LA-UR- 97-122

CONF-970937--2

Title: MUTATION AND MOLECULES TOWARDS MUTATION-BASED COMPUTATION

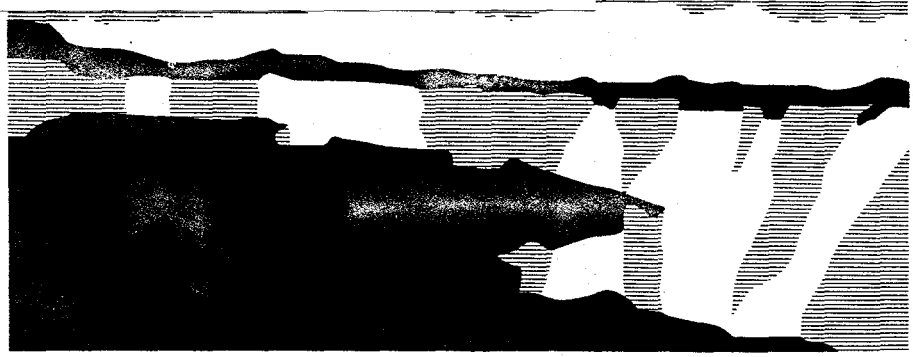
Author(s): CHRISTIAN M. REIDYS
CHRISTOPHER L. BARRETT

Submitted to: Optimization and Simulation Conference
Singapore
September 1-4, 1997

## 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.

MASTER



Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Form No. 836 R5
ST 2629 10/91

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

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

# Mutation and Molecules
# Towards Mutation based Computation

By

Christian M. Reidys and Christopher L. Barrett

Los Alamos National Laboratory
TSA/DO-SA, 87548 New Mexico

*Mailing Address:
Los Alamos National Laboratory
Mailstop: TA-0, SM-1237, MS M997
TSA/DO-SA, 87548 New Mexico, USA
Phone: **(505) 665-0911 Fax: **(505) 665-7464
E-Mail: ducktsasa.lanl.gov

**Abstract:** In this paper recent results on random graphs are used as a framework for a theory of mutation based computation. The paradigm for mutation based computation will be the evolution of molecular structures. The mathematical structure of "folding maps" into molecular structures is shown to guarantee an effective search by point-mutations. Detailed mathematical models for these mappings are discussed. We will show that combinatorial structures consisting of (i) a (random) contact graph and (ii) a family of relations imposed on its adjacent vertices allow for efficient search by point mutations. We will determine the graph structure of the contact-graph and dicuss its relation to the optimization process. Mappings of sequences into random structures are constructed. Here, the set of all sequences that map into a particular random structure is modeled as a random graph in sequence space, the neutral network. We will analyze the graph structure of neutral networks and show how they are embedded in sequence space. Explicitely we discuss connectivity and density of neutral networks and prove that any two neutral networks come close in sequence space. Finally several experiments are shown that illustrate the prospective of using this molecular computation method.

**Key words.** random structure, sequence-structure mapping, random graph, connectivity, giant component, optimization, evolution

# 1 Introduction

Molecular structures have evolved in time by randomizations on sequence level. The randomizations were essentially of "local" nature, like, for example point mutations caused by radiation. It is highly nontrivial that this kind of "local" variation is sufficient to find a certain target structure. One could think of a sort of "local-optimum" in which a population gets stuck because all fitter structures cannot be reached by point mutations. The main point of this paper consists in showing that molecular folding landscapes exhibit a generic "mathematical" structure that allows practically to find *any* target structure by point mutations. Therefore point mutations are a powerful search method in all landscapes that exhibt an analogous mathematical structure. Mutations and molecules are in this sense strongly correlated objects. Our analysis focusses on two main points. First we discuss what is meant by "structure"; in fact we will construct a probability space of structures and determine their main properties. Second we determine the basic properties of preimages of the above random structures. Again we will work in a probability space of possible preimages. In particular we determine how the preimages, the so called neutral networks, are embedded in sequence space. In the course of this analysis we will introduce a mathematical model for generating an analyzing mappings from sequences into structures.

The term "structure" can reflect different levels of coarse graining. For us "structure" will consist of a list of pairs of coordinates of the sequence that are paired by means of chemical

bonds. More precisely it will be a *contact graph* and a *multi-set of relations* imposed on the extremities of its edges. The contact graph is a random graph whence we refer to the above structures as *random structures* [2]. One important feature of molecular structures is their robustness with respect to point mutations. Consequently many 1 or 2 mutant neighbors of a sequence will fold into the same structure and sequences realising this structure form a neutral network in sequence space. This robustness is a well studied phenomenon in molecular biology and has been discussed in the context of *neutral evolution* [?]. It is closely related to the structure of the contact graph of the molecule. Suppose a mutation occurs in a component of the contact graph. Then, according to the relations (rules) associated with the edges (bonds), a high fraction of all nucleotides of this component has to be changed in order to stay compatible. The probability of the sequence remaining compatible with the molecule decreases exponentially with the size of the component in which the mutation occurred.

The paper is structured as follows: first we introduce the concept of random structures and compatibility. Second, we construct preimages (neutral networks) of random structures as random subgraphs of $\mathcal{Q}_\alpha^n$ and third we analyze how neutral networks are embedded in sequence space. Finally some computer experiments are presented that illustrate how the set of structures is searched.

## 2 Random structures and compatibility

A *graph* $X$ consists of a tuple $(vX, eX)$ and a map $o \otimes t : eX \longrightarrow vX \times vX$. $vX$ is called the *vertex set* and $eX$ the *edge set*. An element $P \in X$ is called a *vertex* of $X$; an element $y \in X$ is called an *edge*. The vertex $o(y)$ is called the *origin* of $y$ and the vertex $t(y)$ is called the *terminus* of $y$; $o(y), t(y)$ are called the *extremities* of some edge $y$. There is an obvious notion of $Y$ being a *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 *path* in $X$ is a sequence $(Q_1, y_1, Q_2, y_2, \ldots, y_n, Q_{n+1})$, where $Q_i \in X$, $y_i \in X$, $o(y_i) = Q_i$ and $t(y_i) = Q_{i+1}$. A path such that $Q_1 = Q_{n+1}$ is called a *cycle*. $X$ is called *connected* if any two vertices are vertices of a path of $X$. A connected graph without cycles is called a *tree*. Being connected is an equivalence relation in $X$, and the maximal connected subsets of vertices are called *components* of $X$. For $Y < X$, the *closure* of $Y$ in $X$, $\overline{Y}$, is the induced subgraph of all vertices of $X$ that are adjacent to some vertices of $Y$. A subgraph $Y \to X$ 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 $y \in X$.
A sequence $V \in \mathcal{Q}_\alpha^n$ is a tuple $(P_1, \ldots, P_n)$ where $\mathcal{Q}_\alpha^n$ is a generalized $n$-cube. The $\mathcal{Q}_\alpha^n$-vertices are sequences of length $n$ over the alphabet $\mathcal{A}$ of size $\alpha$ and two sequences are adjacent in $\mathcal{Q}_\alpha^n$ if they differ in exactly one coordinate. Let $1 \geq c_1, c_2 > 0$ be positive constants and suppose $m(n) \in \mathbb{N}$ fulfills $2m(n)/n \nearrow c_1$ ($m$ corresponds to the number of secondary bonds of the molecule). Let now $X_1$ be a partial 1-factor graph on $2m$

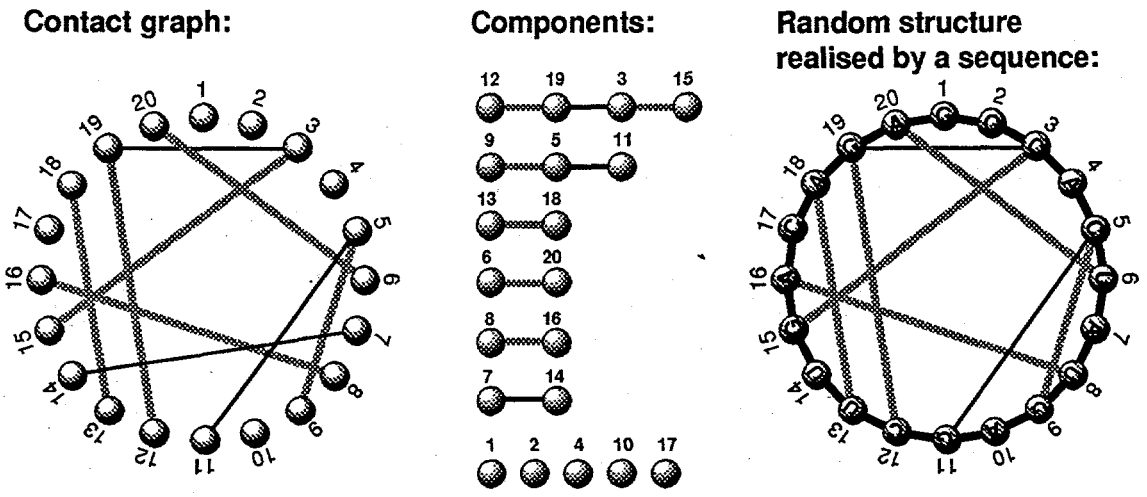



Figure 1: *A contact graph, consisting of an ordered set of vertices (numbered), between which there can be either secondary (gray) or tertiary (thin black) edges, together with its set of components. On the right hand side, the bases of one compatible sequence are shown on the vertices, indicating that, in the random structure, certain relations associated with the edges have to be fulfilled (in this case Watson–Crick base-pairing rules).*

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 with $k$ edges has probability $p^k(1-p)^{\binom{n}{2}-m-k}$ with $p = c_2/n$ [1]. The graphs $X_1, X_2$ induce the 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. $X_1 \otimes X_2$ is called the *contact graph*. The probability space formed by the graphs $X_1 \otimes X_2$ will be referred to as $\Gamma^n_{m,c_2}$. A *random structure*, $s_n$, on $n$ nucleotides of a finite alphabet $\mathcal{A}$ consists of the following pieces of data:
$(i)$ the contact graph $X_1 \otimes X_2$ and
$(ii)$ a family of symmetric relations $(\mathcal{R}_*, \mathcal{R}_y)_{y \in X_2}$, where $\mathcal{R}_*, \mathcal{R}_y \subset \mathcal{A} \times \mathcal{A}$.
Each $\mathcal{R}_y$ is supposed to have the property: for all $a \in \mathcal{A}$ there exists one $b \in \mathcal{A}$ with the property: $a\mathcal{R}_y b$. 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.
A vertex (sequence) $V \in \mathcal{Q}^n_\alpha$ is called *compatible* to $s_n$ if and only if

- for all bonds $y$ of the partial 1-factor graph $X_1$ its nucleotides indexed by the extremities $\{o(y), t(y)\}$ have the property $P_{o(y)}\mathcal{R}_* P_{t(y)}$ (note that since $\mathcal{R}_*$ is symmetric we also have $P_{t(y)}\mathcal{R}_* P_{o(y)}$)

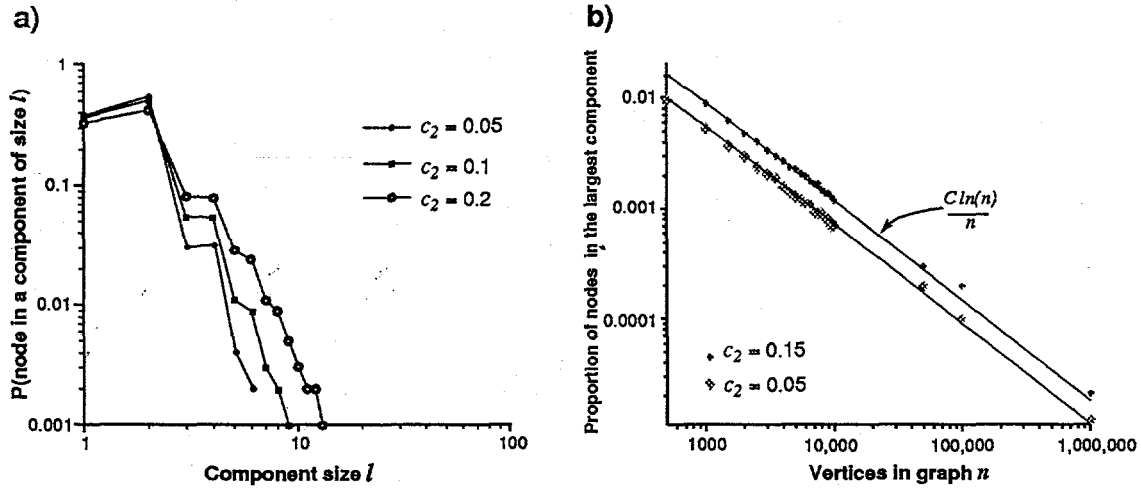



Figure 2: *a) The distribution of vertices in components of size $\ell$ in a single contact graph, for different values of $c_2$ (the fraction of tertiary bonds). b) The scaling of the largest component in a single contact graph with sequence length (number of vertices n) for $c_2 =$ 0.05 and 0.15. Also shown are fitted lines with $y = C\ln(n)/n$, with $C = 0.8$ and 1.3 respectively.*

- its nucleotides fulfill for all tertiary bonds $y \in X_2$: $P_{o(y)}\mathcal{R}_y P_{t(y)}$.

The set of compatible vertices with respect to the random structure $s_n$ is called $C(s_n)$. By construction there are $n - 2m$ vertices not incident to an $X_1$-edge and there are asymptotically $[n-2m]e^{-c_2}$ isolated vertices in $X_1 \otimes X_2$.

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

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

*There exists a constant $C(c_1, c_2) > 0$ such that a.a.s. all paths in $\Gamma^n_{m,c_2}$ 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 $\Gamma^n_{m,c_2}$, $T$, have the property $|T| \leq C\ln(n)$.*

Accordingly contact graphs decompose with probability 1 into small tree components. In this context some bounds for $c_1, c_2$ that are observed in RNA and protein structures might be of interest: $0.4 \leq c_1 \leq 0.7$ and $0 \leq c_2 \leq 0.2$. Plugging this in we obtain that most nucleotides of the contact graph $(X_1 \otimes X_2)$ are contained in very small components.

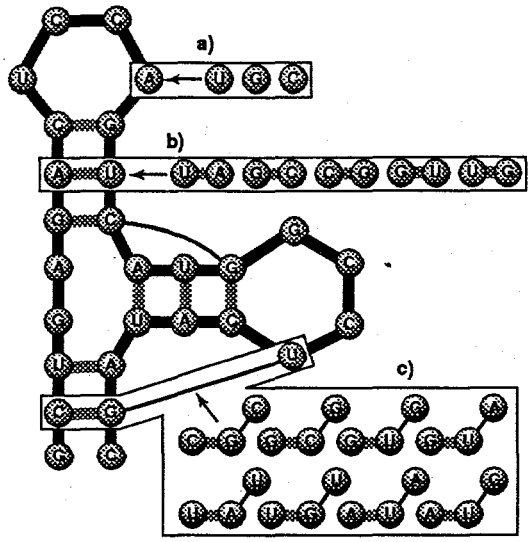



Figure 3: *This figure illustrates the decomposition of a sequence with respect to the components of the contact graph of the structure. Nucleotides in three types of components (single nucleotides* **a**, *nucleodides in secondary interactions* **b**, *and nucleotides in secondary and tertiary interactions* **c**) *are shown. For each of the three classes we monitor all possible compatible segments.*

# 3 Neutral Networks

In this section we establish a mapping between sequences and random structures and construct preimages of structures as random graphs in sequence space. Let $s_n$ be a random structure. Its preimage is necessarily contained in $C(s_n)$, the set of compatible sequences. We have shown that the underlying contact graph has almost all vertices in tree components of at most logarithmic size. It induces moreover a partition of the indices $\{1,\dots,n\}$ into its components. We can regroup the indices of the nucleotides of a compatible sequence into the components of the contact graph. Formally we can now consider each multi-set $(P_{i_1},\dots,P_{i_k})$, consisting of nucleotides whose indices belong to a component of the contact graph, to be an element of a new alphabet, $\mathcal{A}_k$. Accordingly we can rewrite a compatible sequence as $(A_{i_1},\dots,A_{i_\ell})$ ($\ell$ being the number of components of the contact graph).

To illustrate this let us consider an RNA secondary structure with respect to the biophysical alphabet **A,U,G,C**. The latter has a contact graph whose edges are exactly the paired positions $\{(i_1,i_k),\dots,(i_m,i_j)\}$. These are also all nontrivial components of the contact graph. The Watson-Crick base pairing rule, **AU,UA,GC,CG,GU,UG**, corresponds to the induced alphabet. Accordingly, a compatible sequence $(P_1,\dots,P_n)$ can be rewritten as $(P_{i_1},\dots,P_{i_k},(P_{j_1},P_{j_2}),\dots,(P_{j_r},P_{j_{r+1}}))$ where each pair $(P_{j_k},P_{j_{k+1}})$ fulfills the Watson-Crick base pairing rules. Here the set of compatible sequences is the vertex

set of $\mathcal{Q}_\alpha^{n-\ell} \times \mathcal{Q}_\beta^\ell$, where $\ell$ is the number of base pairs.
In general the set of compatible sequences is the vertex set of

$$\prod_{i=1}^{h} \mathcal{Q}_{\alpha_i}^{n_i} \quad \text{where} \quad \sum_i i \cdot n_i = n\,.$$

$\alpha_i = |\mathcal{A}_i|$, $h$ is the number of components, and $n_i$ the length of the $i$-th component of the contact graph. Next we construct the preimage of the random structure $s_n$. It will be a random induced graph by selecting the vertices in each factor $\mathcal{Q}_{\alpha_i}^{n_i}$ with independent probability $\lambda_i$. Note that "vertex" here corresponds to a multi-set $(P_{i_1}, \ldots, P_{i_k})$ consisting of nucleotides whose indices belong to a component of the contact graph of $s_n$. In this sense "vertex" can be viewed as a certain segment of the sequence. $\lambda_i$ ($i$ being the index of a component) can be interpreted as the stability of the random structure with respect to a mutation that has (i) occurred in the $i$-th component and that has (ii) led to a compatible sequence. In this context the structure of randomly induced subgraphs of generalized $n$-cubes is of particular interest:

**Theorem 2.** *[4] Let $\mathcal{Q}_\alpha^n$ be a generalized $n$-cube and $\Gamma_n$ an induced subgraph with $\mu_n\{\Gamma_n\} = \lambda^{|\Gamma_n|}(1-\lambda)^{\alpha^n - |\Gamma_n|}$. Then*

$$\lim_{n\to\infty} \mu_n\{\Gamma_n \text{ is } \mathcal{Q}_\alpha^n\text{-dense}\} = \begin{cases} 1 & \text{for} \quad \lambda > 1 - \sqrt[\alpha-1]{\alpha^{-1}} \\ 0 & \text{for} \quad \lambda < 1 - \sqrt[\alpha-1]{\alpha^{-1}}. \end{cases}$$

Moreover, the number of isolated vertices in random graphs $\Gamma_n$ is Poisson with mean $\mu = \alpha^n(1-\lambda)^{\alpha(n-1)}$. $\lambda^*$ is furthermore the threshold value for connectivity.

**Theorem 3.** *[4] Let $\mathcal{Q}_\alpha^n$ be a generalized $n$-cube and $\Gamma_n < \mathcal{Q}_\alpha^n$ a random induced subgraph with $\mu_n\{\Gamma_n\} = \lambda^{|\Gamma_n|}(1-\lambda)^{\alpha^n - |\Gamma_n|}$. Then*

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

The next theorem shows that random induced subgraphs exhibit giant components for surprisingly small probabilities, $\lambda_n = \frac{c\ln(n)}{n}$.

**Theorem 4.** *[3] Let $\mathcal{Q}_\alpha^n$ be a generalized $n$-cube, $\lambda_n = \frac{c\ln(n)}{n}$ and $\mu_n$ a measure on $\mathcal{G}(\mathcal{Q}_\alpha^n)$ such that $\mu_n\{\Gamma_n\} = \lambda_n^{|\Gamma_n|}(1-\lambda_n)^{\alpha^n - |\Gamma_n|}$. Then we can choose constants $c, C > 0$ such that the largest $\Gamma_n$-component, $C_n^{(1)}$, is the induced subgraph of all $\Gamma_n$-vertices that are contained in $\Gamma_n$-components of size $\geq n^3$. Further $C_n^{(1)}$ has the property*

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

*The second largest $\Gamma_n$-component, $C_n^{(2)}$, is of size $\leq Cn/\ln(n)$.*

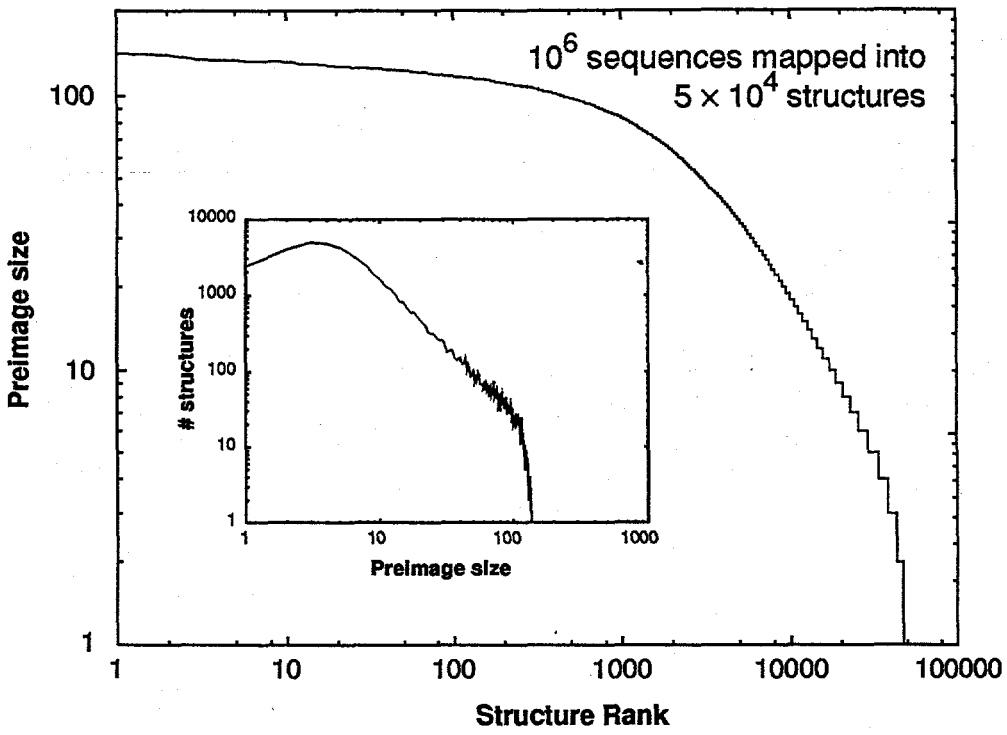



Figure 4: *The distribution of the sizes of neutral networks for a mapping in random structures with* $n = 30$, $\lambda = 0.8$ *and fraction of tertiary bonds,* $c_2 = 0.05$. *The curve bases on the data of* $10^6$ *sequences being mapped into* $5 \times 10^4$ *random structures. The inserted curve dispays the frequency distribution of preimage sizes.*

The above results imply that above a certain threshold value neutral networks of structures are denss and connected within the set of compatible sequences. As long as the underlying contact graphs have almost all components of length 1, 2 and 3. Sequences can easily move by single digit error probabilities $p$ such that $pn \approx 1$.

We obtain mappings $f : Q_\alpha^n \longrightarrow \{s_n\}$ by constructing the corresponding preimages as random graphs iteratively, i.e. we first choose a mapping $r : \{s_n\} \to N$ where $j \leq i \implies r(s_j) \geq r(s_i)$ and set

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

Figure 4 monitors the distribution of the sizes of neutral networks of random structures. Analogous to RNA data [?] and also to data on small proteins [?] these curves obey a generalized Zipf law $f(x) = a(1 + x/b)^{-c}$ where $x$ is the rank of the structure and $a, b, c$ are positive constants. $b$ can be interpreted as the number of frequent structures and $c$ describes the power-law decay for rare structures. Note that structures with low ranks exhibit neutral networks that have giant components and percolate sequences space.

# 4 Searching by point mutations

One central question for the search process on molecules is how well the set of structures is searched by point-mutations. In other words to what extend are point mutations a sufficient variation mechanism?

A *population* $\mathbf{V}$, of size $N$, is a (finite) multi-set of sequences $(V_i \,|\, i \in N)$ where $\{ V_i \,|\, i \in N \} \subset Q_\alpha^n$ and $N > 1$. The time evolution of $\mathbf{V}$ is obtained by a mapping from $(V_i \,|\, i \in N)$ to the family $(V_i' \,|\, i \in N)$ as follows: we select an ordered pair $(V_l, V_k)$ where $V_l, V_k \in \{ V_i \,|\, i \in N \}$. The first coordinate $V_l$ of the pair is chosen with a probability that is the fitness of $V_l$ relative to the average firness in the population among the elements of $\mathbf{V}$. The second coordinate is selected with uniform probability on $(V_i \neq V_l \,|\, i \in N)$, i.e. $1/(N-1)$. We select those pairs of sequences at equidistant time steps, and for a population of size $N$ we refer to a *generation* as $N$ such time steps. Next we map the first sequence, $V_l = (x_1, ..., x_n)$, into the sequence $V^* = (x_1', ..., x_n')$. This is performed by assigning to each coordinate $x_i$a $x_i' \neq x_i$ with probability $p$ where all $x_i' \neq x_i$ are equally distributed and leave the coordinate fixed otherwise. This random mapping $_l \mapsto v^*$ is called *replication.* Finally, we delete the second coordinate of the pair $(V_l, V_k)$, that is $V_k$ and have a mapping $(V_l, V_k) \mapsto (V_l, V^*)$. Thereby we obtain a new family by substituting the $V_k$ by the $V^*$.

We will next explain why the graph structure of the union of two contact graphs is of particular importance for the search in the set of structrues and then analyse its basic graph structure.
Suppose $s_n$ is a random structure that has a high fraction of a *population* of sequences on its neutral network and that $s_n'$ is the target structure. Then, how likely will sequences that fold into $s_n$ be mutated into sequences folding into the target structure? If we take the union of all bonds of the contact graphs of $s_n$ and $s_n'$ we obtain a new graph. This graph will allow to describe how "close" the above two structures come and plays thereby a central role for the answer to our question. Formally we could view the union graph as a new type of contact graph. We could then use this to determine sequences that fulfill the constraints imposed by *both* the constituent random structures. Now, if on the one hand the union graph decomposes in small components, the above arguments discussed in relation to single contact graphs apply; it is highly likely that "bi-compatible" sequences exist, which are capable of folding into both structures. On the other hand, if the union graph exhibits a large component of order $n$ it is unlikely that bi-compatible sequences will exist. The key result for random structures reads:

**Theorem** 5.*[2] Suppose* $\Gamma^n_{m,c_2}$ *and* $\Gamma'^n_{m,c_2}$ *are two random contact-graphs with* $\lim_{n\to\infty} \frac{2m}{n-1} = c_1 > 0$ *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* $\Gamma^n_{m,c_2} \cup \Gamma'^n_{m,c_2}$ *are contained in components that are line-graphs. There exists a constant* $C > 0$ *with the property that a.a.s. all components in* $\Gamma^n_{m,c_2} \cup \Gamma'^n_{m,c_2}$ *have lengths* $\leq C \ln(n)$.

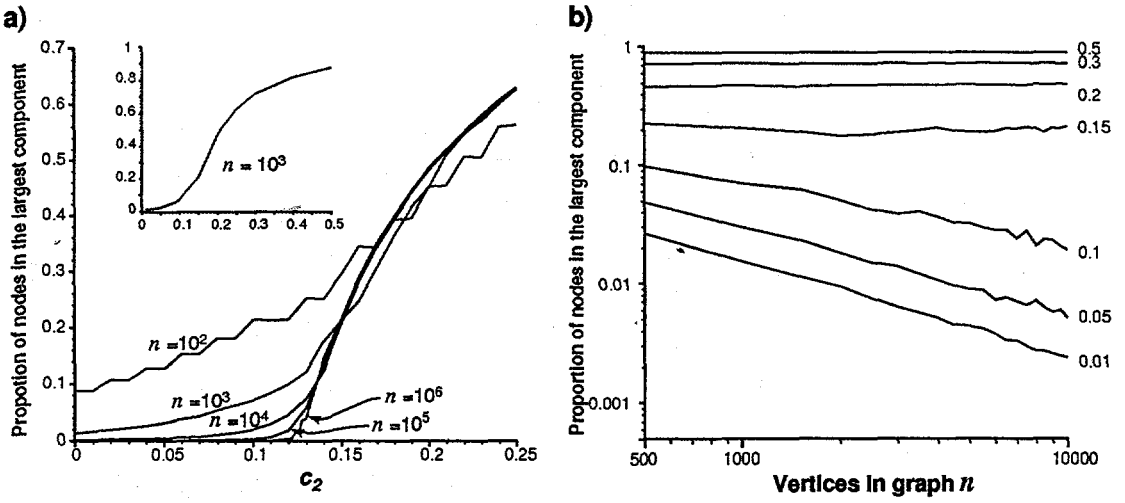



Figure 5: *Illustration of theorem 4: a) The size of the largest component in the union of two contact structures for $c_1 = 0.6$ as a function of $c_2$. Curves are shown for graphs with $10^2$–$10^6$ nodes with $0 \geq c_2 \geq 0.25$ and, inset, for a graph of 1000 nodes with $c_2 \leq 0.5$. b) The size of the largest component as a function of sequence length n. Curves for various values of $c_2$ are displayed.*

*(ii) Suppose $8c_1[2-c_1]c_2 > 1$ and $\xi \neq 0$ solves $(1-x) = e^{-8c_1[2-c_1]c_2 x}$. Then $\Gamma^n_{m,c_2} \cup \Gamma'^n_{m,c_2}$ has a.a.s. components $C^n$ with the property*

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

According to this theorem there is a phase transition in $(X_1 \cup X_1') \otimes (X_2 \cup X_2')$. Below the critical value for $c_2$ (which is 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 of emerges. Figure 4 illustrates this dramatic change of the graph structure of the union-graph [?]. Below the critical value for $c_2$ a population can switch between any two structures doing point mutations.

We finally introduce a class of generic fitness landscapes on random structures in which we let a population of $N$ sequences search for a certain target structure. This target structure will always be the fittest one. For this purpose let $\varphi : \mathcal{Q}^n_\alpha \to \{s_n\}$ be a fixed sequence to structure map. Set $f_{\eta,\xi} : \{s_n\} \to \mathbb{R}$, $\eta, \xi \in \mathbb{R}_+$:

$$(i) \qquad \mu\{f_{\eta,\xi} = k\xi\} = \frac{\eta^k}{k!} e^{-\eta}.$$

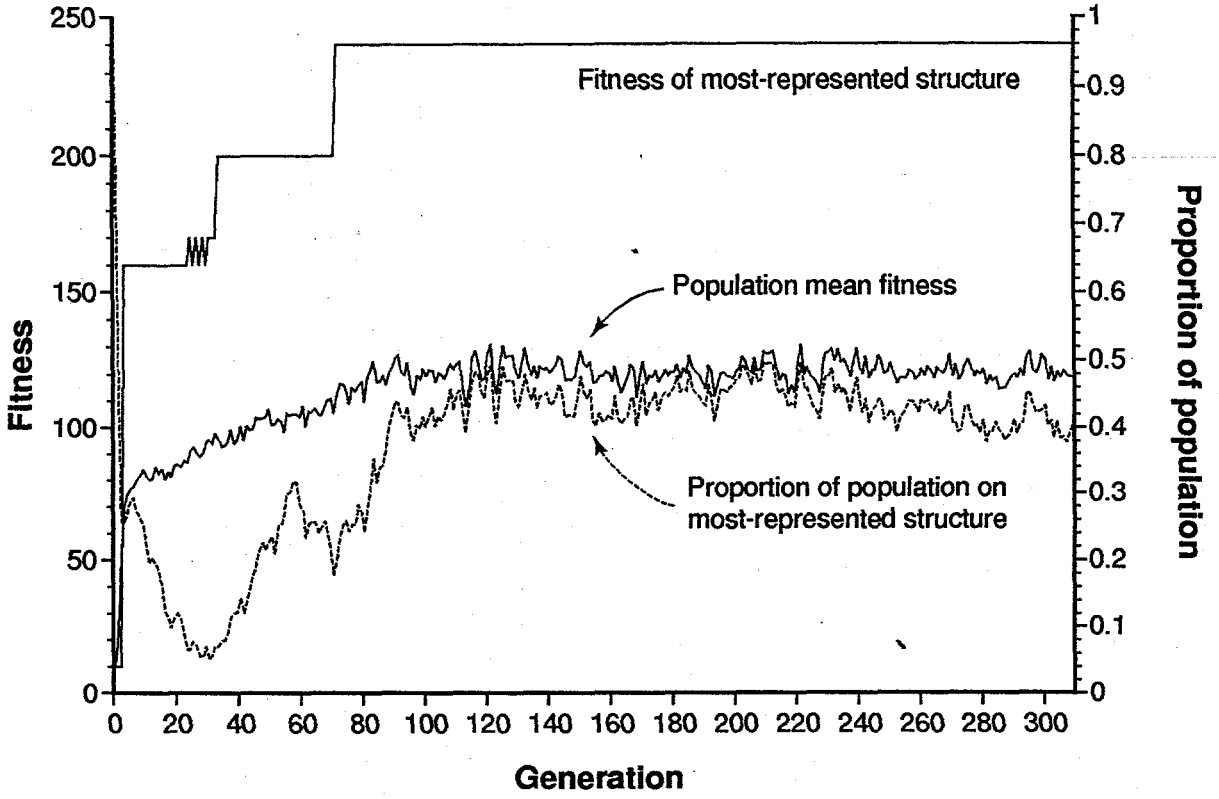



Figure 6: *The time evolution of a population of* 2000 *sequences of length* 40 *in a Poisson landscape with* $\eta, \xi = 10$. *The following three curves are shown: the mean fitness, the fraction of the population realizing the most represented structure (mrs) in the population and the fitness of the mrs as functions of time. The highest possible fitness in this experiment was* 250 *and the run was terminated after* 300 *generations, where the population had one of* 3 *possible structures with fitness* 240 *as mrs. In the experiment the probability of hitting the target structure with a random sequence read* $10^{-4}$.

## 5 Conclusion

Point mutations induce local variations of sequences. Under the basic assumption that fitness is defined on structural level, molecular structures turn out to be well suited for the action of point mutations. In fact this a generic property of molecular folding landscapes which turn out to have a distinct mathematical structure. First, on a certain structural level, contact graphs of structures have been proven to be robust with respect to point mutations. Structures with large neutral networks, so called frequent structures, have highly connected, extended, percolating neutral networks in sequence space. The number of frequent structures, however, scales exponentially with sequence length. These results can be proven rigorously using random graph theory and rely mostly on combinatorial rather

than biophysical properties of base pairs. The robustness guarantees that fit phenotypes can be preserved while the underlying genotypes, the sequences continuously perform neutral mutations.
Second, within the parameter range observed for molecular structures, populations can realise new phenotypes simply by doing point mutations i.e. they can perform transitions between the associated neutral networks. Thus "innovation" is obtained by point mutations–all frequent structures are inevitably found.
Molecular folding landscapes serve as a paradigm of landscapes in which point mutations are fully sufficient for search. It might be speculated that molecules and mutations have coevolved in time. The natural question to ask here is to what extend also other fitness landscapes can be searched by mutations and how to determine those landscapes. In this class of landscapes neutral networks are generic features which make sure that the optimum will be found as long as the latter has a neutral network that is dense and connected. New inside is needed for the analysis of fitness landscapes and in general to obtain results on the relation between landscape and variation method.

**Acknowledgments** We want to thank Simon Fraser for discussions and his support in particular for creating the figures ..... .

## References

[1] Béla Bollobás. *Random Graphs.* ACADEMIC PRESS, 1985.

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

[3] C.M. Reidys. Random induced subgraphs of generalized $n$-cubes. *Advances in Applied Mathematics*, 1997. accepted, Dec. 1996.

[4] C.M. Reidys, F.P. Stadler, and P.K. Schuster. Generic properties of combinatory maps and neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 1995. in press.