ln(c_1)}$. Obviously $d > d'$ implies $\lim_{n \to \infty} E[P_{d'}] \leq \lim_{n \to \infty} E[P_d]$ whence $\lim_{n \to \infty} \sum_{h \geq 1} 2h \ln(n) E[P_d] \leq \lim_{n \to \infty} 2 n^{2+C \ln(c_1)}$. Then choosing $C$ such that $2 + C \ln(c_1) < 0$ implies that the corresponding expected number of paths of length $\geq C \ln(n)$ tends to 0. In particular the expected number of components (i.e. line-graphs) of finite size $k + 1$ formed by $X_1$ and $X'_1$-edges is bounded by $\left[ \frac{2m}{n} \right]^k$.

(ii) We consider a branching process on the vertices of $X_1 \cup X'_1$. First each vertex $P \in X_2$ is not contained in $X_1 \cup X'_1$ with probability $(1 - \frac{2m}{n})^2$ whence $\mu = 4m - n(\frac{2m}{n})^2$ is the mean number of vertices contained in $X_1 \cup X'_1$. Let now $N_t, L_t, D_t \subset X_1 \cup X'_1$ be sets of sizes $n_t, x_t, d_t$ called neutral, alive, dead vertices and $t$ is time. We initialize the process by selecting a vertex $P_0$ setting $L_0 := \{P_0\}, N_0 := v[X_1 \cup X'_1] \setminus \{P_0\}$ and $D_0 := \emptyset$. From $L_{t-1}$ we pass to $L_t$ as follows. First we
select a vertex \( P \in L_{t-1} \). Second to any \( P' \in N_{t-1} \) we check whether there is an \( X_1, X'_t, X_2 \)-or \( X'_t \)-edge having extremities \( \{P, P'\} \). Third we set \( P \) to be dead and add all extremities \( P'' \) found by the process to \( L_t \). The process ends if \( L_t = \emptyset \).

There are three situations possible: \( \{P, P'\} \), \( \{P, Q\} \), \( \{Q, P'\} \) or \( \{P, Q\}, \{P', Q'\} \) are extremities of \( X_1, X'_t \)-edges. Clearly in the first two cases we can add \( P' \) in the next timestep and \( Q, P' \) in the next two time steps to be alive. In the third case there are at least 4 edges in \( K_n \) connecting \( P, P' \) that can be selected as \( X_2, X'_t \)-edges in the following two time steps. Thus to prove \((ii)\) it suffices to show that the above branching process on the random graph \( G_{\frac{\alpha n}{n}} \) on \( \mu \) vertices produces components of size \( K\mu \) where \( K > 0 \) is a constant. On \( G_{\frac{\alpha n}{n}} \) the r.v. \( n_t \) is \( B([\mu], [(1 - \frac{2\alpha}{n})^t]) \) distributed since a vertex \( P \in X_1 \cup X'_t \) stays neutral \( t \) times with probability \( [(1 - \frac{2\alpha}{n})^t] \). Furthermore \( n_t = \mu - l_t - t \) (i.e. \( l_t > 0 \) is equivalent to \( n_t - \mu + t > 0 \)) whence \( \mu \{l_t > 0\} = \mu \{n_t > \mu - t\} \). For times \( t = \lambda \mu \) where \( \lambda > 0 \) we observe

\[
\lim_{n \to \infty} \frac{[(1 - \frac{2\alpha}{n})^t]^\lambda}{e^{-\frac{\lambda t}{n}}} = 1.
\]

The Moivre-Laplace theorem \([4]\) implies that \( n_t \) becomes asymptotically localized at its mean \( \mu e^{-\frac{\lambda t}{n}} \).

We next consider the equation \( e^{-\frac{\lambda t}{n}} = (1 - x) \) which is easily seen to have two solutions \( 1, \xi \) iff \( \frac{\alpha \mu}{n} > 1 \) and distinguish two cases: (I) \( 1 - \lambda < e^{-\frac{\lambda t}{n} \xi} \) and (II) \( 1 - \lambda > e^{-\frac{\lambda t}{n} \xi} \). Clearly case (I) implies already \( n_t > (1 - \lambda)\mu \) whence

\[
\lim_{n \to \infty} \mu \{l_t > 0\} = \lim_{n \to \infty} \mu \{n_t > \mu - t\} = 1.
\]

In case (II), \( 1 - \lambda > e^{-\frac{\lambda t}{n} \xi} \), using the Moivre-Laplace theorem \([3]\), we observe that there exists a constant \( K > 0 \) such that

\[
\lim_{n \to \infty} \mu \{n_t > \mu - t\} < e^{-Kn}.
\]

Since there are \( \mu \) vertices to select in order to initialize the process it follows that the expectation value for components of size \( \geq (1-\lambda)\mu \), where \( 1 - \lambda > e^{-\frac{\lambda t}{n} \xi} \), tends to zero. Accordingly, choosing \( \lambda \) such that \( 1 - \lambda < e^{-\frac{\lambda t}{n} \xi} \) we obtain by the above branching process a.a.s. a component of size \( \leq \mu(1 - \lambda) \). Since \( \lim_{n \to \infty} \mu = nc_1[2 - c_1] \), \((ii)\) follows and the proof of the theorem is complete.
4. Applications

In this section the results of this paper will first be applied to mappings in RNA-secondary structures. Secondary structures are commonly understood as lists of Watson-Crick base pairs. They do not have any further contacts (tertiary interactions). First one finds that in the limit of long sequences there are about 60% of the nucleotides in Watson-Crick base pairs [7, 18, 13]. Moreover secondary structures show an remarkable amount of stability with respect to point-mutations. Typically about 30% of all neighboring sequences are again mapped into the same structure.

In this case the preimages of RNA secondary structures are predicted to form an extended network. It should be composed of a small number of components, depending on the fraction of structure-neutral point-mutants. According to theorem 1, secondary structures are supposed to be stable with respect to random point-mutations and hence should allow for neutral evolution [15]. According to theorem 2 for any choice of \(c_1 < 1\), the two 1-regular subgraphs \(X_1, X'_1\) do not induce components bigger that logarithmic size. Consequently mappings in secondary structures are supposed to guarantee an effective search by the image space by random point-mutations. It should be easy to establish new secondary structures i.e. to switch between two networks.

Schuster and coworkers [6, 11, 18] have studied e.g. the lengths of so called neutral walks i.e. paths in sequence space where each vertex maps into one fixed secondary structure. It turned out that in case of AUGC-sequences one can practically walk into maximal Hamming-distance without changing the structure. Recently in [9, 10] complete mappings of GC-sequences in secondary structures have been reported where first exact numerical data on preimages have been obtained. According to those data the preimages decompose in a few (mostly 1,2,4) number of components, as predicted by the results on random-structures. Finally in [8] so called transitions between secondary structures have been observed. Using a biophysical folding algorithm [12] it was observed that complete populations switch between two preimages.

Moreover the results on random-structures might be helpful for designing 3D-folding algorithms for RNA. In fact theorem 2 shows how crucial the fraction of tertiary interactions is as regards the suitability of the mappings for evolutionary optimization. Known 3D-structures (e.g. t-RNA) have about 4-6% nucleotides involved in tertiary interactions which matches theorem 2.

Finally the results suggest an optimization method for e.g. Boolean networks. Using theorems 1 and 2, mappings from sequences into Boolean nets can be constructed that would allow for effective search by task space without getting trapped into local optima.
Acknowledgments: I want to thank Chris Barret, Simon Fraser, Martin Huynen, Stuart Kauffman, Christopher Langton, Erik Nimwegen, Steen Rasmussen, Peter Schuster, Peter F. Stadler and Andrew Wuensche for stimulating discussions. Furthermore I want to thank the Santa Fe Institute for an excellent collaboration.
Bibliography


