LA-UR-97: 4454

Approved for public release; distribution is unlimited.

Title: EVOLUTION ON FOLDING LANDSCAPES IN COMBINATORIAL STRUCTURES

CONF-980133--

Author(s): S. M FRASER
C. M. REIDYS

Submitted to: THIRD INTERNATIONAL SYMPOSIUM ON ARTIFICAL LIFE AND ROBOTICS
OITA 870-11, JAPAN
JANUARY 19-21, 1998

Los Alamos
NATIONAL LABORATORY

DISCLAIMER

Evolution on Folding Landscapes in Combinatorial Structures

S.M. Fraser
Santa Fe Institute
Hyde Park Road
87504 New Mexico

C.M. Reidys
Los Alamos National Laboratory
TSA/DO-SA
87548 New Mexico, USA
Mailstop: TA-0, SM-1237, MS M997

Abstract

In this paper we investigate the evolution of molecular structures by random point mutations. We will consider two types of molecular structures: (a) (RNA) secondary structures and (b) random structures. In both cases structure consists of (i) a contact graph and (ii) a family of relations imposed on its adjacent vertices. The vertex set of the contact graph is simply the set of all indices of a sequence, and its edges are the bonds. The corresponding relations associated with the edges are viewed as secondary base pairing rules and tertiary interaction rules respectively. Mappings of sequences into secondary and random structures are modeled and analyzed. Here, the set of all sequences that map into a particular structure is modeled as a random graph in the sequence space, the so called neutral network and we study how neutral networks are embedded in sequence space. A basic replication–deletion experiment reveals how effective secondary and random structures can be searched by random point mutations and to what extend the structure effects the dynamics of this optimization process. In particular we can report a non-linear relation between the fraction of tertiary interactions in random structures, and the times taken for a population of sequences to find a high-fitness target random structure.

1 Introduction

Evidently, the term "structure" can reflect different levels of coarse graining. In biophysics "structure" is defined in terms of some physical conditions, for example minimum free energy or kinetic parameters; it can also be defined as the set of all affine coordinates of the atoms in a molecule. Alternatively, a structure can be described as a list of all pairs of coordinates of the sequence that are joined by means of chemical bonds. In this paper we will consider "structure" as such a correlation scheme. In particular we will not assume that this scheme has to fulfill constraints that might arise from an embedding in the three dimensional space. In order to investigate different aspects of the process, we consider the following two types of mappings from RNA sequences: first, mappings into RNA secondary structures, and second, into random structures which are described below. For both RNA secondary structures and random structures, the structure consists of (i) a graph (the *contact graph*), in which the vertices correspond to the indices of the nucleotides and the edges correspond to the bonds, and (ii) a set of relations that represent the base-pairing rules.

2 Structures and compatible sequences

Definition 1 *A secondary structure [7] is a vertex-labeled graph on n vertices $\{1,\dots,n\}$ with an adjacency matrix $A=(a_{i,k})_{1\leq i,k\leq n}$ such that*

- $a_{i,i+1}=1$ *for* $1\leq i\leq n-1$
- *for each i there is at most a single $k\neq i-1,i+1$ such that $a_{i,k}=1$*
- *if $a_{i,j}=a_{k,l}=1$ and $i<k<j$ then $i<l<j$.*

We call an edge $\{i,k\}$, $|i-k|\neq 1$ a *bond* or *base pair*. A vertex i connected only to $i-1$ and $i+1$ is called *unpaired*. The number of base pairs and the number of unpaired bases in a secondary structure s is $n_p(s)$ and $n_u(s)$ respectively. Let $\mathcal{A}=\{a_1,\dots,a_m\}$ be an finite alphabet. A *pairing rule* Π over $\mathcal{A}$ is a symmetric, binary relation over $\mathcal{A}$. Let s be a secondary structure and $\Pi(s)=\{\{i,k\}\,|\,a_{i,k}=1,\,k\neq i-1,i+1\}$ its *set of contacts*. The graph $(\{1,\dots,n\},\Pi(s))$ is called the *contact graph* of the secondary structure s. It neglects the backbone bonds that are listed in the corresponding adjacency matrix. Suppose an alphabet $\mathcal{A}$ and a base pairing rule Π is fixed. A secondary structure induce a partition of sequence space as follows: a vertex $P\in\Omega_\alpha^n$ is *compatible to* s_n 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 $\{i,j\} \in \Pi(s)$. Let $C[s_n]$ be the set of all compatible sequences, it exhibits a graph structure as follows

$$\mathcal{C}[s_n] = \Omega_\alpha^{n_u(s)} \times \Omega_\beta^{n_p(s)}. \tag{2.1}$$

Accordingly, in $\mathcal{C}[s_n]$ two sequences are adjacent if and only if they differ in (i) in a single position i which is unpaired in s, or (ii) in two positions i and j which form a base pair $\{i,j\} \in \Pi(s)$. Secondary structures have no tertiary bonds. Hence generalized structures have been introduced [3]. Random structures allow for a probabilistic analysis of structural properties which is not feasible for secondary structures. Random structures allow for two types of contacts; first the secondary bonds which form a partial 1-factor graph and second tertiary bonds that are completely random and occur with independent probability p. More precisely let $1 \geq c_1 > 0$ and $1 \geq c_2 \geq 0$ be positive constants. Suppose that $m(n)$ is a monotonously increasing map $\mathbb{N} \to \mathbb{N}$ such that $\lim_{n\to\infty} \frac{2m}{n} = c_1$. $m(n)$ can be viewed as the number of secondary bonds of the random structure. Writing a sequence $V \in \Omega_\alpha^n$ as $V = (P_1, \ldots, P_n)$, let X_1 be a partial 1-factor graph on $2m(n)$ indices, say, $\{\ell_{i_1}, \ldots, \ell_{i_{2m}}\} \subset \{1, \ldots, n\}$. X_1 is the contact graph formed by all secondary interactions. Next let X_2 be the random graph obtained by selecting all possible edges between the n nucleotides except the secondary edges with probability c_2/n. Clearly, 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 where a graph (X_2) with k edges has probability $\mu_2\{X_2\} = p^k(1-p)^{\binom{n}{2}-m-k}$ with $p = c_2/n$. The graphs X_1, X_2 induce the random graph $X_1 \otimes X_2$ whose vertex set is $\{1, \ldots, n\}$ and whose edge set $e(X_1 \otimes X_2)$ is the (disjoint) union of all X_1, X_2-edges. $X_1 \otimes X_2$ has probability $\mu\{X_1 \otimes X_2\} = \mu_1\{X_1\}\,\mu_2\{X_2\}$ and is called the random *contact graph*. The probability space formed by the graphs $X_1 \otimes X_2$ will be referred to as Γ^n_{m,c_2}.

Definition 2 *A random structure (r. s.), s_n, on n nucleotides over a finite alphabet $\mathcal{A}$ consists of the following pieces of data: (i) a contact graph $X_1 \otimes X_2$ and (ii) a family of symmetric relations $(\mathcal{R}_*, \mathcal{R}_y)_{y\in X_2}$, where $\mathcal{R}_*, \mathcal{R}_y \subset \mathcal{A} \times \mathcal{A}$, such that for all $a \in \mathcal{A}$ there exists one $b \in \mathcal{A}$ such that $a\mathcal{R}_y b$.*

The relation $\mathcal{R}_*$ is motivated by Watson-Crick *base-pairing rules* observed in RNA secondary structures. For $y \in X_2$ the relation $\mathcal{R}_y$ corresponds to a specific (tertiary) interaction rule that might be context dependent.

Analogous to secondary structures a vertex (sequence) $V \in \Omega_\alpha^n$ is called *compatible* to s_n if and only if (i): for all bonds y of the partial 1-factor graph X_1 its nucleotides indexed by the extremities $\{i,k\}$ have the property $P_i \mathcal{R}_* P_k$ and (ii): its nucleotides fulfill for all tertiary bonds $y \in X_2$: $P_i \mathcal{R}_y P_k$. Again we denote the set of compatible vertices of the r. s. s_n by $C[s_n]$. The contact graph $X_1 \otimes X_2$ induces a partition of its vertex set $\{1, \ldots, n\}$ into the vertex sets of its components $(C_\ell^{(i)})$, where $C_\ell^{(i)}$ denotes component i of $X_1 \otimes X_2$ containing exactly ℓ indices. Let $V = (P_1, \ldots, P_n)$ be a compatible sequence. Each component of the contact graph $X_1 \otimes X_2$, $C_\ell^{(i)}$, induces a multi-set $(P_{i_1}, \ldots, P_{i_\ell})$, consisting of nucleotides whose indices belong to $C_\ell^{(i)}$. This multi-set can be viewed to be an element of a new alphabet, $\mathcal{A}_i$, whose elements are all possible multi-sets $(P_{i_1}, \ldots, P_{i_\ell})$ induced by compatible sequences. Accordingly we can rewrite a compatible sequence as (c.f. (2.1))

$$(A_{i_1}, \ldots, A_{i_l}), \tag{2.2}$$

l being the number of components of the contact graph. We next analyze the graph structure of the contact graph of random structures [3].

Theorem 1 *Let $0 \leq c_2, c_1 \leq 1$, $\frac{2m(n)}{n} \nearrow c_1$ and let $\tilde{T}$ be the r.v. representing the number of vertices of a random graph Γ^n_{m,c_2} that are contained in tree components. Then for $[c_1 + c_2] < 1$ asymptotically almost all vertices of Γ^n_{m,c_2} are in tree components, i.e.*

$$\lim_{n\to\infty} \left[\mathbb{E}[\tilde{T}]/n\right] = 1. \tag{2.3}$$

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

Accordingly, contact graphs of r.s. decompose with probability 1 into small components and it might in this context be of interest to state the typical ranges for c_1, c_2 that are observed in RNA and protein structures: $0.4 \leq c_1 \leq 0.7$, $0 \leq c_2 \leq 0.2$. In this parameter range Theorem 1 guarantees that a.s. almost all nucleotides of the contact graph $(X_1 \otimes X_2)$ are contained in small components.

3 Neutral Networks as Random Graphs

In this section we introduce a probabilistic model of sequence to structure maps in secondary and random structures as introduced in [6]. The model is formulated in the language of random graph theory and can be sketched as follows: a secondary structure or random structure s_n determines a set of compatible sequences, as described in (2.1) and (2.2). For each compatible sequence we probabilisticly decide whether or not it maps into s_n by selecting it with independent probability λ. The resulting random subset of the set of compatible sequences induces a Ω_α^n-subgraph to which we refer to as $\Gamma_n[s_n]$.

The above modeling of preimages of structures as random graphs bases to large extent on properties of random induced subgraphs of generalized n-cubes. In the following we state some basic properties of random induced subgraphs of generalized n-cubes [6].

Theorem 2 *Let Ω_α^n be a generalized n-cube and Γ_n an induced subgraph with $\mu_n\{\Gamma_n\} = \lambda^{|\Gamma_n|}(1-\lambda)^{\alpha^n-|\Gamma_n|}$ and let $\lambda^* = 1 - \sqrt[\alpha-1]{\alpha^{-1}}$. Then*

$$\lim_{n\to\infty} \mu_n\{\Gamma_n \text{ is } \Omega_\alpha^n\text{-dense and connected}\} = \begin{cases} 1 & \text{for} \quad \lambda > \lambda^* \\ 0 & \text{for} \quad \lambda < \lambda^* . \end{cases}$$

The next result shows that for $\lambda_n \geq \frac{c\ln(n)}{n}$ there exists a.s. a unique giant component in random induced subgraphs of Ω_α^n [5].

Theorem 3 *Let Ω_α^n be a generalized n-cube, $\lambda_n = \frac{c\ln(n)}{n}$, $c > 0$ and μ_n a measure such that $\mu_n\{\Gamma_n\} = \lambda_n^{|\Gamma_n|}(1-\lambda_n)^{\alpha^n-|\Gamma_n|}$. Then there exist $c > 0$, $h \in \mathbb{N}$ such that the largest Γ_n-component, $C_n^{(1)}$, is the induced subgraph of all Γ_n-vertices that are contained in Γ_n-components of size $\geq n^h$, and*

$$\forall \epsilon > 0 \qquad \lim_{n\to\infty} \mu_n\{\Gamma_n \,|\, |C_n^{(1)}| \geq [1-\epsilon]|\Gamma_n|\} = 1 .$$

Accordingly, for remarkably small picking probabilities, asymptotically with probability 1, for a pair of Γ_n-vertices there exists a Γ_n-path connecting them. The next result investigates paths in random induced subgraphs of generalized n-cubes further. It shows, that for the slightly larger picking probability $\lambda_n \geq n^a$, $0 < a < 1/2$ the random graph distance of two vertices P, Q scales linearly with the corresponding distance of P, Q in Ω_α^n [4].

Theorem 4 *Let Ω_α^n be a generalized n-cube, λ_n a probability such that*

$$\exists n_0 \in \mathbb{N}, \forall n \geq n_0 : \qquad \lambda_n \geq n^{-a} \quad \text{where } 0 < a < 1/2 , \tag{3.1}$$

and $\mu_n\{\Gamma_n\} = \lambda_n^{|\Gamma_n|}(1-\lambda_n)^{\alpha^n-|\Gamma_n|}$ a measure on the set of all induced Ω_α^n-subgraphs. Then, for $k \in \mathbb{N}$ such that $k > \frac{1+3a}{1-2a}$

$$\lim_{n\to\infty} \mu_n\{\Gamma_n \,|\, \text{for } P,Q \in C_n^{(1)} : \text{a.s. } d_{\Gamma_n}(P,Q) \leq [2k+3]\, d_{\Omega_\alpha^n}(P,Q)\} = 1 .$$

4 Sequence to Structure Maps

The intuitive idea of defining probabilistic sequence to structure mappings in combinatorial structures $f : \Omega_\alpha^n \longrightarrow \{s_n\}$ is quite simple: one labels a given set of structures, $\{s_n\}$ and constructs iteratively the neutral networks as random graphs over the remaining sequences. That is, 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

$$f_r^{-1}(s_0) = \Gamma_n[s_0] \quad f_r^{-1}(s_i) = \Gamma_n[s_i] \setminus \bigcup_{j<i} [\Gamma_n[s_i] \cap \Gamma_n[s_j]] . \tag{4.1}$$

One central property of the probabilistic sequence to structure mappings is the embedding of two neutral networks, i.e. what is the Hamming distance between two neutral networks in sequence space? For secondary structures the following is the key result regarding this question [6]:

Theorem 5 *Let Π be a nonempty pairing rule on $\mathcal{A}$ and s_n and s_n' be arbitrary (nonempty) secondary structures. Then*

$$C[s_n] \cap \; C[s_n'] \neq \emptyset . \tag{4.2}$$

Theorem 5, combined with Theorem 2, guarantees that for secondary structures with large neutral networks $\Gamma_n[s_n]$, $\Gamma_n[s_n']$ there are some sequences $P \in \Gamma_n[s_n]$, $Q \in \Gamma_n[s_n']$ with pairwise small Hamming distances $d_{\Omega_\alpha^n}(P,Q)$. With random structures, we can also study the minimum distance between neutral networks analytically in terms of a new graph, which is obtained by taking the superposition of all edges of the contact graphs of two random structures. Let $X_1 \otimes X_2$ and $X_1' \otimes X_2'$ be the contact graphs with respect to two random structures. Then their union

graph, $(X_1 \cup X_1') \otimes (X_2 \cup X_2')$, is obtained by first taking the union of the secondary edges in X_1, X_1', and then all the tertiary X_2, X_2' edges that are not already in X_1, X_1'. The main result reads [3]

Theorem 6 *Let $X_1 \otimes X_2$ and $X_1' \otimes X_2'$ be two random contact graphs with $\lim_{n\to\infty} \frac{2m}{n-1} = c_1 > 0$ and $0 \leq c_2 \leq 1$. Then for $c_1 < 1$ and $c_2 = 0$ a.s. almost all $(X_1 \cup X_1') \otimes (X_2 \cup X_2')$-vertices are contained in line-graph components. Further, there exists a constant $C > 0$ such that a.s. components of $(X_1 \cup X_1') \otimes (X_2 \cup X_2')$ have size $\leq C \ln(n)$.*
Suppose that $8c_1[2-c_1]c_2 > 1$ and that $\xi \neq 0$ solves $(1-x) = e^{-8c_1[2-c_1]c_2 x}$. Then $(X_1 \cup X_1') \otimes (X_2 \cup X_2')$ has a.s. components $C^{(n)}$ with the property

$$|C^{(n)}| \geq (1-\xi)n[\frac{4m}{n} - (\frac{2m}{n})^2]. \qquad (4.3)$$

According to Theorem 6 there is a distinct change in the graph-structure of $(X_1 \cup X_1') \otimes (X_2 \cup X_2')$. Below the critical value for c_2 (for $c_1 \approx 0.6$ $c_2^{crit} \approx 0.13$) the largest component is $\leq C \ln(n)$, $C > 0$, and above the critical value a giant component emerges.

5 Discussion

In some sense random structures are generalizations of biomolecular secondary structures, which are formally planar knot-free graphs together with rules associated with their bonds. Both, random structures and secondary structures ([2]), induce neutral networks in sequence space, i.e. extended, mostly connected subgraphs consisting of all sequences that are all mapped into the random structure. However, random structures differ from secondary structures in two important regards. First, they may include tertiary interactions, and secondly, they need not satisfy such a knot-freeness condition. Random structures induce a natural, tractable probability space, Γ^n_{m,c_2}, and accordingly allow for the formulation of "almost surely "results, like Theorem 6. They also enable investigation of the influence of tertiary contacts. For the biologically-realistic parameter range, contact graphs of random structures exhibit a similar graph structure to the contact graphs of RNA secondary structures (Theorem 1); their largest component scales with the logarithm of sequence length. The computer generated statistics of contact graphs of random structures illustrate the assertions of Theorem 1. These findings imply that RNA secondary structures and random structures exhibit a significant robustness with respect to point-mutations. As noted in the introduction, the stability of these molecular structures with respect to mutations in the sequences that form them is important in the context of [1], since it provides an explanation for the neutrality of many point mutations at the genomic level. It also explains how diffusion on neutral networks by the accumulation of neutral mutations can, at some later stage, allow novel functionality to emerge (i.e. the population discovers a new neutral network). In the above we have used RNA molecules as the model because of the availability of folding algorithms and formal structural descriptions, because of the clear relationship between structure and function, and the history of *in vitro* RNA experiments. But of course many of the same arguments apply to proteins. In the mapping between DNA or RNA sequences and proteins there is an additional source of neutrality, which is the redundancy in the codon–amino acid coding. The robustness of RNA secondary structures and random structures is the key observation for the modeling of their neutral networks. Neutral networks consist of a few components, whose size depends on the fraction of neutral point mutants (with respect to the structure). Key properties of neutral networks are connectivity and path-structure. For random induced subgraphs of generalized n-cubes there exists a threshold value for connectivity 2 and a giant component exists already for $\lambda_n \geq \frac{\ln(n)}{n}$ Theorem 3. In the case of RNA secondary structures connectivity can be formulated with respect to their corresponding graph of compatible sequences $\mathcal{C}[s] \cong \Omega_\alpha^{n_u} \times \Omega_\beta^{n_p}$ (2.1). Accordingly, connectivity is defined with respect to (i) point-mutations and (ii) base pair-mutations. Since the graph $\Omega_\beta^{n_p}$ is simply a generalized n-cube over the alphabet of base pairs all results of random induced subgraphs derived so far (Theorems 2,3 and 4) apply accordingly. An extension of the connectivity properties is the computation of the length of actual paths between two sequences of a neutral network 4. The main result is that for $\lambda_n \geq n^{-a}$, $0 < a < 1/2$ the random graph distance scales linearly with the distance in the generalize n-cube. Because random structures form a probability space, it becomes possible to analyze in detail the graph structure of the union of two contact graphs, which can only be done for RNA secondary structures with a group theoretic argument [3]. This union of two contact graphs contains information about how close the neutral networks of the two constituent structures come. Considering the edges of the contract graphs as constraints that have to be fulfilled for a sequence to be compatible with a particular structure, the union graph encodes those constraints

for two structures simultaneously; a sequence must fulfill all those constraints to be bicompatible. It is in the region of bicompatible sequences that transitions between neutral networks take place, and the more base pairs in a sequence that are incompatible with the union graph, the greater the number of mutations a sequence must undergo to move between the two networks. We have shown (Theorem 6) that the structure of this union graph changes dramatically while varying the fraction of tertiary interactions c_2 in the two constituent graphs, and shows a phase transition phenomenon at certain critical threshold c_2 with the sudden emergence of a giant component in the union graph. This giant component is very likely to contain cycles, which make it unlikely that that bicompatible sequences will exist, and will increase the distance between the neutral networks. Thus, we expect that populations will find it increasingly difficult to make transitions between neutral nets at higher c_2 values. Larger structures, in which the phase transition is more marked, would show even stronger effects of c_2. For a $c_1 = 0.6$ the critical fraction of tertiary interactions for the emergence of a giant component in the union graph of two random structures has $c_2 = 0.15$ as an upper bound (eq. (2)). For $c_1 = 1$ a lower bound on the critical fraction would be 0.125. Known 3D-structures (for example t-RNA) have values of c_2 which are well below this critical threshold, with about 4-6% nucleotides involved in tertiary interactions. In order to investigate the effects of structure on the ability of populations to search, we induce dynamics over a population of sequences using a replication–deletion process which is described above. In these simulations, sequences are assigned a fitness according to the structure they form, reflecting the assessment of fitness at the phenotypic level. The replication–deletion process is a simply scheme in which sequences are replicated according to their fitness, with a constant population size, in a flow reactor-type process. It should be noted that the only genetic operator that is being used here is point mutation—sequences make only local moves in the sequence space, via their mutant offspring. Yet because of the guarantee that for biologically realistic parameter values neutral networks are dense and connected (Theorem 2), populations of sequences evolving in this way are able to search a large proportion of the shape space. The results demonstrate this—populations of 2000 sequences rapidly find the neutral network of one of the few high-fitness structures in a landscape of 10000 structures.

Acknowledgements

We want to thank Christopher L. Barrett for stimulating discussions. Special thanks to Darrell Morgeson for his continuous support. SF is funded by DARPA under grant ONR N0014-95-1-1000.

References

[1] M. Kimura. *The Neutral Theory of Evolution.* Cambridge Univ. Press, 1983.

[2] C.M. Reidys. *Neutral Networks of RNA Secondary Structures.* PhD thesis, Friedrich Schiller Universität, Jena, Math. Faculty, September 1995.

[3] C.M. Reidys. Mapping in random-structures. *SIAM Journal of Discrete Mathematics and Optimization*, 1996. submitted.

[4] C.M. Reidys. Random Graphs and Sequence to Structure Maps. *Combinatorics Probability and Computing.*, 1997. submitted.

[5] C.M. Reidys. Random Induced Subgraphs of Generalized n-Cubes. *Adv. Appl. Math.*, 19(AM970553):360–377, 1997.

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

[7] M.S. Waterman. Combinatorics of RNA hairpins and cloverleaves. *Studies Appl. Math.*, 60:91–96, 1978.

Report Number (14) LA-UR--97-4454
CONF-980133--

Publ. Date (11) 199711
Sponsor Code (18) DOD, XF
UC Category (19) UC-000, DOE/ER

DOE