Generic Properties of Combinatory Maps

Neutral Networks of RNA Secondary Structures

By

Christian Reidys[a,*], Peter F. Stadler[b,c], and Peter Schuster[a,b,c]

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

[b]Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria

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

[*]Mailing Address:
Institut für Molekulare Biotechnologie
Beutenbergstraße 11, PF 100 813, D-07708 Jena, Germany
Phone: **49 (3641) 65 6459 Fax: **49 (3641) 65 6450
E-Mail: duck@imb-jena.de

Abstract

Random graph theory is used to model relationships between sequences and secondary structures of RNA molecules. Sequences folding into identical structures form neutral networks which percolate sequence space if the fraction of neutral nearest neighbors exceeds a threshold value. The networks of any two different structures almost touch each other, and sequences folding into almost all "common" structures can be found in a small ball of an arbitrary location in sequence space. The results from random graph theory are compared with data obtained by folding large samples of RNA sequences. Differences are explained in terms of RNA molecular structures.

1. Introduction

Conventional biophysics is concerned with structure predictions of biopolymers that relates a structure to a given sequence. Structure is defined in the context of some physically defined conditions like, for example, minimum free energy structures fulfilling the common 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 shall not be one-to-one: many sequences may be mapped onto the same structure. The degree of this redundancy shall 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 shall 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 notion of structure apparently is context dependent and thus anything but trivial.

Coarse-grained protein structures are usually expressed in terms of secondary structure elements, for example α-helices, β-sheets and reverse turns, and their arrangements in three-dimensional space. This is illustrated best by the popular "ribbon-models". RNA secondary structures are representative for another type of coarse-graining: they are commonly understood as lists of Watson-Crick (**AU** and **GC**) and **GU** base pairs. Base pairing and base pair stacking constitute the major contributions to the free energy of RNA structure formation and consequently the base pairs of secondary structures are conserved in the three-dimensional structures of the RNA molecules (Figure 1). In addition, biochemists and molecular biologists have successfully used RNA secondary structures for molecular interpretations of RNA function.

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 [22] Straightforward estimates on the numbers of possible secondary structures for given chain length n show high degree of redundancy: there are many more sequences than structures and hence many sequences have to fold into one and the same structure.

The relation between RNA sequences and secondary structures is understood as a (non invertible) mapping from sequence space into shape space [15] The sequence space is a metric space of all sequences where the metric is given by the Hamming distance [19] (counting the number of positions in which two aligned sequences differ). In the case of point mutations the Hamming metric is a measure of relatedness in the sense that close by sequences are closely related since they are inter-converted by

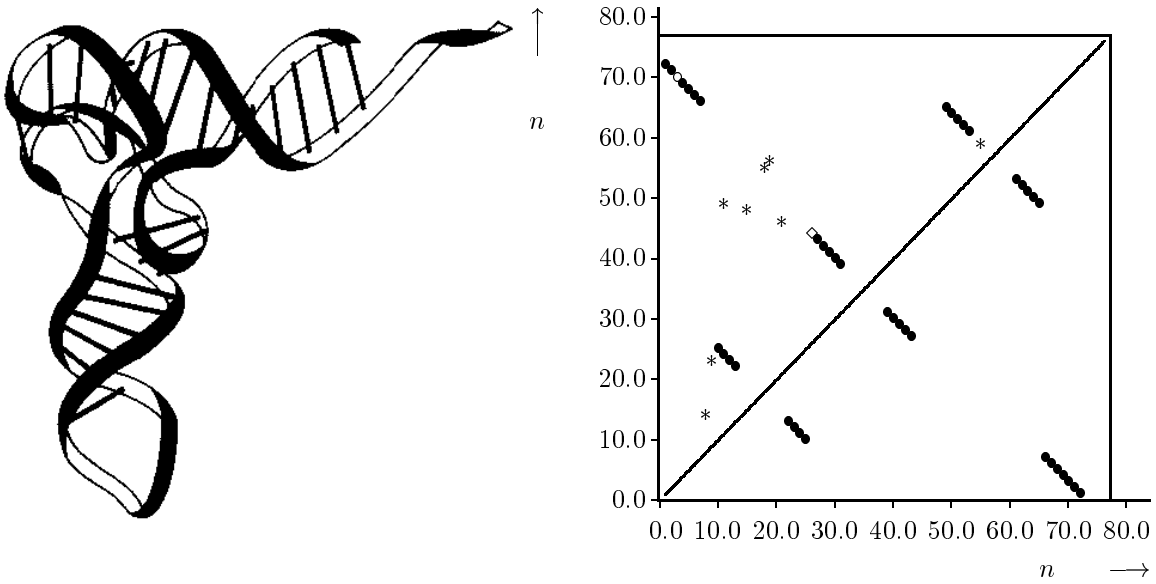


Figure 1: **a)** Three dimensional structure of phenylalanin tRNA from yeast.

b) Contact map of tRNA-phe.

The upper triangle shows the contact of the three-dimensional structure: • Watson-Crick base pairs, ○ GU base pairs, ◇ other non-Watson-Crick base pairs in double helical regions, and ∗ tertiary contacts between bases.

The lower triangle shows the secondary structure consisting of both Watson-Crick and GU base pairs, •.

a small number of mutations. The other metric space used here is an abstract space of all structures with some metric defined according to a concept of relatedness between structures. The concept applied here is based on the idea of converting two structures into each other by means of a set of weighted operations that are illustrated best in a tree representation of secondary structures [15]. For the purposes pursued here the particular representation of secondary structures is not important: the general conclusions drawn here, in essence, depend much more on other factors such as the degree of redundancy, the base pairing alphabet (two letters, e.g. **GC**, or four letters), or the chain length n of the polynucleotide.

The concept of sequence space to shape space mapping allows to study global properties of sequence-structure-relations that are inaccessible otherwise. (An illustrative example is the idea of *shape space covering* described in [35] An understanding of sequence-structure relations, however, is of central importance in biophysics of biopolymer structure. Its knowledge is fundamental to conventional biotechnology, since is represents the basis of rational design of biopolymers and it is required for conceiving efficient optimization experiments in applied molecular evolution.

RNA secondary structures distinguish only paired and unpaired regions irrespective of the particular bases at the individual positions (**G**, **C**, **A**, or **U**). We can expect therefore that many different sequences can meet the base pairing conditions as determined by a given secondary structure. Indeed if two bases capable of forming a base pair (**GC**, **GU**, **CG**, **AU**, **UA**, or **UG**) oppose each other at all positions in base paired regions we are dealing with a *compatible sequence* of the structure under consideration. The number of compatible sequences is readily computed for any given secondary structure S_k, with n_u unpaired bases and n_p base pairs is $4^{n_u} \cdot 6^{n_p}$. Clearly, the chain length of

the biopolymer molecule is simply given by $n = n_u + 2n_p$. 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 [41]

In contrast to counting compatible sequences folding into a given structure by fairly straightforward combinatorics, estimates on the numbers of sequences that form the target structure under given conditions is a particularly hard problem. Two strategies were followed, one consisting in folding all sequences of a given chain length into secondary structures and evaluating the data by enumeration [17], the other using some abstract model for sequence-to-structure mapping and performing rigorous mathematical analysis. Here we pursue the latter approach and make the assumption that apart from biophysical pairing rules the mapping of sequences into structures 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 [3].

This paper is organized as follows: In section 2 we introduce the notion of a combinatory map and argue that a sensible model of biologically relevant sequence-structure maps must account explicitly for the degree of neutrality. In section 3 we define and analyze a random graph model for combinatory maps, which is refined to match the details of RNA secondary structures in section 4. These refined models are compared with computational data from RNA folding in section 5. Section 6 contains our conclusions.

2. Configuration Spaces and Landscapes

2.1. Configuration Spaces

In molecular evolution (and, apart from recombination, in all biology) the basis of variation is simply the limited accuracy of replication. Replication errors or mutations produce RNA sequences which differ from the parental template sequence. In other words, mutations act on the *sequence* of the DNA (or RNA in case of some viruses).

At the level of individual nucleotides we can distinguish *point mutations*, *insertions*, and *deletions*, see figure 2. While insertions and deletions alter the size of the genome, the chain length is kept constant under the action of point mutations.

Figure 2: Three classes of mutations. Point mutations are copying errors with single base exchanges; they leave the chain lengths constant. In case of insertions part of the template sequence is duplicated during replication. A deletion leads to an error copy which is shorter than the original.

A very different type of mutations occurs at the level of genomes. Entire genes can be inserted or deleted, and the genome can be rearranged. Again, insertions and deletions change the size of the genome – now in terms of the number of genes, while rearrangements conserve the number of genes. They lead to a permutation of the genome, see [33].

Mutations can be viewed as "moves" in an abstract set of *configurations*. This suggests a natural "geometrical" arrangement of the configurations (be they DNA sequences or arrangements of genes on a mitochondrial genome): Configurations that can be inter-converted by a single move (mutation) may be viewed as neighbors. Consequently, the smallest number of moves which is necessary to inter-convert to arbitrary configurations u and v can be interpreted as a distance $d(u, v)$. (Of course we shall assume that the neighborhood relation is symmetric: if u can be obtained by a single mutation

from v, then it is also possible to produce v as a direct mutant of u.) The neighborhood relation allows us to view the set of all configurations as an undirected graph: Each configuration is represented by a vertex, and neighboring configurations are connected by an edge. It is trivial to check that the distance measure $d(\ ,\ )$ is a metric – in fact, it coincides with the canonical metric on the graph [5]. A sensible definition of the mutation operators shall lead to a connected graph: every configuration can be reached from any other configuration by a sequence of mutations.

2.2. Combinatory Maps and Landscapes

Let $\mathcal{S}_n$ be the (finite) set of all structures (where the definition of the structures themselves is irrelevant for our discussion) that can be formed from sequences of length n. Then the *shape space* is the discrete metric space $(\mathcal{S}_n, D)$ where D is an arbitrary metric[1]. An example is the set of all *secondary structures*[2] of RNA for given sequence lengths n together with a *tree metric* [37, 38, 12, 22]. We can always choose the trivial metric $D_0 : \mathcal{S}_n \times \mathcal{S}_n \to \mathbb{R}_0$ defined by

$$D_0(u,v) = \begin{cases} 0 & \text{iff} \quad u = v, \\ 1 & \text{otherwise.} \end{cases}$$

Definition 1. *Let $\mathcal{C}_n$ be a configuration space, and $\mathcal{S}_n$ a set of structures. Then a combinatory map is a map $f_n : \mathcal{C}_n \to \mathcal{S}_n$.*

Two vertices x, y of $\mathcal{C}_n$ are said to be equivalent if there is an *automorphism* $\alpha \in \mathbf{Aut}(\mathcal{C}_n)$ such that $y = \alpha(x)$. A graph automorphism is a bijection on the edge set of the graph. A graph is *vertex transitive* if all vertices are equivalent. Applying the automorphism group to *pairs* of vertices we obtain the *symmetry classes* Θ as its orbits:

$$\Theta(x_0, y_0) = \{(x,y) \in \mathcal{C}_n \times \mathcal{C}_n \,\big|\, \exists \alpha \in \mathbf{Aut}[\mathcal{C}_n] : \, (\alpha(x), \alpha(y)) = (x_0, y_0)\}\,.$$

A graph is *distance transitive* if for any two pairs of vertices (x,y) and (u,v) with $d(x,y) = d(u,v)$ there is an automorphism α such that $\alpha(x) = u$ and $\alpha(y) = v$. In this case the symmetry classes are exactly the distance classes of $\mathcal{C}_n$, i.e. $\{(x,y) | d(x,y) = \Delta\}$ is a symmetry class.

The *correlation function* of a combinatory map is given by:

$$\rho(\Theta) = 1 - \frac{\mathbf{E}[D^2(f_n(x), f_n(y)) \,|\, (x,y) \in \Theta]}{\mathbf{E}[D^2(f_n(x), f_n(y)) \,|\, x, y \in \mathcal{C}_n]}\,,$$

where Θ is a symmetry class of $\mathcal{C}_n$. Recall that for distance transitive graphs such as generalized hypercubes, the symmetry classes coincide with the (Hamming) distance classes.

[1] A metric on a set X is a map $D : X \times X \to \mathbb{R}$ with the following properties: For all $x, y, z \in \mathcal{S}_n$ holds $D(x,y) = 0 \iff x = y$, $D(x,y) = D(y,x)$ and $D(x,z) \leq D(x,y) + D(y,z)$.

[2] We give a formal definition of a secondary structure in section 4.1.

Definition 2. *Two configurations $x, y \in \mathcal{C}_n$ are called neutral iff $f_n(x) = f_n(y)$. We call them neutral neighbors if $d(x, y) = 1$. By*

$$\lambda_s(\Theta) \stackrel{\text{def}}{=\!=} \frac{|\{(x,y) \in \Theta |\, f_n(x) = f_n(y) = s\}|}{|\Theta|}$$

we denote the expected fraction of neutral configurations in symmetry class Θ, and the average frequency of neutral neighbors is given by

$$\overline{\lambda}(\Theta) \stackrel{\text{def}}{=\!=} \frac{1}{|\mathcal{C}_n|} \sum_{s \in \mathcal{S}_n} |f_n^{-1}(s)| \lambda_s(\Theta).$$

The *density surfaces* $\boldsymbol{\mu}(D|\Theta)$ are defined as the conditional probability that two structures $f_n(x)$ and $f_n(y)$ have distance D in the shape space provided that the underlying sequences are in symmetry class Θ. These quantities have been used extensively for the investigation of RNA folding. The autocorrelation function $\rho(\Theta)$ can be calculated from the density surface [15]. Furthermore we have by definition $\overline{\lambda}(\Theta) = \boldsymbol{\mu}(0|\Theta)$, independent of the metric D on $\mathcal{S}_n$.

Now consider the trivial metric. We have then $D^2 = D$, and therefore

$$\begin{aligned}
\mathbf{E}[D^2(f_n(x), f_n(y)) \,|\, x, y \in \mathcal{C}_n] &= 1 - \frac{1}{|\mathcal{C}_n|^2} \sum_{s \in \mathcal{S}_n} |f_n^{-1}(s)|^2 \\
\mathbf{E}[D^2(f_n(x), f_n(y)) \,|\, (x,y) \in \Theta] &= \frac{1}{|\mathcal{C}_n|} \sum_{s \in \mathcal{S}_n} |f_n^{-1}(s)| (1 - \lambda_s(\Theta)) = 1 - \overline{\lambda}(\Theta)\,.
\end{aligned}$$

Some simple calculations finally yield

$$\rho(\Theta) = \frac{\overline{\lambda}(\Theta) - \sum_{s \in \mathcal{S}_n} \left(\frac{|f_n^{-1}(s)|}{|\mathcal{C}_n|}\right)^2}{1 - \sum_{s \in \mathcal{S}_n} \left(\frac{|f_n^{-1}(s)|}{|\mathcal{C}_n|}\right)^2}. \tag{1}$$

Hence the correlation function coincides with the neutral neighbor frequency in the case if the trivial metric is used (up to a very small correction).

It is convenient to introduce the notation

$$\wp(s) \stackrel{\text{def}}{=\!=} \frac{|f_n^{-1}(s)|}{|\mathcal{C}_n|}$$

for the *frequency* of s. The *rank* $\imath(s)$ is obtained by sorting the structures with respect to their frequencies. The *rank order function* $\psi : \mathbb{N} \to \mathbb{R}$ of the combinatory map f is then given by $\psi(\imath(s)) \stackrel{\text{def}}{=\!=} \wp(s)$. The rank order is well known for the case of folding RNA into its secondary structure, see section 2.3.

The correction terms in equ. (1) can be recast in terms of the rank order function ψ, yielding

$$\rho(\Theta) = \frac{\overline{\lambda}(\Theta) - \sum_i \psi^2(i)}{1 - \sum_i \psi^2(i)} \approx \overline{\lambda}(\Theta)\,.$$

In general $\sum_i \psi^2(i)$ is small and we are left with $\rho(\Theta) \approx \bar{\lambda}(\Theta)$.

2.3. Random Maps

Let X, Y be sets. Then we can obtain a map $f : X \longrightarrow Y$ by selecting each $\eta \in Y$ with the same probability to be the image for a given $\xi \in X$. We thereby consider *random maps* on finite sets, $f : X \to Y$ and use the abbreviations $x = |\, X \,|$ and $y = |\, Y \,|$. We denote further the set of all maps $f : X \longrightarrow Y$ by

$$\mathbf{Map}(X, Y) \stackrel{\text{def}}{=\!=} \{f \,|\, f : X \longrightarrow Y\}\,.$$

Using the *measure* $\boldsymbol{\mu}\{\, f \,\} = 1/y^x$ we obtain the probability space $(\mathbf{Map}(X, Y), \boldsymbol{\mu})$.

An interesting quantity is the distribution of preimage sizes. To this end we introduce the random variable

$$\hat{Z}_k(f) \stackrel{\text{def}}{=\!=} |\, \{z \in Y \,|\, |\, f^{-1}(z) \,| = k\} \,|,$$

which counts the number of images of given size $|\, f^{-1}(s) \,| = k$. We next set

$$\mathbf{E}[\hat{Z}_k] \stackrel{\text{def}}{=\!=} \frac{1}{y^x} \sum_{k=0}^{y} |\, \{f \,|\, \hat{Z}_k(f) = k\} \,|\,.$$

Claim *For $k \in \mathbb{N}$, $\mathbf{E}[\hat{Z}_k]$ is given by $\mathbf{E}[\hat{Z}_k] = B(k, x, 1/y) \cdot y$.*
To prove the claim we first observe

$$|\, \{f \,|\, |\, f(X) \,| = m \,\wedge\, \hat{Z}_k = x_k \,\}\, | = \binom{y}{m} |\, \mathrm{dis}(x_k, m, x) \,|$$

where $|\, \mathrm{dis}(x_k, m, x) \,|$ is the number of different distributions of x elements in m different cells with x_k cells containing k elements and no cell empty. Then we express $\mathbf{E}[\hat{Z}_k]$ by

$$\mathbf{E}[\hat{Z}_k] = \frac{1}{y^x} \sum_{m=1}^{y} \binom{y}{m} \sum_{x_k=1}^{m} x_k \,|\, \mathrm{dis}(x_k, m, x) \,|\,.$$

The values $|\, \mathrm{dis}(x_k, m, x) \,|$ fulfill the functional equation [31]

$$\sum_{x=0}^{\infty} |\, \mathrm{dis}(x_k, m, x) \,| \frac{z^x}{x!} = \binom{m}{x_k} (\frac{z^k}{k!})^{x_k} ([e^z - 1] - \frac{z^k}{k!})^{m - x_k} \tag{2}$$

The above equation implies the following recursion formula for $|\, \mathrm{dis}(x_k, m, x) \,|$:

$$\frac{m}{x_k} \binom{x}{k} |\, \mathrm{dis}(x_k - 1, m - 1, x - k) \,| = |\, \mathrm{dis}(x_k, m, x) \,|\,.$$

We further obtain

$$\sum_{x=0}^{\infty} \left[\sum_{x_k=1}^{m} x_k \,|\, \mathrm{dis}(x_k, m, x) \,| \right] \frac{z^x}{x!} = m \frac{z^k}{k!} [e^z - 1]^{m-1} \tag{3}$$

and proceed by computing

$$\begin{aligned} \sum_{x=0}^{\infty} \sum_{m=1}^{y} \binom{y}{m} \sum_{x_k=1}^{m} x_k \,|\, \mathrm{dis}(x_k, m, x) \,| \frac{z^x}{x!} &= \left[\frac{z^k}{k!}\right] \sum_{m=1}^{y} \binom{y}{m} m \,[e^z - 1]^{m-1} \\ &= \left[\frac{z^k}{k!}\right] y\, e^{(y-1)\,z}\,. \end{aligned}$$

Setting

$$a_{k,y,x} \stackrel{\text{def}}{=\!=} \sum_{m=1}^{y} \binom{y}{m} \sum_{x_k=1}^{m} x_k \,|\, \mathrm{dis}(x_k, m, x) \,|$$

we can write $\mathbf{E}[\hat{Z}_k] = \frac{1}{y^x} a_{k,y,x}$. By comparison of coefficients of the above identity we derive

$$\mathbf{E}[\hat{Z}_k] = \frac{1}{y^x} \binom{x}{k} (y-1)^{x-k}\, y = B(k, x, 1/y)\, y\,,$$

proving the claim.
Further we obtain using the recursion formula for $|\,\mathrm{dis}(x_k, m, x)\,|$

$$\sum_{x=0}^{\infty} \left[\sum_{m=2}^{y} \binom{y}{m} \sum_{x_k=1}^{m} (x_k^2 - x_k) |\,\mathrm{dis}(x_k, m, x)\,| \right] \frac{z^x}{x!} = \left[\frac{z^k}{k!}\right]^2 y\,(y-1)\, e^{(y-2)\,z}\,,$$

and by comparison of coefficients we end up with

$$\sum_{m=2}^{y} \binom{y}{m} \sum_{x_k=1}^{m} (x_k^2 - x_k)\,|\,\mathrm{dis}(x_k, m, x)\,| = \frac{y\,(y-1)x!}{(k!^2)(x-2k)!}\,(y-2)^{x-2k}\,.$$

This implies for $\mathbf{V}[\hat{Z}_k] \stackrel{\text{def}}{=\!=} \mathbf{E}[\hat{Z}_k^2] - \mathbf{E}[\hat{Z}_k]^2$

$$\mathbf{V}[\hat{Z}_k] = \binom{x}{k} \binom{x-k}{k} (\frac{1}{y})^{2k}\, (1 - \frac{2}{y})^{x-2k} y\,(y-1) + B(k, x, 1/y)\, y - [B(k, x, 1/y)\, y]^2\,.$$

The value $\mathbf{E}[\hat{Z}_k]$ can be interpreted as the *frequency* of the preimage size k in the set $\mathbf{Map}(X, Y)$. It is immediately checked that $\mathbf{E}[\hat{Z}_k]$ is a unimodal curve in k that has at $k = \lceil \frac{x}{y} \rceil$ its point of maximum. Moreover if $(X_n, Y_n)_{n\in\mathbb{N}}$ is a family of finite sets such that $x_n = |\,X_n\,| \nearrow \infty$, $y_n = |\,Y_n\,| \nearrow \infty$ and $\lim_{n\to\infty} \frac{x_n}{y_n} = 0$, $\mathbf{E}[\hat{Z}_k^n]$ becomes a delta function localized at $\lceil \frac{x_n}{y_n} \rceil$.

This fact contradicts computational results about RNA sequence-to-structure maps. RNA sequence to secondary structure mappings 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 [35, 41]. The above parameters a, b and c have the following interpretation: a is a normalization constant, b is the number of frequent structures and c describes the power-law decay for rare structures. Figure 3 shows the dependence of b on the chain length n. For large n we find $a = (c-1)/b$ using the continuous approximation $\int_0^\infty \psi(\imath) d\imath = 1$. Consequently $\int_0^\infty \psi^2(\imath) d\imath = [(c-1)^2/(2c-1)]b$,.

Extending the above result on $\mathbf{E}[\hat{Z}_k]$ by introducing an *a priori* probability $P(s)$ for mapping a ξ to $s \in Y$ we obtain an inhomogeneous spectrum of preimage sizes. Even this ansatz is not sufficient to describe the features observed for RNA-folding maps.
Let us suppose that the set of definition X is a generalized hypercube $\mathcal{Q}_\alpha^n$ and Y the set of RNA secondary structures. The average number of neutral neighbors with respect to the structure s for a mapping that assigns each $\xi \in X$ with the probability $P(s)$ to s is $(\alpha - 1)\, n\, P(s)$. Now computational data on RNA sequence to structure maps exhibit that for all "frequent" structures their corresponding fraction of neutral neighbors is asymptotical (that is for $n \to \infty$) constant. Further those data

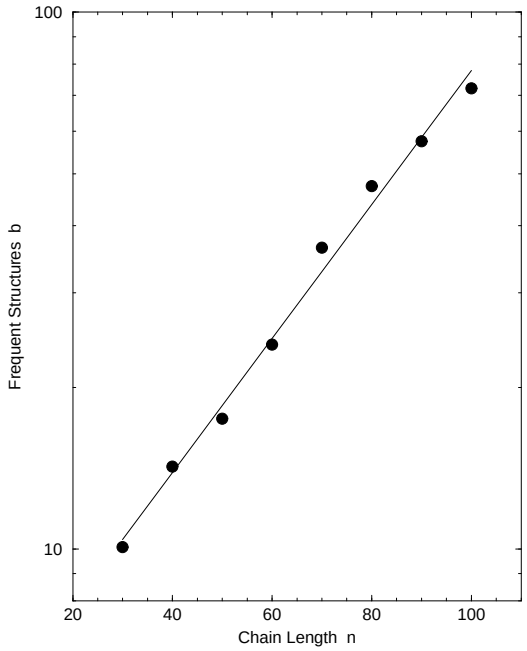


Figure 3: Dependence of the number of "frequent" structures b on the chain length n for RNA secondary structures (biophysical alphabet). These data have been obtained for so-called loop-structures, a coarse grained version of the full secondary structure graphs. For details see [12, 35, 22].

show (see figure 3) that there are exponentially many "frequent" structures. The latter observation contradicts the implications of the above model. If there are exponentially many constant probabilities $P(s_i)$ the expected number of "neutral neighbors" tends to zero in the limit of infinite chain length since each sequence has only $(\alpha - 1)\, n$ adjacent sequences.

We therefore conclude that random maps — even with a non-uniform *a priori* probability for different structures — cannot explain one prominent feature of RNA sequence-to-structure relations: The existence of so called "neutral networks" [35] i.e. extended networks in sequence space that consist of sequences folding into a fixed secondary structure. In the sequel we shall proceed by introducing general models for random induced subgraphs of configuration spaces. These shall finally be used to construct the "neutral networks" formally as random graphs.

3. Random Induced Subgraphs

3.1. Basic Random Graph Models

3.1.1. Basics of Graph and Probability Theory

Before we introduce the basic models we recall some basic facts of graph theory.

Notation. A *graph* G is a pair $(\mathrm{v}[G], \mathrm{e}[G])$, together with two *incidence maps* $\tilde{\tau} : \mathrm{e}[G] \to \mathrm{v}[G]$ and $\tilde{\iota} : \mathrm{e}[G] \to \mathrm{v}[G]$. $\mathrm{v}[G]$ is called the *vertex set* and $\mathrm{e}[G]$ the *edge set*. We can interpret $\tilde{\iota}(e)$ and $\tilde{\tau}(e)$ as the two vertices defining a (directed) edge. For our purposes it shall be more convenient to consider an edge e as the unordered set of vertices $e = \{x, y\}$, $x, y \in \mathrm{v}[G]$. We say x is *incident to* e if $x = \tilde{\iota}(e)$ or $x = \tilde{\tau}(e)$. Further two vertices $x, y \in \mathrm{v}[G]$ are called *adjacent* if and only if $\{x, y\} \in \mathrm{e}[G]$.

- 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 *degree* δ_v of a vertex v is the number of edges $e \in \mathrm{e}[G]$ of the form $e = \{v, v'\}$.
- 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 $(v = v_1, e_1, v_2, ...e_{m-1}, v_m = v')$ where $e_k = \{v_k, v_{k+1}\}$ 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 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

$$\mathrm{Supp}(\pi) \stackrel{\mathrm{def}}{=\!=} \{v \in \mathrm{v}[G] |\, v \,\text{occurs in}\, \pi\}\,.$$

- The *length* of a path $\pi = (v_1, e_1, v_2, ..., e_{m-1}, v_m)$ is $\ell(\pi) \stackrel{\mathrm{def}}{=\!=} m - 1$, 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 *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

$$B_r(v) \stackrel{\mathrm{def}}{=\!=} \{v' \in \mathrm{v}[G] \,|d_G(v, v') = r\}\,.$$

- The *boundary* $\partial_G V$ in G of a set $V \subset \mathrm{v}[G]$ is

$$\partial_G V \stackrel{\mathrm{def}}{=\!=} \{\, v' \in \mathrm{v}[G] \,|\, \exists v \in V :\, d_G(v, v') = 1\,\}.$$

- The *closure* in G of $V \subset \mathrm{v}[G]$, $\overline{V}$, is given by $\overline{V} \stackrel{\mathrm{def}}{=\!=} V \dot{\cup}\, \partial_G V$.

Notation. In the sequel we write ∂ instead of $\partial_{\mathcal{C}_n}$.

We proceed now by giving the mathematical definition of a configuration space which has been discussed in the previous section.

Definition 3. *A sequence of graphs* $(\mathcal{C}_n)_{n\in\mathbb{N}}$ *is called a sequence of configuration spaces if the following assertions hold*

(O) *Each graph* $\mathcal{C}_n$ *is a* γ_n*-regular, connected graph such that*

(i) $\gamma_n \nearrow \infty$

(ii) *For* $\ell \in \mathbb{N}$ *it holds* $|\,\mathcal{C}_n\,|^{-1}\,\gamma_n^{\ell} \searrow 0$.

(P) *For* $v, v' \in \mathrm{v}[\mathcal{C}_n]$ *it holds* $\lim_{n\to\infty} |\,\partial_{\mathcal{C}_n}\{v\} \cap \partial_{\mathcal{C}_n}\{v'\}\,|\,\gamma_n^{-1} = 0$.

(Q) *For all* $v_0 \in \mathrm{v}[\mathcal{C}_n]$ *and* $h \in \mathbb{N}$ *there exists a constant* $c(h) > 0$ *such that* $\lim_{n\to\infty} |\,\{v \in \mathrm{v}[\mathcal{C}_n] | d(v_0, v) = h\}\,|\,\gamma_n^{-h} = c(h)$.

(R) *Let* k *be a fixed natural number and* $v, v' \in \mathrm{v}[\mathcal{C}_n]$ *such that* $1 < d(v, v') = k$. *Then there exists a* $m \in \mathbb{N}$ *and a set of paths* $\mathbf{P}_{\mathcal{C}_n}^{v,v',m} \subset \Pi(\mathcal{C}_n)$ *with the following properties*

(i) $\lim_{n\to\infty} |\,\mathbf{P}_{\mathcal{C}_n}^{v,v',m}\,| = \lim_{n\to\infty} \gamma_n$.

(ii) *For* $\pi \in \mathbf{P}_{\mathcal{C}_n}^{v,v',m}$ *we have* $\mathrm{Supp}(\pi) \cap \partial\{v\} \neq \emptyset$ *and* $\mathrm{Supp}(\pi) \cap \partial\{v'\} \neq \emptyset$.

(iii) *For* $\pi, \pi' \in \mathbf{P}_{\mathcal{C}_n}^{v,v',m}$ *holds*

$$\pi \neq \pi' \implies \mathrm{Supp}(\pi) \cap \mathrm{Supp}(\pi') = \emptyset \quad \textit{and} \quad d(v, v') \leq \ell(\pi) \leq d(v, v') + m\,.$$

(iv) *Let* ℓ *be a fixed natural number,* $\Phi_v \subset \partial\{v\}$ *and* $\Phi_{v'} \subset \partial\{v'\}$ *such that for all* $m_1 \in \mathbb{N}$ *we have* $\lim_{n\to\infty} |\,\Phi_v\,| > m_1$ *and* $\lim_{n\to\infty} |\,\Phi_{v'}\,| > m_1$. *Then there exist* ℓ *pairs of vertices* $((v_1^{(i)}, v_1'^{(i)}))_{1\leq i\leq \ell}$, $v_1^{(i)} \in \Phi_v$, $v_1'^{(i)} \in \Phi_{v'}$ *with the following property:* $\forall\, 1 \leq i \neq j \leq \ell$:

$$|\,\{(\pi, \pi') \in \mathbf{P}_{\mathcal{C}_n}^{v_1^{(i)}, v_1'^{(i)}, m} \times \mathbf{P}_{\mathcal{C}_n}^{v_1^{(j)}, v_1'^{(j)}, m} \,|\, \mathrm{Supp}(\pi) \cap \mathrm{Supp}(\pi') \neq \emptyset\,\}\,| = \theta_n \gamma_n$$

where $\lim_{n\to\infty} \theta_n = 0$.

Example: Let $\mathcal{A}$ be a finite set with $|\,\mathcal{A}\,| \stackrel{\mathrm{def}}{=\!=} \alpha$ and $\mathcal{Q}_\alpha^n$ be the graph with following vertex and edge set:

$$\mathrm{v}[\mathcal{Q}_\alpha^n] \stackrel{\mathrm{def}}{=\!=} \{(x_1, ..., x_n) \,|\, v_i \in \mathcal{A},\, 1 \leq i \leq n\,\}$$
$$\mathrm{e}[\mathcal{Q}_\alpha^n] \stackrel{\mathrm{def}}{=\!=} \{\{(x_1, ..., x_n), (x_1', ..., x_n')\} \,|\, \text{for exactly one index} \, 1 \leq i \leq n \text{ holds } x_i \neq x_i'\,\}$$

Then we call $\mathcal{Q}_\alpha^n$ a *generalized hypercube* over the "alphabet" $\mathcal{A}$ with "sequence-length" n. We write for short $v = (x_1, ..., x_n)$ and proceed by verifying that the sequence of graphs $(\mathcal{Q}_\alpha^n)$ is in fact a sequence of configuration spaces as introduced in definition 3.

We have first $|\,\partial\{v\}\,| = \delta_v = (\alpha - 1)\,n$, hence $\mathcal{Q}_\alpha^n$ is a $(\alpha - 1)\,n$ regular graph. Furthermore $\mathcal{Q}_\alpha^n$ is obviously connected–for given $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ one only has to substitute "step by step" (in arbitrary order) the coordinates in which the vertices differ.

- (O) Clearly we have $\lim_{n\to\infty}(\alpha - 1)\,n \nearrow \infty$ and since $|\,\mathcal{Q}_\alpha^n\,| = \alpha^n$ we inspect $\lim_{n\to\infty}[(\alpha - 1)\,n]^{-i}\,\alpha^n \nearrow \infty$ $i \in \mathbb{N}$.
- (P) $|\,\partial\{v\} \cap \partial\{v'\}\,| = (\alpha - 2)$ since being adjacent to the two vertices v, v' implies that there are exactly $\alpha - 2$ possible choices for the corresponding coordinate left.
- (Q) For arbitrary $r \in \mathbb{N}_n$ holds $|\,B_r(v)\,| = \binom{n}{r}\,(\alpha - 1)^r$ therefore: $\lim_{n\to\infty} |\,B_r(v)\,| = \lim_{n\to\infty}(\alpha - 1)^r\,n^r$.
- (R) For $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ with $d(v, v') = k$ we write v, v' as

$$v = (x_1, ..., x_n) \quad \text{and} \quad v' = (x_1', ..., x_k', x_k, ..., x_n)\,.$$

Let $v_1 \in \partial\{v\} \cap B_{1+k}(v')$ we set

$$g_j(v_1) \stackrel{\text{def}}{=\!=} (x_1, ..., x_j, x'_{j+1}, ..., x'_k, x_{k+1}, .., \hat{x}_r, ..., x_n) \qquad 0 \leq j \leq k \;\; \hat{x}_r \neq x_r \tag{4}$$

and inspect $g_k(v_1) = v_1$ and $g_0(v_1) \in B_1(v') \cap B_{1+k}(v)$. We shall show that we can choose $m = 0$ and $\mathbf{P}_{\mathcal{Q}^n_\alpha}^{v,v',0}$ to be the following set of paths

$$\pi(v_1) \stackrel{\text{def}}{=\!=} (g_k(v_1), g_{k-1}(v_1), ..., g_1(v_1), g_0(v_1))\,, \quad v_1 \in B_1(v) \cap B_{1+k}(v')\,.$$

First $\lim_{n\to\infty} |\, B_1(v) \cap B_{k+1}(v') \,| = (\alpha - 1) \lim_{n\to\infty} n$ implies (R) (i) and (R) (ii) is trivial. For $v_1, \tilde{v}_1 \in B_1(v) \cap B_{1+k}(v')$ we observe immediately

$$v_1 \neq \tilde{v}_1 \quad \Longrightarrow \mathrm{Supp}(\pi(v_1)) \cap \mathrm{Supp}(\pi(\tilde{v}_1)) = \emptyset\,,$$

proving (R) (iii). Further we have for all $v_1 \in B_1(v) \cap B_{1+k}(v')$ $\ell(\pi(v_1)) = k$ verifying that all paths have equal length i.e. $m = 0$.
To verify R (iv) we first inspect $|\, \partial\{v\} \cap \partial\{v'\} \,| = (\alpha - 2)$ where $v, v' \in \mathrm{v}[\mathcal{Q}^n_\alpha]$. Let us assume $\Phi_v \subset \partial\{v\}, \Phi_{v'} \subset \partial\{v'\}$ and $\lim_{n\to\infty} |\, \Phi_v \,| > m_1$, $\lim_{n\to\infty} |\, \Phi_{v'} \,| > m_1$ for arbitrary $m_1 \in \mathbb{N}$. Then there are $2\,\ell$ vertices $v_1^{(i)} \in \partial\{v\}$ $v_1'^{(i)} \in \partial\{v'\}$, $1 \leq i \leq \ell$ that differ from v and v' respectively in exactly 2ℓ *different* coordinates.
The corresponding paths $\mathbf{P}_{\mathcal{Q}^n_\alpha}^{v_1^{(i)}, v_1'^{(i)}, 0}$ $1 \leq i \leq \ell$ have all equal length $k+2$ and fulfill (R) (iv):

$$\text{For } i \neq j: \quad |\, \{(\pi, \pi') \in \mathbf{P}_{\mathcal{Q}^n_\alpha}^{v_1^{(i)}, v_1'^{(i)}, 0} \times \mathbf{P}_{\mathcal{Q}^n_\alpha}^{v_1^{(j)}, v_1'^{(j)}, 0} \,|\, \mathrm{Supp}(\pi) \cap \mathrm{Supp}(\pi') \neq \emptyset \,\}\,| \leq (k+2)\,(\alpha - 1)\,.$$

The above example implies that the following sequences of graphs are sequences of configuration spaces:

(1) the family of Boolean hypercubes $(\mathcal{Q}^n_2)_{n\in\mathbb{N}}$;
(2) the family of generalized hypercubes $(\mathcal{Q}^n_\alpha)_{n\in\mathbb{N}}$;
(3) the family of the canonical configuration spaces for the graph bipartitioning problem [39] i.e. the Johnson graphs $J(n, n/2)$.

Next, for the convenience of the reader we recall some basic terminology from probability theory.

Notation. A *probability space* $(A, \mathcal{A}, \boldsymbol{\mu})$ is a triple consisting of a *point set* A, a *Borel–algebra* $\mathcal{A}$ and a *(probability) measure* $\boldsymbol{\mu}$. In our situation A is a finite set and the Borel-algebra is simply the *power set* $\mathcal{P}(A)$ of A. The measure of an arbitrary set $M \in \mathcal{P}(A)$ is then given as the sum over the point measures: $\boldsymbol{\mu}\{\, M \,\} = \sum_{a\in M} \boldsymbol{\mu}\{\, a \,\}$.
A *random variable* $\hat{X}$ is a $\boldsymbol{\mu}$–measurable function on A. The *distribution* of the random variable $\hat{X}$ is determined by the (cumulative) distribution function $F(x) = \boldsymbol{\mu}\{\, \hat{X} < x \,\}$, where $-\infty < x < \infty$. In the case of integer-valued random variables, we can specify them as well by the probability density function $f(x) = \boldsymbol{\mu}\{\, \hat{X} = x \,\}$.
The *expectation value* of $\hat{X}$ is given by $\mathbf{E}[\hat{X}] = \int x dF(x)$, which reduces to $\mathbf{E}[\hat{X}] = \sum_x k f(x)$ on finite sets A. The *variance* of $\hat{X}$ is defined by $\mathbf{V}[\hat{X}] = \mathbf{E}[(\hat{X} - \mathbf{E}[\hat{X}]\,)^2]$.
We shall make use of the r–th *factorial moment* of a positive, integer valued random variable $\hat{Y}$ which is defined by:

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

where $[j]_r = j \cdot (j-1) \cdot ... \cdot (j - (r-1))$. If $\hat{Y}$ counts vertices with a certain property, then $\mathbf{E}[\hat{Y}]_{\,r}$ can be interpreted as the expected number of *ordered* r tuples of vertices which all have this property.

3.1.2. Model and some Basic Facts

We shall now introduce a probability space that is composed by random induced subgraphs.

Model *Let H be a finite graph. Each subset $X \subset \mathrm{v}[H]$ induces the subgraph $H[X]$ and we have a one-to-one correspondence between $X \subset \mathrm{v}[H]$ and $H[X]$. Let us denote the set of all induced subgraphs of H by $\mathcal{G}(H)$. Further we suppose $0 \leq \lambda \leq 1$ to be given and set for $\Gamma \in \mathcal{G}(H)$*

$$\boldsymbol{\mu}_\lambda\{\Gamma\} \stackrel{\text{def}}{=\!=} \lambda^{|\,\mathrm{v}[\Gamma]\,|}(1-\lambda)^{|\,\mathrm{v}[H]\,|-|\,\mathrm{v}[\Gamma]\,|}\,.$$

Obviously $\boldsymbol{\mu}_\lambda$ fulfills $\sum_\Gamma \boldsymbol{\mu}_\lambda\{\Gamma\} = 1$ and we thereby obtain the probability space

$$\Omega_n \stackrel{\text{def}}{=\!=} (\,\mathcal{G}(H), \mathcal{P}(\mathcal{G}(H)), \boldsymbol{\mu}_\lambda\,)\,.$$

Remark. We can construct each $\Gamma \in \mathcal{G}^{\mathrm{II}}(H)$ by selecting each vertex $v \in H$ with the independent probability $0 < \lambda \leq 1$. This leads to the set V_λ. Then Γ is the induced subgraph of V_λ in H i.e.

$$\Gamma = H[V_\lambda]\,.$$

Remark. In the sequel we shall consider a sequence of graphs, more precisely a sequence of configuration spaces $(\mathcal{C}_n)_{n\in\mathbb{N}}$ as introduced in definition 3. This leads to a family of corresponding probability spaces Ω_n. If we want to emphasize that we are working with random subgraphs of a graph $\mathcal{C}_n$ we shall write Γ_n. Accordingly we shall refer to the underlying probability measures as $\boldsymbol{\mu}_{n,\lambda}$ or simply $\boldsymbol{\mu}_n$.

Notation: Suppose a sequence of configuration spaces $(\mathcal{C}_n)$ is given and Ω_n is the associated sequence of probability spaces according to the above model.
A *property of graphs* $\mathbf{Q}_n$ is a *subset* $\mathbf{Q}_n \subset \Omega_n$ such that $\Gamma_n \in \mathbf{Q}_n$ and $\Gamma_n \cong \Gamma'_n$ then $\Gamma'_n \in \mathbf{Q}_n$.
If $\mathbf{Q}_n$ is a property of graphs and $\Gamma_n \in \mathbf{Q}$ then we say "Γ_n has the property $\mathbf{Q}_n$". We shall further use the term "almost every (no) random graph Γ_n has the property $\mathbf{Q}_n$" if and only if $\boldsymbol{\mu}_n\{\mathbf{Q}_n\} = 1$ ($\boldsymbol{\mu}_n\{\mathbf{Q}_n\} = 0$). The term "asymptotically almost every (no) random graph Γ_n has the property $\mathbf{Q}_n$" means $\lim_{n\to\infty} \boldsymbol{\mu}_n\{\mathrm{Q}_n\} = 1$ ($\lim_{n\to\infty} \boldsymbol{\mu}_n\{\mathrm{Q}_n\} = 0$) and we express this by saying *a.a.s. Γ_n has property $\mathbf{Q}_n$*.

Next we show that for the random graphs Γ_n the distribution of vertex degrees is asymptotically *invariant*. In the language of statistical physics this means that the vertex degree is a *self-averaging quantity*. Let $k = k(n) > 0$ be an integer valued function. We introduce the random variables

$$\hat{X}_{n,k} \stackrel{\text{def}}{=\!=} |\,\{v \in \mathrm{v}[\Gamma_n] \,|\, \hat{\delta}_v = k\}\,|\,, \tag{5}$$

counting the number of vertices with degree k in a random graph Γ_n.

Proposition 1. *Let $(\mathcal{C}_n)_n$ be a sequence of configuration spaces such that $\lim_{n\to\infty} B(k,\gamma_n,\lambda)\,\mathbf{E}[\hat{\omega}_n] = \infty$. Then we have*

$$\lim_{n\to\infty} \boldsymbol{\mu}_{n,\lambda}\{\hat{X}_{n,k} = B(k,\gamma_n,\lambda)\,\lambda\,|\,\mathcal{C}_n\,|\,\} = 1\,.$$

Proof. We first observe

$$\lim_{n\to\infty} \mathbf{E}[\hat{X}_{n,k}] \;=\; \lim_{n\to\infty} \sum_{\ell=1}^{|\mathcal{C}_n|} \boldsymbol{\mu}\{\,\hat{\delta}_v = k\,\}\,\ell\,\boldsymbol{\mu}_{n,}\{\hat{\omega}_n = \ell\} \;=\; \lim_{n\to\infty} \mathcal{B}(k,\gamma_n,\lambda)\,\mathbf{E}[\hat{\omega}_n]\;.$$

Then we compute the variance of $\hat{X}_{n,k}$ by evaluating the second factorial moment $\mathbf{E}[\hat{X}_k]_{\,2}$ (i. e. the expectation value of ordered 2-tuples (v,v') in random graphs Γ_n of vertices v with vertex degree $\delta_v = k$). Property (P) of definition 3 implies that the probability for selecting a pair of vertices with degree k is asymptotically given by $\mathcal{B}(k,\gamma_n,\lambda)^2$ and therefore

$$\lim_{n\to\infty} \mathbf{E}[\hat{X}_{n,k}]_{\,2} = \lim_{n\to\infty} \sum_{\ell=2}^{|\mathcal{C}_n|} \ell\,(\ell-1)\,\mathcal{B}(k,\gamma_n,\lambda)^2\,\boldsymbol{\mu}_{n,}\{\hat{\omega}_n = \ell\}\,,$$

since there are $\ell\,(\ell-1)$ ordered pairs of vertices of degree k in a graph Γ_n with $|\,\Gamma_n\,| = \ell$. Rewriting this as $\lim_{n\to\infty}\mathbf{E}[\hat{X}_{n,k}]_{\,2} = \lim_{n\to\infty}(\mathcal{B}(k,\gamma_n,\lambda)\,\mathbf{E}[\hat{\omega}_n]\,)^2$ we end up with
$\mathbf{E}[\hat{X}^2_{n,k}] \;= \mathbf{E}[\hat{X}_{n,k}]_{\,2} + \mathbf{E}[\hat{X}_{n,k}]$ and thus

$$\lim_{n\to\infty} \mathbf{V}[\hat{X}_{n,k}] \;=\; \lim_{n\to\infty} (\mathbf{E}[\hat{X}_{n,k}] \;+ [\theta_n\,\mathbf{E}[\hat{X}_{n,k}]\,]^2) \quad \text{where } \lim_{n\to\infty}\theta_n = 0.$$

By assumption we have $\lim_{n\to\infty}\mathbf{E}[\hat{X}_{n,k}] \;= \infty$ and the proposition follows. ∎

3.2. Density

In this section we shall suppose that a family of configuration spaces $(\mathcal{C}_n)$ is given. According to the model we consider for each graph $\mathcal{C}_n$ the set of all induced subgraphs Γ_n of $\mathcal{C}_n$ and the probability measure

$$\boldsymbol{\mu}_{n,\lambda}\{\Gamma_n\} = \lambda^{\mathrm{v}[\Gamma_n]}(1-\lambda)^{\mathrm{v}[\mathcal{C}_n]-\mathrm{v}[\Gamma_n]}\,.$$

Assuming that $\boldsymbol{\mu}_{n,\lambda}$ and $\mathcal{C}_n$ are fixed we speak of *random graphs* Γ_n. We shall write in the sequel Ω_n and $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n,\lambda}$.

Definition 4. *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 shall discuss the in this section the density property of random graphs $\Gamma_n < \mathcal{C}_n$ for underlying $0 < \lambda < 1$ and establish the existence of a "critical" λ-value, λ^* that has the following property:

- for $\lambda < \lambda^*$ in the limit almost no random graph Γ_n is dense and
- for $\lambda > \lambda^*$ in the limit almost every random graph Γ_n is dense.

For this purpose let

$$\hat{Z}_n(\Gamma_n) \;\stackrel{\text{def}}{=\!=}\; |\,\{v \in \mathrm{v}[\mathcal{C}_n] \mid v \notin \overline{\mathrm{v}[\Gamma_n]}\}\,| \tag{6}$$

defined on Ω_n, be the random variable that counts the number of vertices having *no* adjacent vertex v in $\mathcal{C}_n$ with $v \in \mathrm{v}[\Gamma_n]$. We first compute the asymptotical distribution of the following sequence of random variables $(\hat{Z}_n)$ associated to the sequence of probability spaces (Ω_n).

Lemma 1. *Let $(\mathcal{C}_n)_n$ be a family of configuration spaces. Suppose*

$$\mu \stackrel{\text{def}}{=\!=} \lim_{n\to\infty} (|\,\mathcal{C}_n\,|\,(1-\lambda)^{\gamma_n+1}) \in I\!R_+ \cup \{0\} \cup \{\infty\}\,.$$

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

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n = \ell\} = \frac{\mu^\ell}{\ell!}\, e^{-\mu} \tag{7}$$

and in particular we have

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n = 0\} = e^{-\mu}, \qquad \textit{and} \qquad \lim_{n\to\infty} \mathbf{E}[\hat{Z}_n] = |\,\mathcal{C}_n\,|\,(1-\lambda)^{\gamma_n+1}\,.$$

For $\mu = \infty$ and $\ell \in I\!N$ holds

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

Proof. We first consider the case $\mu \in \mathbb{R}_+ \cup \{0\}$. According to corollary 9 in the appendix it suffices to show that $\lim_{n\to\infty} \mathbf{E}[\hat{Z}_n]_{\,r} = \mu^r$ holds for all $r \in \mathbb{N}$.
For each ordered r-tuple $(v_1, ..., v_r)$ of vertices the number $N(r)$ of adjacent vertices in $\mathrm{v}[\mathcal{C}_n]$, fulfills $r\,(\gamma - r) \leq N(r) \leq r\,\gamma$. There are at most $\binom{|\,\mathcal{C}_n\,|}{r-1}\, r\,\gamma$ sets of r-vertices with less than $\gamma\, r$ adjacent vertices in $\mathrm{v}[\mathcal{C}_n]$: In fact, we may choose $r-1$ vertices arbitrarily and the last one has to be adjacent to at least one of the others. Since each vertex $v \in \mathrm{v}[\mathcal{C}_n]$ fulfills $v \in \mathrm{v}[\Gamma_n]$ with the independent probability λ and r is fixed

$$\sum_{\ell=1}^{|\,\mathcal{C}_n\,|} [\,|\,\mathcal{C}_n\,| - |\,\Gamma_n\,|\,]_r\; (1-\lambda)^{\gamma_n\, r} \boldsymbol{\mu}_n\{|\,\Gamma_n\,| = \ell\} \leq \mathbf{E}[\hat{Z}_n]_{\,r} \leq \sum_{\ell=1}^{|\,\mathcal{C}_n\,|} R_{n,r}\, \boldsymbol{\mu}_n\{|\,\Gamma_n\,| = \ell\}\,,$$

where

$$R_{n,r} \stackrel{\text{def}}{=\!=} \Big[[\,|\,\mathcal{C}_n\,| - |\,\Gamma_n\,|\,]_r (1-\lambda)^{\gamma_n\, r} + (|\,\mathcal{C}_n\,| - |\,\Gamma_n\,|)^{r-1} (r-1)\gamma_n (1-\lambda)^{(\gamma_n - r)r}\Big]\,.$$

In the proof of proposition 1 we have verified that $\mathbf{E}[\hat{\omega}_n] = \lambda\,|\,\mathcal{C}_n\,|$ and the above inequality becomes

$$\begin{aligned}(1 - a\,\frac{r^2}{|\,\mathcal{C}_n\,|})\,[(1-\lambda)\,|\,\mathcal{C}_n\,|\,(1-\lambda)^{\gamma_n}]^r &\leq \mathbf{E}[\hat{Z}_n]_{\,r} \\ &\leq [(1-\lambda)\,|\,\mathcal{C}_n\,|\,(1-\lambda)^{\gamma_n}]^r \,\Big[1 + a'\,|\,\mathcal{C}_n\,|^{-1}\,(r-1)\,\gamma_n (1-\lambda)^{-r^2}\Big] \quad a, a' \in \mathbb{R}_+\,.\end{aligned}$$

Taking the limit $n \nearrow \infty$ this implies

$$\lim_{n\to\infty} \mathbf{E}[\hat{Z}_n]_{\,r} = \lim_{n\to\infty} \big[\,|\,\mathcal{C}_n\,|\,(1-\lambda)^{\gamma_n+1}\big]^r$$

and the first statement is proved.
For $\mu = \infty$ the above argument shows that

$$\mathbf{E}[\hat{Z}_n] \nearrow \infty \quad \text{ and } \quad \lim_{n\to\infty} \mathbf{E}[\hat{Z}_n]_{\,2} = \lim_{n\to\infty} \mathbf{E}[\hat{Z}_n]^{\,2}\,.$$

Since $\mathbf{E}[\hat{Z}_n]_{\,2} + \mathbf{E}[\hat{Z}_n] - \mathbf{E}[\hat{Z}_n]^{\,2} = \mathbf{V}[\hat{Z}_n]$ we obtain $\mathbf{V}[\hat{Z}_n] = \theta_n\, \mathbf{E}[\hat{Z}_n]^{\,2}$ with $\lim_{n\to\infty} \theta_n = 0$. Therefore for every $\ell \in \mathbb{N}$ holds $\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n \geq \ell\} = 1$ and the lemma is proved. ∎

Theorem 1. *Let $(\mathcal{C}_n)_n$ be a sequence of configuration spaces and for each $n \in \mathbb{N}$ we have*

$$0 < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1\,.$$

Suppose $\lambda^* \stackrel{\text{def}}{=\!=} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$ then for $\lambda > \lambda^*$ holds

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \textit{ is dense in } \mathcal{C}_n\} = 1$$

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

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

In the basic terminology of random graph theory λ^* is called *threshold value* for the density property. For $\lambda > \lambda^*$ almost every random graph Γ_n is dense in $\mathcal{C}_n$ and almost no Γ_n is dense in $\mathcal{C}_n$ for $\lambda < \lambda^*$.

Proof. According to lemma 1 we have $\lim_{n\to\infty} \mathbf{E}[\hat{Z}_n] = \lim_{n\to\infty} |\,\mathcal{C}_n\,|\,(1-\lambda)^{\gamma+1}$. We further inspect from the lemma that

$$\lim_{n\to\infty} \mathbf{E}[\hat{Z}_n] = \begin{cases} 0 & \text{for } \lambda > \lambda^* \\ \infty & \text{for } \lambda < \lambda^*\,. \end{cases}$$

Therefore for $\lambda > \lambda^*$ we have the case $\mu = 0$ of lemma 1 and thus $\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n = 0\} = 1$ since $\hat{Z}_n$ is Poisson.
For $\lambda < \lambda^*$ we have $\mu = \infty$ of lemma 1 and obtain for $\ell \in \mathbb{N}$ $\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n \geq \ell\} = 1$. By definition of $\hat{Z}_n$ holds $\{\hat{Z}_n = 0\} = \{\Gamma_n \text{ is dense }\}$ and the theorem is proved. ∎

Remark. Suppose $\mathcal{C}_n$ is a generalized hypercube i.e. $\mathcal{C}_n = \mathcal{Q}^n_\alpha$, the formula for the density threshold reads

$$\lambda^* = 1 - \sqrt[\alpha^{-1}]{\alpha^{-1}}\,.$$

3.3. Connectedness and the Sequence of Components

3.3.1. Definitions and two Auxiliary Lemmata

We assume in this section that for each $\mathcal{C}_n$ holds $0 < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$. Furthermore we set

$$\lambda^* \stackrel{\text{def}}{=\!=} 1 - |\,\mathcal{C}_n\,|^{\frac{-1}{\gamma_n}}\,.$$

Let G be a finite graph. Recall that two vertices $v, v' \in \mathrm{v}[G]$ are connected if there exists a path in G in which v and v' occur. Further a graph G is connected, if for all pairs of vertices $v, v' \in \mathrm{v}[G]$, there exists a path in G in which v, v'' occur and it is *disconnected* otherwise. 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 $V = \emptyset$ then $G[\emptyset]$ 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.

Definition 5. *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 $|\mathcal{X}| \geq 2/3\,|G|$.*

For a random graph Γ_n an *isolated vertex* $v \in \mathrm{v}[\Gamma_n]$ is a vertex with the property $\partial\{v\} \cap \mathrm{v}[\Gamma_n] = \emptyset$. We consider the random variable $\hat{U}_I$ defined on Ω_n that counts the total number of components $\mathcal{X}$ in a random graph Γ_n with orders in the interval $I \subseteq \mathbb{N}$. Analogously we make use of the notation $\hat{U}_\ell$ for the number of components of order ℓ. The following lemma implies already that Γ_n is a.a.s. disconnected for $\lambda < \lambda^*$.

Lemma 2. *Suppose $k \in I\!N$ $v \in \mathrm{v}[\mathcal{C}_n]$ and $0 \leq \lambda \leq 1$. Then we have for $\lambda > \lambda^*$*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \textit{ contains no components with } 1 \leq |\mathcal{X}| \leq \gamma_n\} = 1$$

and for $\lambda < \lambda^*$ and $\ell \in I\!N$ holds

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \textit{ contains at least } \ell \textit{ components with } 1 \leq |\mathcal{X}| \leq \gamma_n\} = 1\,.$$

Proof. Obviously the smallest nontrivial component is an isolated vertex. For $v \in \mathrm{v}[\Gamma_n]$ the probability of $\partial\{v\} \cap \mathrm{v}[\Gamma_n] = \emptyset$ is $(1-\lambda)^{\gamma_n}$. Let $\hat{I}_n$ be the random variable defined by

$$\hat{I}_n(\Gamma_n) \stackrel{\text{def}}{=\!=} |\,\{v \in \mathrm{v}[\Gamma_n] \,|\, \partial\{v\} \cap \mathrm{v}[\Gamma_n] = \emptyset\,\}\,|\,.$$

Then $\mathbf{E}[\hat{I}_n] = \mathbf{E}[\hat{\omega}_n]\,(1-\lambda)^{\gamma_n}$ and in complete analogy to the proof of theorem 1 it can be shown that

$$\mathbf{E}[\hat{I}_n] = \begin{cases} 0 & \text{for } \lambda > \lambda^* \\ \infty & \text{for } \lambda < \lambda^*\,. \end{cases}$$

Further we inspect for $\lambda < \lambda^*$: $\lim_{n\to\infty} \mathbf{E}[\hat{I}_n]_2 = \lim_{n\to\infty} \mathbf{E}[\hat{I}_n]^2$ and thus $\mathbf{V}[\hat{I}_n] = \theta_n\,\mathbf{E}[\hat{I}_n]^2$ with $\lim_{n\to\infty}\theta_n = 0$. This fact implies that for $\lambda < \lambda^*$ and $\ell \in \mathbb{N}$ holds

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ has more than } \ell \text{ isolated vertices}\} = 1\,.$$

What remains to be proven is the nonexistence of components smaller than γ_n in the case $\lambda > \lambda^*$. Clearly there are $\binom{|\mathcal{C}_n|}{\ell}$ different subsets of vertices X with $|X| = \ell$ to select. To each $X \subset \mathrm{v}[\mathcal{C}_n]$ there corresponds the induced subgraph in $\mathcal{C}_n$: $\mathcal{C}_n[X]$. We now proceed by evaluating an upper bound on the number of different subgraphs $\mathcal{C}_n[X]$ that are connected. For the first vertex we can choose every $v \in \mathrm{v}[\mathcal{C}_n]$ and for each subsequent vertex there are at most γ_n possible choices. Therefore

$$|\,\{X \,|\, \mathcal{C}_n[X] \text{ is connected}\}\,| \leq |\,\mathcal{C}_n\,|\,\gamma_n^{|\mathcal{C}_n[X]|-1}$$

and there exists a $b \in \mathbb{R}_+$ such that $|\,\{X \,|\, \mathcal{C}_n[X] \text{ is connected}\}\,| \leq e^{b\ln(\gamma_n)\,|\mathcal{C}_n[X]|}$.
Suppose now $|X| \leq \gamma_n$ and $\mathcal{C}_n[X]$ is connected. The probability that $\mathcal{C}_n[X]$ is a component of a random graph Γ_n is bounded from above by $(1-\lambda)^{|\partial X|}$ since necessarily holds $\partial X \cap \mathrm{v}[\Gamma_n] = X$. In other words no vertex $v \in \mathrm{v}[\Gamma_n]$ can be contained in the $\mathcal{C}_n$-boundary ∂X.
Suppose $|X| \leq \gamma_n$ then we inspect using property (Q) of definition 3

$$\exists\, a \in \mathbb{R}_+ : \qquad a\,\gamma_n |X| \leq |\partial X|\,.$$

Accordingly, $(1-\lambda)^{a\,\gamma_n\,|\mathcal{C}_n[X]\,|}$ serves as an upper bound on the probability for the existence of a component of order $|\,\mathcal{C}_n[X]\,|$ in a random graph Γ_n. We obtain

$$\lim_{n\to\infty}\mathbf{E}[\hat{U}_{[0,\gamma_n]}] \leq \lim_{n\to\infty}\sum_{\ell=1}^{\gamma_n} e^{b\,\ln\gamma_n\,\ell}\;(1-\lambda)^{a\,\gamma_n\,\ell}=0 \quad a,b\in\mathbb{R}_+$$

and the proof of the lemma is complete. ∎

We proceed investigating under what conditions the random graphs $\Gamma_n < \mathcal{Q}_\alpha^n$ are connected. We shall establish that the parameter $\lambda^* = 1-|\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$ is not only a "critical"parameter for the density but also a "critical" parameter for the connectedness.

According to lemma 2 we can restrict ourselves to the case $\lambda > \lambda^*$ since for $\lambda < \lambda^*$ holds

$$\forall\ell\in\mathbb{N}: \quad \lim_{n\to\infty}\boldsymbol{\mu}_n\{\Gamma_n \text{ has more than } \ell \text{ nontrivial components}\}=1\,.$$

Thus it remains to prove that $\lambda > \lambda^*$ implies

$$\lim_{n\to\infty}\boldsymbol{\mu}_n\{\Gamma_n \text{ is connected}\}=1\,.$$

The first step is

Lemma 3. *Let $\Gamma_n < \mathcal{C}_n$ be a random graph $\Gamma_n\in\Omega_n$ and*

$$1-|\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}<\lambda$$

Then

$$\forall\ell\in I\!N \quad \lim_{n\to\infty}\boldsymbol{\mu}_n\{\Gamma_n\,|\,\min_{v\in\mathrm{v}[\Gamma_n]}\delta_v\geq\ell\}=1\,.$$

Proof. For $k\in\mathbb{N}$ we consider the random variable $\hat{Y}_k$ on Ω_n whose states are the numbers of vertices $v\in\mathrm{v}[\Gamma_n]$ with $\delta_v=k$. We proceed by estimating the expectation value $\mathbf{E}[\hat{Y}_k]$. An upper bound for $\mathbf{E}[\hat{Y}_k]$ is given by $|\,\mathcal{C}_n\,|\binom{\gamma_n}{k}\lambda^k\,(1-\lambda)^{\gamma_n-k}$ since there are at most $|\,\mathcal{C}_n\,|$ vertices to select. Further λ^* solves for x:

$$\lim_{n\to\infty}\left[|\,\mathcal{C}_n\,|\binom{\gamma_n}{k}x^k\,(1-x)^{\gamma_n-k}\right]=1$$

and for $\lambda>\lambda^*$ we obtain

$$\forall\,\ell\in\mathbb{N}: \quad \lim_{n\to\infty}\left[\sum_{k=0}^{\ell}|\,\mathcal{C}_n\,|\binom{\gamma_n}{k}\lambda^k\,(1-\lambda)^{\gamma_n-k}\right]=0\,.$$

Therefore

$$\forall\ell\in\mathbb{N} \quad \lim_{n\to\infty}\boldsymbol{\mu}_n\{\Gamma_n\,:\,\min_{v\in\mathrm{v}[\Gamma_n]}\delta_v\geq\ell\}=1$$

and the lemma is proved. ∎

The same argument implies

Corollary 1. *Let $\Gamma_n < \mathcal{C}_n$ be a random graph $\Gamma_n \in \Omega_n$ and $1 - |\mathcal{C}_n|^{-\frac{1}{\gamma_n}} < \lambda$. Then*

$$\forall \ell \in I\!N \quad \lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \min_{v\in \mathrm{v}[\mathcal{C}_n]} |\,\partial\{v\} \cap v[\Gamma]\,| \geq \ell\,\} = 1\,.$$

In order to make the following discussion more transparent, we first analyze connectedness for a special class of configuration spaces namely the general hypercubes (Hamming graphs).

3.3.2. Connectedness in Generalized Hypercubes

In this subsection we shall assume that the sequence of configuration spaces $(\mathcal{C}_n)$ is given by $(\mathcal{Q}^n_\alpha)_{n\in\mathbb{N}})$. Then according to our model the corresponding probability spaces (Ω_n) are formed by all subgraphs Γ_n of $\mathcal{Q}^n_\alpha$ that are of the form $\Gamma_n = \mathcal{Q}^n_\alpha[V]$ where $V \subset \mathrm{v}[\mathcal{Q}^n_\alpha]$ and the probability measure $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n,\lambda}$ is

$$\boldsymbol{\mu}_{n,\lambda}(\Gamma_n) = \lambda^{|\,\mathrm{v}[\Gamma_n]\,|}(1-\lambda)^{\alpha^n - |\,\mathrm{v}[\Gamma_n]\,|}\,.$$

For brevity we shall refer to Γ_n simply as a random graph and we shall write $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n,\lambda}$. Recall that a generalized hypercube $\mathcal{Q}^n_\alpha$ is a $\gamma_n \stackrel{\text{def}}{=\!=} [(\alpha-1)\,n]$-regular graph and

$$\lambda^* = 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} = 1 - \alpha^{\frac{-1}{\alpha-1}}\,.$$

We restate that for $v \in \mathrm{v}[\mathcal{Q}^n_\alpha]$ $B_r(v)$ is the ball centered at v with radius r, i.e.,

$$B_r(v) \stackrel{\text{def}}{=\!=} \{v' \,|\, v' \in \mathrm{v}[\mathcal{Q}^n_\alpha] \wedge d(v,v') = r\}\,.$$

First we estimate "how many independent paths there are in $\mathcal{Q}^n_\alpha$ that connect the balls $B_2(v)$ and $B_2(v')$ under the assumption that v, v' have finite distance k".

Let v and v' be two vertices of $\mathrm{v}[\mathcal{Q}^n_\alpha]$ with $d(v,v') = k$ where k is a fixed natural number. By renumbering of the coordinates we can write

$$v = (x_1, x_2, \ldots, x_n) \quad \text{and} \quad v' = (x'_1, x'_2, \ldots, x'_k, x_{k+1}, ..., x_n)\,. \tag{8}$$

We shall write vertices $v_2 \in B_2(v) \cap B_{2+k}(v')$ and $v'_2 \in B_2(v') \cap B_{2+k}(v)$ as

$$v_2 = (x_1, ..., x_k, x_{k+1}, ..., x_{r-1}, \hat{x}_r, x_{r+1}, ..., x_{s-1}, \hat{x}_s, x_{s+1}, ..., x_n)$$
$$v'_2 = (x'_1, ..., x'_k, x_{k+1}, ..., x_{t-1}, \hat{x}_t, x_{t+1}, ..., x_{u-1}, \hat{x}_u, x_{u+1}, ..., x_n)\,.$$

with $\hat{x}_r \neq x_r$, $\hat{x}_s \neq x_s$, $\hat{x}_t \neq x_t$ and $\hat{x}_u \neq x_u$.

For $0 \leq j \leq k$ and $v_2 \in B_2(v) \cap B_{2+k}(v')$ we set:

$$f_j(v_2) \stackrel{\text{def}}{=\!=} (x_1, ..., x_j, x'_{j+1}, ..., x'_k, v_{k+1}, ..., x_{r-1}, \hat{x}_r, x_{r+1}, ..., x_{s-1}, \hat{x}_s, x_{s+1}, ...)\,. \tag{9}$$

Clearly $f_k(v_2) = v_2$, $f_0(v_2) \in B_2(v') \cap B_{2+k}(v)$ and each family $(f_j(v_2))_{0\leq j\leq k}$ can be identified with a path of length k in $\mathcal{Q}^n_\alpha$, $\pi(v_2)$. By definition $\pi(v_2)$ has a nonempty intersection with $B_2(v) \cap B_{2+k}(v')$ and $B_2(v') \cap B_{2+k}(v)$. Let $v_2, \tilde{v}_2 \in B_2(v) \cap B_{2+k}(v')$ then we have

$$v_2 \neq \tilde{v}_2 \implies \mathrm{Supp}(f_j(v_2)) \cap \mathrm{Supp}(f_j(\tilde{v}_2)) = \emptyset\,,$$

in other words for different $v_2, \tilde{v}_2 \in B_2(v) \cap B_{2+k}(v')$ the corresponding paths $\pi(v_2), \pi(\tilde{v}_2)$ are pairwise disjoint. In fact we have a mapping

$$\begin{array}{rccc} \psi^{v,v'}_{n,k} : & B_2(v) \cap B_{2+k}(v') & \longrightarrow & \Pi(\mathcal{C}_n) \\ & v_2 & \mapsto & \psi^{v,v'}_{n,k} \stackrel{\text{def}}{=\!=} \pi(v_2)\,. \end{array}$$

Next we introduce the random variable $\hat{Z}^{v,v'}_{n,k}$ for $v, v' \in \mathcal{Q}^n_\alpha$ with fixed $d(v,v') = k$.

$$\hat{Z}^{v,v'}_{n,k}(\Gamma_n) \stackrel{\text{def}}{=\!=} \begin{cases} |\,\{\pi(v_2) \,|\, \pi \in \Pi(\Gamma_n),\ v_2 \in B_2(v) \cap B_{2+k}(v')\,\}\,| & \text{for } v, v' \in \mathrm{v}[\Gamma_n] \\ 0 & \text{otherwise}\,. \end{cases}$$

The result is that a.a.s. a random graph Γ_n has the property that for each pair of vertices $v, v' \in \mathrm{v}[\Gamma_n]$ with $d(v,v') = k$, $k \in \mathbb{N}$ there exists on the order of n^2 pairwise disjoint paths in Γ_n "connecting" $B_2(v)$ and $B_2(v')$.

Lemma 4. *Let $\mathcal{Q}^n_\alpha$ be a generalized hypercube and $\Gamma_n < \mathcal{Q}^n_\alpha$ a random graph with underlying probability measure $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n,\lambda}$ and $0 < \lambda \leq 1$. Then for arbitrary $\chi \in \mathbb{R}_+$ holds*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \forall v, v' \in \mathrm{v}[\mathcal{Q}^n_\alpha] : d(v,v') = k,\, k \in \mathbb{N} : \; |\, \hat{Z}^{v,v'}_{n,k} - \mathbf{E}[\hat{Z}^{v,v'}_{n,k}] \,|\, < \chi\, n^2 \,\} = 1\,.$$

Proof. We consider the random variable $\hat{Z}^{v,v'}_{n,k}$. For each $v_2 \in B_2(v) \cap B_{2+k}(v')$ there corresponds the path $\pi(v_2)$ obtained according to equation (9) in $\mathcal{Q}^n_\alpha$. By definition $\pi(v_2)$ has a nonempty intersection with $B_2(v') \cap B_{2+k}(v)$ and $B_2(v) \cap B_{2+k}(v')$. $\pi(v_2)$ is already a path in a random graph Γ_n with probability λ^{k+1}.
For different vertices $v_2, \tilde{v}_2 \in B_2(v) \cap B_{2+k}(v')$ the paths $\pi(v_2)$ and $\pi(\tilde{v}_2)$ fulfill $\mathrm{Supp}(\pi(v_2)) \cap \mathrm{Supp}(\pi(\tilde{v}_2)) = \emptyset$. Accordingly, the random variable $\hat{Z}^{v,v'}_{n,k}$, counting the number of those paths is binomial distributed with expectation value

$$\mathbf{E}[\hat{Z}^{v,v'}_{n,k}] = \lambda^{k+1} \binom{n-k}{2} (\alpha-1)^2\,.$$

By applying corollary 9 in the appendix we obtain

$$\forall\, \chi \in \mathbb{R}_+ \,\exists\, b \in \mathbb{R}_+ : \quad \boldsymbol{\mu}_n\{|\, \hat{Z}^{v,v'}_{n,k} - \mathbf{E}[\hat{Z}^{v,v'}_{n,k}] \,| \geq \chi\, n^2\} \leq e^{-b\,n^2}\,.$$

On the other hand there are at most $\binom{n}{k}\,(\alpha-1)^k\,\alpha^n$ different pairs of vertices in $\mathrm{v}[\mathcal{Q}^n_\alpha]$ with $d(v,v') = k$. We immediately inspect

$$\lim_{n\to\infty} \left[\binom{n}{k} (\alpha-1)^k\, \alpha^n\, \boldsymbol{\mu}_n\{|\, \hat{Z}^{v,v'}_{n,k} - \mathbf{E}[\hat{Z}^{v,v'}_{n,k}] \,| \geq \chi\, n^2\} \right] = 0$$

and consequently for arbitrary $\chi \in \mathbb{R}_+$ holds

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \forall v, v' \in \mathrm{v}[Q^n_\alpha],\, d(v,v') = k \quad |\, \hat{Z}^{v,v'}_{n,k} - \mathbf{E}[\hat{Z}^{v,v'}_{n,k}] \,|\, < \chi\, n^2 \,\} = 1$$

proving the lemma. ∎

Next we show that 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 make use of special paths in $\mathcal{Q}^n_\alpha$ which we introduce now:

For $v, v' \in \mathrm{v}[\mathcal{Q}^n_\alpha]$ with $d(v,v') = k$ we assume v, v' to be given by equation (8). For $v_1 \in \partial\{v\} \cap B_{1+k}(v')$ we set

$$g_j(v_1) \stackrel{\mathrm{def}}{=\!=} (x_1, ..., x_j, x'_{j+1}, ..., x'_k, x_{k+1}, .., \hat{x}_r, ..., x_n) \qquad 0 \leq j \leq k \;\; \hat{x}_r \neq x_r \tag{10}$$

and inspect $g_k(v_1) = v_1$, $g_0(v_1) \in B_1(v') \cap B_{1+k}(v)$. In fact we have a mapping

$$\begin{array}{ccc} \partial\{v\} \cap B_{k+1}(v') & \longrightarrow & \Pi(\mathcal{Q}^n_\alpha) \\ v_1 & \mapsto & \pi(v_1) \stackrel{\mathrm{def}}{=\!=} (g_k(v_1), g_{k-1}(v_1), ..., g_1(v_1), g_0(v_1))\,. \end{array}$$

Let $M^{v,v'}_{n,k}(\Gamma_n)$ be the set of paths

$$M^{v,v'}_{n,k}(\Gamma_n) \stackrel{\mathrm{def}}{=\!=} \{\pi(v_1) \,|\, \pi(v_1) \in \Pi(\Gamma_n)\,\}\,.$$

Then for $v_1, \tilde{v}_1 \in B_1(v) \cap B_{1+k}(v')$ we have

$$v_1 \neq \tilde{v}_1 \quad \Longrightarrow \quad \mathrm{Supp}(\pi(v_1)) \cap \mathrm{Supp}(\pi(\tilde{v}_1)) = \emptyset$$

and further for all $v_1 \in B_1(v) \cap B_{1+k}(v')$ we have $\ell(\pi(v_1)) = k$.
We introduce the random variable

$$\hat{Y}^{v,v'}_{n,k} \stackrel{\mathrm{def}}{=\!=} \begin{cases} |\, M^{v,v'}_{n,k}(\Gamma_n) \,| & \text{for } v, v' \in \mathrm{v}[\Gamma_n] \\ 0 & \text{otherwise}\,. \end{cases}$$

The paths $\pi(v_1)$ in $\mathcal{Q}^n_\alpha$ are pairwise disjoint and each of them is a path in Γ_n with probability λ^{k+1}. Therefore for $v, v' \in \mathrm{v}[\Gamma_n]$ $\hat{Y}^{v,v'}_{n,k}$ is binomially distributed and $|\, B_1(v) \cap B_{1+k}(v') \,| = (\alpha-1)\,(n-k)$ implies

$$\mathbf{E}[\hat{Y}^{v,v'}_{n,k}] = \lambda^{k+1}\,(\alpha-1)\,(n-k)\,.$$

Remark. The following lemma is not simply implied by corollary 9 of appendix A as for example the previous lemma. Here we shall make use of lemma 5 in order to be able to apply our main argument "independently finitely many times".

Lemma 5. *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*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \forall v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n],\, d(v,v') = k : \exists\, v_1 \in \partial\{v\},\, v_1' \in \partial\{v'\} :\, \hat{Y}^{v_1,v_1'}_{n,d(v_1,v_1')} > 0\,\} = 1\,.$$

Proof. Corollary 1 implies that for arbitrary $\ell \in \mathbb{N}$ holds:

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n | \forall\, v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n],\, \exists\, v_1^{(1)}, ..., v_1^{(\ell)} \in \partial\{v\} \cap \mathrm{v}[\Gamma_n] \wedge \exists\, v_1'^{(1)}, ..., v_1'^{(\ell)} \in \partial\{v'\} \cap \mathrm{v}[\Gamma_n]\,\} = 1\,. \quad (11)$$

For $w, w' \in \mathrm{v}[\Gamma_n]$ $\hat{Y}_{n,k}^{w,w'}$ is binomially distributed with expectation value $\mathbf{E}[\hat{Y}_{n,k}^{w,w'}] = \lambda^{k+1}\,(\alpha-1)\,(n-k)$. We apply corollary 9 (see appendix A) and obtain

$$\forall b' \in \mathbb{R}_+ \,\exists\, b \in \mathbb{R}_+ : \qquad \lim_{n\to\infty} \boldsymbol{\mu}_n\{\,|\,\hat{Y}_{n,k}^{w,w'} - \mathbf{E}[\hat{Y}_{n,k}^{w,w'}]\,|\, \geq b'\,n\} \leq e^{-b\,n}\,. \quad (12)$$

Let $\ell \in \mathbb{N}$ be fixed and P be the probability for the existence of a pair of vertices $v, v' \in \mathrm{v}[\mathcal{C}_n]$ such that there are ℓ pairs of vertices $(v_1^{(i)}, v_1'^{(i)})$, $v_1^{(i)} \in B_1(v) \cap \mathrm{v}[\Gamma_n]$, $v_1'^{(i)} \in B_1(v') \cap \mathrm{v}[\Gamma_n]$, $1 \leq i \leq \ell$ with the property:

$$\text{For } 1 \leq i \leq \ell,\, b' \in \mathbb{R}_+ \qquad \hat{Y}^{v_1^{(i)}, v_1'^{(i)}}_{n,d(v_1^{(i)}, v_1'^{(i)})} = 0\,.$$

Claim: There exists a $t \in \mathbb{R}_+$ such that $\lim_{n\to\infty} P \leq \lim_{n\to\infty} e^{-\ell\, t\, n}$.
To prove the claim we first observe that according to equation (11) a.a.s. for each pair $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ and arbitrary $\ell \in \mathbb{N}$ there exist ℓ pairs of vertices $v_1^{(i)}, v_1'^{(i)}$ where $1 \leq i \leq \ell$ such that $v_1^{(i)} \in B_1(v) \cap \mathrm{v}[\Gamma_n]$ and $v_1'^{(i)} \in B_1(v') \cap \mathrm{v}[\Gamma_n]$.
The above $2\,\ell$ vertices $v_1^{(i)}, v_1'^{(i)}$, $1 \leq i \leq \ell$ differ by definition from v, v' in exactly one coordinate. Hence there is a mapping that assigns to each vertex the index of the coordinate in which it differs from v or v' respectively. Since ℓ is finite and for any $v \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ holds $\partial\{v\} \nearrow \infty$ we can assume w.l.o.g. that the $2\,\ell$ vertices yield exactly to $2\,\ell$ different coordinates. In particular we observe $d(v_1^{(i)}, v_1'^{(i)}) = k+2$.
For all pairs $v, w \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ holds

$$|\,\partial\{v\} \cap \partial\{w\}\,| \leq \alpha - 2 \quad (13)$$

and for each pair $(v_1^{(i)}, v_1'^{(i)})$ there are $(\alpha-1)\,(n-(k+2))$ vertices $w^{(i)} \in B_1(v_1^{(i)}) \cap B_{1+k}(v_1'^{(i)})$ with corresponding paths $\pi(w^{(i)})$ in $\mathcal{Q}_\alpha^n$ of length $\ell(\pi(w^{(i)})) = k+2$.
Writing $v_1^{(i)} = (x_1^{(i)}, ..., x_k^{(i)}, x_{k+1}^{(i)}, x_{k+2}^{(i)}, ..., x_n^{(i)})$ and $v_1'^{(i)} = (x_1'^{(i)}, ..., x_k'^{(i)}, x_{k+1}'^{(i)}, x_{k+2}'^{(i)}, x_{k+3}^{(i)}, ..., x_n^{(i)})$, to each pair $(v_1^{(i)}, v_1'^{(i)})$ there corresponds the path

$$(g_j(v_1^{(i)})_{0\leq j\leq k+2}), \quad g_j(v_1^{(i)}) \stackrel{\text{def}}{=\!=} (x_1^{(i)}, ..., x_j^{(i)}, x_{j+1}'^{(i)}, ..., x_n^{(i)})\,. \quad (14)$$

Clearly, for different pairs $(v_1^{(i)}, v_1'^{(i)})$, $(v_1^{(j)}, v_1'^{(j)})$ the above paths are pairwise disjoint. We further inspect that there exist at most $2\,\ell\,(k+2)\,[\alpha-1]$ vertices such that

$$\xi \in \bigcup_{w^{(i)} \in B_1(v_1^{(i)}) \cap B_{1+k}(v_1'^{(i)})} \left[\mathrm{Supp}(\pi(w^{(i)})) \cap \left(\bigcup_{j\neq i} \mathrm{Supp}(\pi(w^{(j)})) \right) \right]\,.$$

Therefore each pair $(v_1^{(i)}, v_1'^{(i)})$ leads to at least $[(\alpha-1)\,(n-(k+2))] - [2\,\ell\,(k+2)\,[\alpha-1]]$ paths $\pi(w^{(i)}) \in \Pi(\mathcal{Q}_\alpha^n)$, $w^{(i)} \in B_1(v_1^{(i)}) \cap B_{1+k}(v_1'^{(i)})$ that are completely disjoint to each path of the form $\pi(w^{(j)})$, $w^{(i)} \in B_1(v_1^{(i)}) \cap B_{1+k}(v_1'^{(i)})$, $i \neq j$ (see figure 4).
Thus the probability that none of the remaining pairwise disjoint paths of length $k+2$ is contained in $\Pi(\Gamma_n)$ is given by

$$(1-\lambda^{k+2})^{[(\alpha-1)\,(n-(k+2))]-[2\,\ell\,(k+2)\,[\alpha-1]]}\,.$$

We therefore obtain

$$\exists t \in \mathbb{R}_+ : \qquad \lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \hat{Y}_{n,k}^{v_1^{(i)},v_1'^{(i)}} = 0,\ 1 \le i \le \ell\,\} = \lim_{n\to\infty} e^{-t\,\ell\,n}\,.$$

We make use of equation (12) and choose ℓ sufficiently large such that

$$\lim_{n\to\infty} \left[\gamma^{\ell-1}\,\alpha^n\,e^{-t\,n\,\ell}\right] = 0 \quad b \in \mathbb{R}_+\,.$$

From the above equation we deduce that in the limit at least one pair $(v_1^{(i)}, v_1'^{(i)})$ fulfills

$$\hat{Y}_{n,k+2}^{v_1^{(i)},v_1'^{(i)}} > 0$$

and the proof of the lemma is complete. ∎

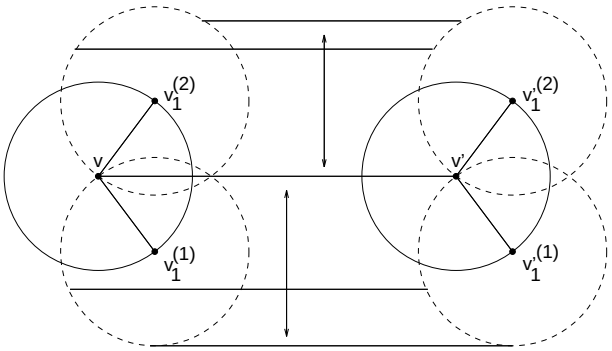


Figure 4: An illustration for the proof of lemma 5. For given $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ each pair of vertices $(v_1^{(i)}, v_1'^{(i)})$ leads to "sufficiently many" independent pairwise disjoint paths in $\Pi(\mathcal{Q}_\alpha^n)$.

Corollary 2. *Let k be a fixed positive integer and let $\hat{\Lambda}_{n,k}$ be the following random variable:*

$$\hat{\Lambda}_{n,k} \stackrel{\text{def}}{=\!=} \begin{cases} 1 & \text{if all pairs } v, v' \in \mathrm{v}[\Gamma_n] \text{ with } d(v,v') < k \text{ occur in a path of } \Gamma_n \\ 0 & \text{otherwise}\,. \end{cases}$$

Under the assumptions of lemma 5 holds $\lim_{n\to\infty} \boldsymbol{\mu}_n\{\,\Gamma_n \,|\, \hat{\Lambda}_{n,k} = 1\,\} = 1$.

Proof. According to equation (11) we observe

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n : \forall v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n] : \exists\, v_1^{(1)}, \ldots, v_1^{(\ell)} \in \partial\{v\} \cap \mathrm{v}[\Gamma_n] \wedge \exists\, v_1'^{(1)}, \ldots, v_1'^{(\ell)} \in \partial\{v'\} \cap \mathrm{v}[\Gamma_n]\,\} = 1,$$

and consequently in the limit a. s. for each pair $v, v \in \mathrm{v}[\Gamma_n]$ there exist $v_1 \in \partial\{v\} \cap \mathrm{v}[\Gamma_n], v_1' \in \partial\{v'\} \cap \mathrm{v}[\Gamma]$ such that $\hat{Y}_{n,k+2}^{v_1,v_1'} > 0$. Clearly, $\{v, v_1\}, \{v', v_1'\} \in \mathrm{e}[\Gamma_n]$ and hence by definition of $\hat{Y}_{n,k+2}^{v_1,v_1'}$ there is at least one path in $\Pi(\Gamma_n)$ in which v and v' occur whence the corollary. ∎

Now we are prepared to state the main result of this subsection:

Theorem 2. *Let $(\mathcal{Q}_\alpha^n)$ be a sequence of generalized hypercubes and $\Gamma_n < \mathcal{Q}_\alpha^n$ randomly induced subgraphs. Then*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is connected}\} = \begin{cases} 1 & \text{for} \quad \lambda > 1 - \sqrt[\alpha-1]{\alpha^{-1}} \\ 0 & \text{for} \quad \lambda < 1 - \sqrt[\alpha-1]{\alpha^{-1}}, \end{cases} \tag{15}$$

thus λ^* is a threshold value for connectedness.

Proof. The existence of components $\mathcal{X}$ with $|\,\mathcal{X}\,| \leq \gamma_n$ with has been investigated in lemma 2, where we proved that $\lambda^* = 1 - \sqrt[\alpha-1]{\alpha^{-1}}$ is a threshold value in Ω_n for the existence of nontrivial components whose orders are smaller than γ_n.

We choose $k \in \mathbb{N}$ then to each pair $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ the connectedness of $\mathcal{Q}_\alpha^n$ guarantees that there exists a path (in $\Pi(\mathcal{Q}_\alpha^n)$) say $(v_i)_{0\leq i\leq d(v,v')}$ in which v, v' occur. We can assume that $v = v_0, v' = v_{d(v,v')}$ and $d(v_i, v_{i+1}) < k$. We now consider the sets $B_2(v_i)$ for $0 \leq i \leq d(v, v')$ and introduce

$$\hat{Z}_{n,j}^{v,v'}(\Gamma_n) \stackrel{\text{def}}{=\!=} |\,\mathrm{v}[\Gamma_n] \cap B_2(v_j)\,|\,.$$

Claim: $\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \forall\, v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n],\ j \leq d(v, v') : |\,\hat{Z}_{n,j}^{v,v'} - \mathbf{E}[\hat{Z}_{n,j}^{v,v'}]\,| < \frac{1}{2}\,\mathbf{E}[\hat{Z}_{n,j}^{v,v'}]\,\} = 1\,.$
We only have to observe that $\hat{Z}_{n,j}^{v,v'}$ is binomially distributed and $\mathbf{E}[\hat{Z}_{n,j}^{v,v'}] = \binom{n}{2}(\alpha-1)^2\,\lambda$. We apply corollary 9 and the claim follows from

$$\lim_{n\to\infty} (\alpha^{2\,n}\, e^{-b\,n^2}) = 0 \quad b \in \mathbb{R}_+\,.$$

Therefore in all balls $B_2(v_i)$, $0 \leq i \leq d(v, v')$ we simultaneously find vertices of a random graph a.a.s.. The elements of $B_2(v_j) \cap \mathrm{v}[\Gamma_n]$ and $B_2(v_{j+k}) \cap \mathrm{v}[\Gamma_n]$ have in the limit pairwise finite distance (in $\mathcal{Q}_\alpha^n$) and are according to corollary 2 a.a.s. connected by a path in Γ_n.

Therefore a.a.s. in a random graph Γ_n for all $v, v' \in \mathrm{v}[\Gamma_n]$ there exists a $k_0 \in \mathbb{N}$ and a sequence of vertices $(v'_j)_{0\leq j\leq d(v,v')}$ such that $v'_j \in \mathrm{v}[\Gamma_n]$, $d(v'_i, v'_{i+1}) < k$ and $v_0 = v$, $v_{d(v,v')} = v'$. This sequence corresponds to a path in $\Pi(\Gamma_n)$ in which v and v' occur proving the theorem. ∎

Remark. A related result in the special case of the Boolean hypercube can be found in [3]. The corresponding subgraphs A_p are constructed as follows: We set $\mathrm{v}[A_p] \stackrel{\text{def}}{=\!=} \mathrm{v}[\mathcal{Q}_2^n]$ 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 [20, 3]. 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]. We shall discuss the sequence of components in the next subsection in the general case of configuration spaces.

3.3.3. Connectedness and Giant Components in General Configuration Spaces

We shall now generalize theorem 2 to general configuration spaces and investigate the emergence of giant components in the random graphs $\Gamma_n < \mathcal{C}_n$. For any $0 < \lambda < 1$ there exists a.a.s. a giant component in a random graph Γ_n. As in the previous subsections we assume that for the sequence of configuration spaces holds $0 < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$ and $\lambda^* = 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$.

Generalizing the strategy we have used for the generalized hypercubes we introduce the random variable

$$\hat{Y}^{v,v'}_{n,k}(\Gamma_n) \stackrel{\text{def}}{=\!=} \begin{cases} |\,\{\pi \in \mathbf{P}^{v,v',m}_{\mathcal{C}_n} \,|\, \pi \in \Pi(\Gamma_n)\,\}\,| & \text{for} \quad v, v' \in \mathrm{v}[\Gamma_n] \\ 0 & \text{otherwise} \end{cases}$$

Note that there are two facts implying that $\hat{Y}^{v,v'}_{n,k} \equiv 0$, namely (i) that $v \vee v' \notin \mathrm{v}[\Gamma_n]$ or (ii) that there exists no path $\pi \in \Pi(\Gamma_n)$ in which both vertices occur.
Property (R.iii) guarantees that for $v, v' \in \mathrm{v}[\Gamma_n]$ $\hat{Y}^{v,v'}_{n}$ is in the limit Gaussian (using the Moivre Laplace theorem) [2] since each path $\pi \in \mathbf{P}^{v,v',m}_{\mathcal{C}_n}$ is a path in a random graph Γ_n with the independent probability p such that

$$\lambda^{d(v,v')+m+2} \leq \mathrm{p} \leq \lambda^{d(v,v')+2}\,.$$

In complete analogy to lemma 5 we prove:

Lemma 6. *Let k be a natural number, $(\mathcal{C}_n)$ be a sequence of configuration spaces and $\Gamma_n < \mathcal{C}_n$ a random graph with underlying $\lambda > 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$. Then we have*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n | \forall\, v, v' \in \mathrm{v}[\mathcal{C}_n],\, d(v,v') = k : \exists v_1 \in \partial\{v\}, v'_1 \in \partial\{v'\} :\ \hat{Y}^{v_1,v'_1}_{n,d(v_1,v'_1)} > 0\,\} = 1\,.$$

Proof. The proof is completely analogous to that of lemma 5. We first observe that corollary 1 implies for arbitrary $\ell \in \mathbb{N}$

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n | \forall v, v' \in \mathrm{v}[\mathcal{C}_n] : \exists v_1^{(1)}, ..., v_1^{(\ell)} \in \partial\{v\} \cap \mathrm{v}[\Gamma_n] \wedge \exists\, v_1'^{(1)}, ..., v_1'^{(\ell)} \in \partial\{v'\} \cap \mathrm{v}[\Gamma_n]\,\} = 1\,. \quad (16)$$

Suppose $w, w' \in \mathrm{v}[\Gamma]$ then $\hat{Y}^{w,w'}_{n,k}$ is asymptotically Gaussian and according to corollary 9 there exists a $b > 0$ such that:

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\,\hat{Y}^{w,w'}_{n,k} = 0\,\} \leq e^{-b\,\gamma_n}\,. \quad (17)$$

Now let $\ell \in \mathbb{N}$ and P be the probability for the existence of a pair of vertices $v, v' \in \mathrm{v}[\mathcal{C}_n]$ with the following property:
all ℓ pairs of vertices $(v_1^{(i)}, v_1'^{(i)})$ where $v_1^{(i)} \in \partial\{v\} \cap \mathrm{v}[\Gamma_n], v_1'^{(i)} \in \partial\{v'\} \cap \mathrm{v}[\Gamma_n]$, $1 \leq i \leq \ell$ fulfill

$$\text{for } 1 \leq i \leq \ell : \qquad \hat{Y}^{v_1^{(i)}, v'^{(i)}_1}_{n,d(v_1^{(i)}, v'^{(i)}_1)} = 0\,.$$

Next we claim that there exists a $b \in \mathbb{R}_+$ such that $\lim_{n\to\infty} P \leq \lim_{n\to\infty} e^{-\ell\, b\, \gamma_n}$ a.a.s. .
Equation (17) implies that a.a.s. the number of vertices of $\mathrm{v}[\Gamma_n]$ that are adjacent to a pair (v, v'), $v, v' \in \mathrm{v}[\mathcal{C}_n]$ exceeds any finite constant. In other words there exist sets Φ_v and $\Phi_{v'}$ that fulfill condition (R.iv).
According to (R.i)-(R.iv) there exist ℓ pairs of vertices $(v_1^{(i)}, v_1'^{(i)})$ such that:
(a) for the random variables $\hat{Y}^{v_1^{(i)}, v'^{(i)}_1}_{n,d(v_1^{(i)}, v'^{(i)}_1)}$ we have $v_1^{(i)}, v'^{(i)}_1 \in \mathrm{v}[\Gamma_n]$.

(b) the pairs of vertices $(v_1^{(1)}, v'^{(1)}_1), ..., (v_1^{(\ell)}, v'^{(\ell)}_1)$ induce paths

$$\bigcup_{1\leq i\leq \ell} \mathbf{P}_{\mathcal{C}_n}^{v_1^{(i)}, v_1'^{(i)}, m} \subset \Pi(\mathcal{C}_n)$$

fulfilling (R.iii) and in particular $k-2 \leq \ell(\pi) \leq k+2+m$.

(c) for each pair $(v_1^{(i)}, v_1'^{(i)})$ we have

$$\lim_{n\to\infty} |\, \mathbf{P}_{\mathcal{C}_n}^{v_1^{(i)}, v_1'^{(i)}, m} \,| = \lim_{n\to\infty} \gamma_n \,.$$

(d) there are at most $2\,\ell\,\theta_n\,\gamma_n$ paths π' in $\bigcup_{1\leq j\neq i\leq \ell} \mathbf{P}_{\mathcal{C}_n}^{v_1^{(j)}, v_1'^{(j)}, m}$ such that for $\pi \in \mathbf{P}_{\mathcal{C}_n}^{v_1^{(i)}, v_1'^{(i)}, m}$ we have $\mathrm{Supp}(\pi) \cap \mathrm{Supp}(\pi') \neq \emptyset$.

Form the above properties it follows that for each pair $(v_1^{(i)}, v_1'^{(i)})$ there corresponds a.a.s. at least $\gamma_n - 2\,\ell\,\theta_n\,\gamma_n$ pairwise disjoint paths (R.iii), where $\lim_{n\to\infty}\theta_n = 0$. The probability that none of them is a path in Γ_n is bounded from above by

$$(1-\lambda^{k+2+m})^{[\gamma_n - 2\,\ell\,\theta_n\,\gamma_n]}\,,$$

since $d(v_1^{(i)}, v_1'^{(i)}) \leq k+2$. We conclude that there exists a $b \in \mathbb{R}_+$ such that

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \hat{Y}^{v_1^{(i)}, v'^{(i)}_1}_{n, d(v_1^{(i)}, v'^{(i)}_1)} = 0,\, 1\leq i\leq \ell\,\} = \lim_{n\to\infty} e^{-b\,\ell\,\gamma_n}\,.$$

The fact $0 < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$ and equation (16) guarantees that we can choose ℓ such that

$$\lim_{n\to\infty}\left[\gamma_n^{\ell-1}\,|\,\mathcal{C}_n\,|\,e^{-b\,\gamma_n\,(\ell+1)}\right] = 0\,.$$

Thus a.a.s. at least one pair $(v_1^{(i)}, v_1'^{(i)})$ fulfills

$$\hat{Y}^{v_1^{(i)}, v'^{(i)}_1}_{n, d(v_1^{(i)}, v'^{(i)}_1)} > 0$$

and the proof of the lemma is complete. ∎

Corollary 3. *Let k be a positive integer and $\hat{\Lambda}_{n,k}$ the random variable:*

$$\hat{\Lambda}_{n,k} \stackrel{\text{def}}{=\!=} \begin{cases} 1 & \text{if all pairs } v, v' \in \mathrm{v}[\Gamma_n] \text{ with } d(v,v') < k \text{ occur in a path of } \Gamma_n \\ 0 & \text{otherwise}\,. \end{cases}$$

Then under the assumptions of lemma 6 we have $\lim_{n\to\infty}\boldsymbol{\mu}_n\{\,\Gamma_n \,|\, \hat{\Lambda}_{n,k} = 1\,\} = 1$.

Proof. According to equation (16) we observe

$$\lim_{n\to\infty}\boldsymbol{\mu}_n\{\Gamma_n : \forall v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n] : \exists v_1^{(1)}, ..., v_1^{(\ell)} \in \partial\{v\}\cap \mathrm{v}[\Gamma_n] \,\wedge\, \exists\, v_1'^{(1)}, ..., v_1'^{(\ell)} \in \partial\{v'\}\cap \mathrm{v}[\Gamma_n]\} = 1$$

and consequently a.a.s. for each pair $v, v \in \mathrm{v}[\Gamma_n]$ there exist $v_1 \in \partial\{v\}\cap \mathrm{v}[\Gamma_n]$, $v_1' \in \partial\{v'\}\cap \mathrm{v}[\Gamma_n]$ with the property

$$\hat{Y}^{v_1, v_1'}_{n, d(v_1, v_1')} > 0\,.$$

Clearly, $\{v, v_1\}, \{v', v_1'\} \in \mathrm{e}[\Gamma_n]$ and there exists by definition of $\hat{Y}^{v_1, v_1'}_{n, d(v_1, v_1')}$ at least one path in $\Pi(\Gamma_n)$ in which v and v' occur. ∎

The following corollary shall be used in the proof of theorem 4 that is concerned with the existence of a giant component in randomly induced subgraphs Γ_n.

Corollary 4. *Let $(\mathcal{C}_n)$ be a sequence of configuration spaces and $0 < \lambda \leq 1$. For $k \in \mathbb{N}$ we set*

$$U_n(\Gamma_n) \stackrel{\text{def}}{=\!=} \{v \in \mathrm{v}[\Gamma_n] \,|\, \delta_v \geq \ln(\gamma_n)\}\,. \tag{18}$$

Then we have

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \Gamma_n[U_n(\Gamma_n)] \textit{ is a component of } \Gamma_n\} = 1\,.$$

Theorem 3. *Suppose $(\mathcal{C}_n)$ is a sequence of configuration spaces such that $0 < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$ and $\lambda > 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$. Then*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is connected}\} = \begin{cases} 1 & \text{for} \quad \lambda > 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} \\ 0 & \text{for} \quad \lambda < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}\,. \end{cases} \tag{19}$$

That is λ^* is a threshold value for the connectedness property.

Proof. We essentially reorganize the proof of theorem 2. The case of small components follows from lemma 2 whence it remains to prove connectedness under the assumption $\lambda > \lambda^*$.
According to corollary 3 a.a.s. the random graphs have the property that each pair of vertices (v, v'), $v, v' \in \mathrm{v}[\Gamma_n]$ with $\lim_{n\to\infty} d(v, v') < k$ where $k \in \mathbb{N}$ occurs in a path in $\Pi(\Gamma_n)$.
For arbitrary $v, v' \in \mathrm{v}[\mathcal{C}_n]$ the connectedness of $\mathcal{C}_n$ guarantees that there exists a path $\eta_{v,v'} \in \Pi(\mathcal{C}_n)$ in which v and v' occur. Let $\eta_{v,v'}$ be given by $(v_j)_{0\leq j\leq d(v,v')}$ we consider the sets $B_2(v_j) \subset \mathrm{v}[\mathcal{C}_n]$ for $1 \leq j \leq d(v, v')$ and show in complete analogy to the proof of theorem 2 that

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \forall\, 0 \leq j \leq d(v, v'),\, B_2(v_j) \cap \mathrm{v}[\Gamma_n] \neq \emptyset\} = 1\,. \tag{20}$$

Clearly, the probability that for one index $0 \leq j \leq d(v, v')$ holds $B_2(v_j) \cap \mathrm{v}[\Gamma_n] = \emptyset$ is given by $(1-\lambda)^{|\,B_2(v_j)\,|}$ and according to property (Q) of definition 3 we have $\lim_{n\to\infty} |\,B_2(v_j)\,| = c\,\gamma_n^2$, $c \in \mathbb{R}_+$. Since $d(v, v') \leq |\,\mathcal{C}_n\,|$ and $0 < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$ we obtain

$$\lim_{n\to\infty} |\,\mathcal{C}_n\,|^3\,(1-\lambda)^{c\,\gamma_n^2} = 0\,,$$

proving that equation (20) holds.
For $k \in \mathbb{N}$ all elements of $B_2(v_j)$ and $B_2(v_{j+k})$ have pairwise finite distance in $\mathcal{C}_n$ and according to corollary 3 to given $k \in \mathbb{N}$ there exists a.a.s. a sequence of vertices (v_i) where $v_i \in \mathrm{v}[\Gamma_n]$, $d(v_i, v_{i+1}) < k_0$, $v_0 = v$ and $v_{d(v,v')} = v'$. ∎

Theorem 4. *Let $(\mathcal{C}_n)$ be a sequence of configuration spaces such that $0 < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$ and $0 < \lambda \leq 1$. Then we have*

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

Proof. It remains to consider the case $\lambda < \lambda^*$ since for $\lambda > \lambda^*$ theorem 3 immediately implies

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is connected}\} = 1\,.$$

Let $\lambda < \lambda^*$ and $k = k(n)$. We consider $V_k \stackrel{\text{def}}{=\!=} \{v \in \mathrm{v}[\Gamma_n] \,|\, \delta_v \leq k\,\}$ and inspect for $k(n) \leq \ln(\gamma_n)$

$$|\,V_k\,| \leq [(1-\lambda)]^{\gamma_n}\,[1+\frac{\lambda}{1-\lambda}]^k\,\gamma_n^k\,k\,|\,\mathcal{C}_n\,|.$$

We choose $k = \ln(\gamma_n)$ and obtain setting $\xi_n \stackrel{\text{def}}{=\!=} (1-\lambda)^{\gamma_n}\,[1+\frac{\lambda}{1-\lambda}]^{\ln(\gamma_n)}\,\gamma_n^{\ln(\gamma_n)}\,\ln(\gamma_n)$

$$\forall \ell \in \mathbb{N}: \quad \lim_{n\to\infty} \gamma_n^\ell\,\xi_n = 0\,. \tag{21}$$

We shall show that $\Gamma_n[U_n(\Gamma_n)]$ (see equation (18)) forms a giant component in Γ_n. For this purpose we first consider the case $v, v' \in U_n(\Gamma_n)$ and $\lim_{n\to\infty} d(v,v') < k$, $k \in \mathbb{N}$.
In this situation we can apply corollary 4 and obtain that a.a.s. in the random graphs Γ_n all pairs of vertices $v, v' \in U_n(\Gamma_n)$ that have finite distance occur in a path in $\Pi(\Gamma_n)$.
For arbitrary $v, v' \in U_n(\Gamma_n)$ the connectedness of $\mathcal{C}_n$ implies that there exists a path $(v_j)_{0\leq j\leq d(v,v')}$ in $\mathcal{C}_n$ in which v and v' occur. It remains to show that

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \forall\, 0 \leq j \leq d(v,v'),\; B_2(v_j) \cap U_n(\Gamma_n) \neq \emptyset\,\} = 1\,.$$

Then the theorem follows by application of corollary 4 since the vertices of $B_2(v_j)$ and $B_2(v_{j+k})$ have pairwise finite distance. The probability that for one index $0 \leq j \leq d(v,v')$ holds $B_2(v_j) \cap U_n(\Gamma_n) = \emptyset$ is given by $(1-\lambda+\xi_n)^{|\,B_2(v_j)\,|}$ and taking the limit we compute

$$\lim_{n\to\infty} (1-\lambda+\xi_n)^{|\,B_2(v_j)\,|} = \lim_{n\to\infty} (1-\lambda)^{|\,B_2(v_j)\,|}\; e^{\frac{|\,B_2(v_j)\,|\,\xi_n}{1-\lambda}}\,.$$

According to equation (21) and property (Q) of definition 3 we have $\lim_{n\to\infty} |\,B_2(v_j)\,|\,\xi_n = 0$ and finally end up with

$$\lim_{n\to\infty} |\,\mathcal{C}_n\,|^3\,(1-\lambda+\xi_n)^{|\,B_2(v_j)\,|} = 0\,.$$

Therefore the subgraph induced by the vertex set

$$U_n = \{v \in \mathrm{v}[\Gamma_n] \,|\, \delta_v \geq \ln(\gamma_n)\,\}$$

is a giant component fulfilling $\mathcal{C}_n[U] < \mathcal{X}_1$ whence the theorem. ∎

4. Neutral Networks of RNA Secondary Structures

4.1. RNA Secondary Structures and Compatible Sequences

The mapping defined by the folding of RNA molecules into their (spatial) shapes has received special attention during the last few years. While a prediction of true 3D structures is far beyond the capacity of present-day computers, secondary structures are readily accessible. A large body of computational data has been published [11, 15, 12, 35, 4, 41] on this example of a sequence-structure mapping.

The shape space consists of all secondary structure graphs as defined below. A variety of different algorithms [29, 46, 45, 27, 26], and different sets of thermodynamic parameters [32, 16, 43] have been used for the prediction of RNA secondary structures. Fortunately, it has been shown recently [42] that the qualitative features of the sequence-structure mappings are independent of algorithm and parameter set.

Definition 6. [44] *A secondary structure is a vertex-labeled graph on n vertices with an adjacency matrix A fulfilling*
(1) $a_{i,i+1} = 1$ *for* $1 \leq i \leq n-1$*;*
(2) For each i there is at most a single $k \neq i-1, i+1$ such that $a_{ik} = 1$;
(3) 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 and write $[i,k] \in s$. A vertex i connected only to $i-1$ and $i+1$ shall be called unpaired. We shall denote the number of base pairs and the number of unpaired bases in a secondary structure s by $n_p(s)$ and $n_u(s)$, respectively. The stickiness of the pair-alphabet is $\mathrm{p} \stackrel{\text{def}}{=\!=} \beta/\alpha^2$*, which is the probability that two arbitrarily chosen letters are capable of forming a base pair. (We denote the size of the alphabet by α and the number of distinct base pairs by β).*

Note that $n_u(s) + 2n_p(s) = n$, i..e., the chain length of the molecule. Let $\mathcal{A}$ be an arbitrary alphabet. A *pairing rule* Π on $\mathcal{A}$ is a set of pairs $[x,y] \in \mathcal{A} \times \mathcal{A}$, such that $[x,y] \in \Pi$ implies $[y,x] \in \Pi$. In the following we shall consider secondary structures over arbitrary alphabets with arbitrary pairing rules.

Definition 7. *Let s be a RNA secondary structure and*

$$\Pi(s) \stackrel{\text{def}}{=\!=} \{[i,k] \,|\, a_{i,k} = 1,\, k \neq i-1, i+1\}$$

its set of contacts. A sequence x is said to be compatible to s if and only if for all $[i,j] \in \Pi(s)$ holds $[x_i, x_j] \in \Pi$, i.e., the nucleotides x_i and x_j can in fact form a base pair for each pair $[i,j] \in \Pi(s)$. We call the set $\mathbf{C}[s]$ the set of compatible sequences.

Remark. For the size of a compatible set we obtain $|\,\mathbf{C}[s]\,| = \alpha^{n_u}\,\beta^{n_p} = \alpha^n \mathrm{p}^{n_p}$.

In order to investigate the structure of compatible sets the following algebraic framework shall be useful:

Definition 8. *Let S_n be the symmetric group in n letters. We write a transposition $\tau \in S_n$ as $\tau = (i,k)$. Then*

$$\begin{array}{rccl} \imath: & \mathcal{S}_n & \to & S_n \\ & s & \mapsto & \imath(s) \quad \stackrel{\text{def}}{=\!=} \prod_{[i,k]\in\Pi(s)}(i,k)\,. \end{array}$$

The map $\imath$ is an *embedding* and we have $\imath(s)^2 = 1$, i.e., the images are *involutions*. A *dihedral group*, D_m, is a group generated by two involutions [36]. The map $\imath$ naturally gives rise to the mapping

$$\begin{array}{rccl} \jmath: & \mathcal{S}_n \times \mathcal{S}_n & \longrightarrow & \{D_m < S_n\} \\ & (s,s') & \mapsto & \jmath(s,s') \stackrel{\text{def}}{=\!=} \langle \imath(s), \imath(s')\rangle\,. \end{array}$$

The structure of $\langle \imath(s), \imath(s')\rangle$ is easily seen to be a *semi-direct product* of the form

$$\langle \imath(s), \imath(s')\rangle \cong \langle \imath(s)\rangle \ltimes \langle \imath(s) \circ \imath(s')\rangle\,.$$

Theorem 5. *[Intersection–Theorem] Let s and s' be arbitrary secondary structures. Then we have*

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

Proof. If the alphabet allows a symmetric base pair $[XX]$ there is nothing to prove: poly-X is compatible with all structures. Suppose therefore that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both, s and s'. Then $\jmath(s,s') \cong D_m$ operates on the set of all positions $\{x_1,..,x_n\}$. Since we have the operation of a dihedral group the orbits are either cycles or chains and the cycles have even order. A constraint for the sequence compatible to both structures appears only in the cycles where the choice of bases is not independent. It remains to be shown that there is a valid choice of bases for each cycle, which is obvious since these have even order. Therefore it suffices to choose an alternating sequence of the pairing partners X and Y. Thus there are at least 2 different choices for the first base in the orbit. ∎

Remark. A generalization of the statement of theorem 5 for 3 different structures is false.

Corollary 5. *Suppose the alphabet $\mathcal{A}$ of length α admits at least one type of complementary base pair. Then $|\,\mathbf{C}[s_1] \cap \mathbf{C}[s_2]\,| \geq \alpha^{|\,\Phi\,|}$ where Φ is the set of orbits induced by the operation of $\langle \imath(s), \imath(s')\rangle$.*

Consider a combinatory map $f_n : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$. We know *a priori* that the vertex set of the preimage $f_n^{-1}(s)$ which consists of all sequences folding into the secondary structure s is contained in the set of *compatible sequences*. In particular, all neutral neighbors of a sequence x are located in the set $\mathbf{C}[f_n(x)]$. Unfortunately, the induced subgraph $\mathcal{Q}_\alpha^n[\mathbf{C}[f_n(x)]]$ is not connected — it decomposes into "hyper-planes" defined by a particular choice of the base pairs.[3] Therefore we introduce the graph $\mathcal{C}[s]$:

Definition 9. *Let s be a secondary structure, then the graph of compatible sequences is*

$$\mathcal{C}[s] \stackrel{\text{def}}{=\!=} \mathcal{Q}_\alpha^{n_u(s)} \times \mathcal{Q}_\beta^{n_p(s)}\,.$$

[3] Even with $[GU]$ pairs the corresponding graphs are still not connected: there is no path of (subsequent point mutations) that would, for instance, convert a $[GC]$ pair into a $[CG]$ pair.)

Remark. Obviously $\mathcal{C}[s]$ has the vertex set $\mathbf{C}[s]$. By definition of the *product of graphs*, two sequences $x, y \in \mathbf{C}[s]$ are neighbors if they differ either

- in a single position i which is unpaired in s, or
- in two positions i and j which form a base pair $[i,j] \in s$.

Remark. Note that two graphs $\mathcal{C}[s], \mathcal{C}[s']$ are *isomorphic as graphs* iff both have the same *number* of unpaired and paired bases. Accordingly, two different secondary structures $s, s' \in \mathcal{S}_n$ can lead to isomorphic graphs of compatible sequences i.e. $\mathcal{C}[s] \cong \mathcal{C}[s']$.

4.2. Neutral Networks as Random Induced Subgraphs

Definition 10. *Suppose $f_n : \mathcal{Q}_\kappa^n \to \mathcal{S}_n$ is a mapping and $s \in \mathcal{S}_n$ a fixed RNA secondary structure. Then the neutral network with respect to s, $\Gamma_n[s]$, is the induced subgraph of $f_n^{-1}(s)$ in $\mathcal{C}[s]$, i.e.,*

$$\Gamma_n[s] \stackrel{\text{def}}{=\!=} \mathcal{C}[s]\left[f_n^{-1}(s)\right] .$$

We shall now construct neutral networks as random graphs by means of a simple random process. More precisely we consider randomly induced subgraphs of the graph product $\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ that are induced by certain subsets of vertices. *A priori* there is no reason why the frequency of neutral neighbor should be the same for single base mutations and base pair mutations, respectively. Numerical data for RNA free energy structures show that in fact the differences are significant. This motivates the following modification of our basic random graph model:

Model A: *Let s be a secondary structure with corresponding graph of compatible sequences $\mathcal{C}[s] = \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$. We consider the set of all subgraphs $G < \mathcal{C}[s]$ that can be written as $G = \mathcal{C}[s][V]$ where $V \subset \mathrm{v}[\mathcal{C}[s]]$. In other words the graphs G are induced subgraphs of vertex sets. Writing $\chi_{u,p} \stackrel{\text{def}}{=\!=} \lambda_u + \lambda_p - \lambda_u\,\lambda_p$ we set*

$$\boldsymbol{\mu}_{n,\lambda_u,\lambda_p}(G) \stackrel{\text{def}}{=\!=} \chi_{u,p}^{|\,\mathrm{v}[G]\,|}\,(1-\chi_{u,p})^{\alpha^{n_u}\beta^{n_p}-|\,\mathrm{v}[G]\,|} .$$

$\sum_G \boldsymbol{\mu}_{n,\lambda_u,\lambda_p}(G) = 1$, thus $\boldsymbol{\mu}_{n,\lambda_u,\lambda_p}$ is a probability measure. Then we define a neutral network as an induced random subgraph of $\mathcal{C}[s]$ where $\boldsymbol{\mu}_{n,\lambda_u,\lambda_p}$ is the underlying measure i.e.:

$$\Gamma_n^{\mathrm{A}}[s] < \mathcal{C}[s] .$$

Remark. We can intuitively construct a neutral network $\Gamma_n^{\mathrm{A}}[s]$ as follows:
writing each $v \in \mathrm{v}[\mathcal{C}[s]]$ as $v = (v_u, v_p)$ we select each $v \in \mathrm{v}[\mathcal{C}[s]]$ with the independent probability $\lambda_u + \lambda_p - \lambda_u\,\lambda_p$. Or equivalently–we select each coordinate v_u with the probability λ_u, v_p with corresponding probability λ_p and finally select a $v = (v_u, v_p) \in \mathrm{v}[\mathcal{C}[s]]$ if either v_u or v_p have been chosen.

Before proceed with the analysis of model A we introduce some more terminology.

Definition 11. *Let G_1, G_2 be graphs, Γ a subgraph of $G_1 \times G_2$ and $(x,y) \in \mathrm{v}[\Gamma]$. The fibers of Γ $\Phi_x^{\Gamma,\Phi_y^\Gamma}$ in $G_1 \times G_2$ are the following induced subgraphs in G_2 and G_1:*

$$\Phi_x^\Gamma \stackrel{\mathrm{def}}{=\!=} G_1 \times G_2[\{y \in \mathrm{v}[G_2] \,|\, (x,y) \in \mathrm{v}[\Gamma]\}] \quad \textit{and}$$
$$\Phi_y^\Gamma \stackrel{\mathrm{def}}{=\!=} G_1 \times G_2[\{x \in \mathrm{v}[G_1] \,|\, (x,y) \in \mathrm{v}[\Gamma]\}] .$$

The theory developed in section 3 can now be used to derive a sufficient criterion for the density and the connectedness of random subgraphs $\Gamma_n^{\mathrm{A}}[s] < \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$.

The connection to the theory in section 3 is established as follows. For $(x,y) \in \mathrm{v}[\Gamma_n^{\mathrm{A}}[s]]$:

$$\Phi_x^{\Gamma_n^{\mathrm{A}}[s]} \cong \Gamma_{n_p} \qquad \text{and} \qquad \Phi_y^{\Gamma_n^{\mathrm{A}}[s]} \cong \Gamma_{n_u}$$

with underlying probability measures

$$\boldsymbol{\mu}_{n_u,\lambda_u}(\Gamma_{n_u}) = \lambda_{n_u}^{|\,\mathrm{v}[\Gamma_{n_u}]\,|}\,(1-\lambda_u)^{\alpha^{n_u}-|\,\mathrm{v}[\Gamma]_{n_u}\,|} \quad \text{and}$$
$$\boldsymbol{\mu}_{n_p,\lambda_p}(\Gamma_{n_p}) = \lambda_{n_p}^{|\,\mathrm{v}[\Gamma_{n_p}]\,|}\,(1-\lambda_p)^{\beta^{n_p}-|\,\mathrm{v}[\Gamma]_{n_p}\,|} .$$

Theorem 6. *Let $\Gamma_n^{\mathrm{A}}[s] < \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ be a random graph constructed according to model A. Suppose that*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\,\Gamma_n^{\mathrm{A}}[s] \,|\, \forall (v_1,v_2) \in \mathrm{v}[\Gamma_n^{\mathrm{A}}[s]] :\ \Phi_{v_1}^{\Gamma_n^{\mathrm{A}}[s]},\ \Phi_{v_2}^{\Gamma_n^{\mathrm{A}}[s]} \text{are dense and connected}\,\} = 1 .$$

Then $\Gamma_n^{\mathrm{A}}[s]$ is a.a.s. dense and connected in $\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$.

Proof. We write for short $\Gamma \stackrel{\mathrm{def}}{=\!=} \Gamma_n^{\mathrm{A}}[s]$. The statement concerning the density-property is obvious. Let k be a natural number we first show that for each pair $(v_1,v_2),(v_1',v_2') \in \mathrm{v}[\Gamma]$ with distance[4] $d((v_1,v_2),(v_1',v_2')) < k$, there exists a path in Γ in which both vertices occur.
For this purpose we consider $\Phi_{v_1}^\Gamma, \Phi_{v_1'}^\Gamma$ which are by assumption a.a.s. connected for arbitrary v_1, v_1'. The probability for selecting a pair $(v_1,x),(v_1',x)$ is λ_p for each $x \in \mathrm{v}[\mathcal{Q}_\beta^{n_p}]$ by definition. Let $\hat{X}_{(v_1,v_2),(v_1',v_2')}$ be the random variable

$$\hat{X}_{(v_1,v_2),(v_1',v_2')}(\Gamma) \stackrel{\mathrm{def}}{=\!=} \begin{cases} |\,\{x \in \mathrm{v}[\mathcal{Q}_\beta^{n_p}] \,|\, (v_1,x),(v_1',x) \in \mathrm{v}[\Gamma]\,\}\,| & \text{for } (v_1,v_2),(v_1',v_2') \in \mathrm{v}[\Gamma] \\ 0 & \text{otherwise.} \end{cases}$$

Then $\hat{X}_{(v_1,v_2),(v_1',v_2')}$ is binomially distributed and $\mathbf{E}[\hat{X}_{(v_1,v_2),(v_1',v_2')}] = c\,e^{a\,n}$ with $a \in \mathbb{R}_+$. Applying corollary 9 of the appendix we observe that

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma | \forall (v_1,v_2),(v_1',v_2') \in \mathrm{v}[\mathcal{C}[s]] \,|\hat{X}_{(v_1,v_2),(v_1',v_2')} - \mathbf{E}[\hat{X}_{(v_1,v_2),(v_1',v_2')}]\,| < \frac{1}{2}\mathbf{E}[\hat{X}_{(v_1,v_2),(v_1',v_2')}]\ \} = 1$$

The pair $(v_1,x),(v_1',x) \in \mathrm{v}[\Gamma]$ is a.a.s. connected by a path of finite length of the form

$$(v_1,x),(w_1,x),\ldots,(w_1',x),(v_1',x)$$

since all fibers $\Phi_x^\Gamma \cong \Gamma_{n_u}$ are by assumption a.a.s. connected. Along these lines we inspect that a.a.s. there exists a path of finite length of the form

$$(v_1',x),(v_1',y),\ldots,(v_1',z),(v_1',v_2')$$

since $\Phi_{v_1'}^\Gamma \cong \Gamma_{n_p}$ is by assumption a.a.s. connected.
In the general case we repeat the argument used in the proofs of the theorems 3 and 4 respectively. We choose a path (ξ_i) in the connected graph $\mathcal{C}[s]$ such that $d(\xi_i,\xi_{i+1}) < k$ and in which $(v_1,v_2),(v_1',v_2')$ occur. Then we show that a.a.s. in each ball $B_2(\xi_i)$ there is a vertex of Γ. Finally we apply the first part implying that there exists a.a.s. a path $\pi \in \Pi(\Gamma)$ in which both vertices occur. ∎

[4]The distance in a direct product of two graphs G_1 and G_2 is given by $d((u_1,u_2),(v_1,v_2)) = d_1(u_1,v_1) + d_2(u_2,v_2)$, where d_1 and d_2 are the distances on G_1 and G_2, respectively.

Remark. Note that density and connectedness of $\Gamma_n^{\mathrm{A}}[s]$ does not allow the general conclusion that the corresponding fibers are dense and connected.

Corollary 6. *Let $\Gamma_n^{\mathrm{A}}[s] < \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ be a random subgraph obtained from model A such that $\lambda_u > 1 - \sqrt[\alpha-1]{\alpha^{-1}}$ and $\lambda_p > 1 - \sqrt[\beta-1]{\beta^{-1}}$. Then*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n^{\mathrm{A}}[s] \text{ is dense and connected } \} = 1\,.$$

Proof. According to theorem 6 it remains to verify the condition

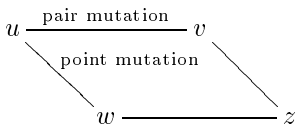
$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\,\Gamma_n^{\mathrm{A}}[s] \,|\, \forall (v_1,v_2) \in \mathrm{v}[\Gamma_n^{\mathrm{A}}[s]] : \Phi_{v_1}^{\Gamma_n^{\mathrm{A}}[s]}, \Phi_{v_2}^{\Gamma_n^{\mathrm{A}}[s]} \text{ are dense and connected}\,\} = 1\,.$$

From the proof of lemma 5 we inspect immediately that the probability for the existence of a disconnected random subgraph for $\lambda > \lambda^*$ has the upper bound $e^{-b\,\ell\,n}$ with arbitrary $\ell \in \mathbb{N}$ and $b \in \mathbb{R}_+$. Since we can only have α^n different fibers of the form $\Phi_{v_1}^{\Gamma_n^{\mathrm{A}}[s]}$ and accordingly β^n fibers $\Phi_{v_2}^{\Gamma_n^{\mathrm{A}}[s]}$, we can choose ℓ sufficiently large such that

$$\lim_{n_u\to\infty} \alpha^{n_u}\, e^{-b\,\ell\,n_u} = 0 \quad \text{and} \quad \lim_{n_p\to\infty} \beta^{n_p}\, e^{-b\,\ell\,n_p} = 0\,.$$

Consequently the above condition is implied by $\lambda_u > 1 - \sqrt[\alpha-1]{\alpha^{-1}}$, $\lambda_p > 1 - \sqrt[\beta-1]{\beta^{-1}}$, and the corollary follows. ∎

Before we proceed with the analysis of model A we consider the following situation: Suppose for a (combinatory) map $f_n : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$ holds $f(u) = f(v) = f(w)$ where v differs from u by a point mutation, while w differs from u by a pair mutation. Then there is a unique sequence z which differs from u by both the point mutation and the pair mutation:



In model A we have assumed that there is asymptotically no correlation between point mutations and pair mutations. Each fiber $\Phi_{v_2}^{\Gamma_n^{\mathrm{A}}[s]}$ is isomorphic to a random graph Γ_{n_u} and accordingly $\Phi_{v_1}^{\Gamma_n^{\mathrm{A}}[s]} \cong \Gamma_{n_p}$. The other extreme is to consider these two types of mutations as completely correlated in the following sense:
If any three vertices in the diagram above are chosen, the fourth vertex has to be chosen as well.
In other words we have $\forall (v_1,v_2) \in \mathrm{v}[\Gamma_n^{\mathrm{A}}[s]]$ holds $\Phi_{v_2}^{\Gamma_n^{\mathrm{A}}[s]} \cong \Gamma_{n_u}$ and $\Phi_{v_1}^{\Gamma_n^{\mathrm{A}}[s]} \cong \Gamma_{n_p}$.

Model B *Let $\Gamma_{n_u} < \mathcal{Q}_\alpha^{n_u}$ and $\Gamma_{n_p} < \mathcal{Q}_\beta^{n_p}$ be random subgraphs as introduced in our basic model. We set $\Gamma_n^{\mathrm{B}}[s] = \Gamma_{n_u} \times \Gamma_{n_p}$ and*

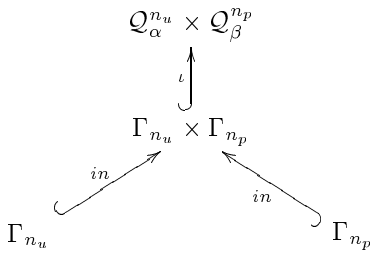
$$\boldsymbol{\mu}_{\lambda_u,\lambda_p}(\Gamma_n^{\mathrm{B}}[s]) \stackrel{\mathrm{def}}{=\!=} \boldsymbol{\mu}_{n_u,\lambda_u}(\Gamma_{n_u}) \times \boldsymbol{\mu}_{n_p,\lambda_p}(\Gamma_{n_p})\,.$$

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

Remark. We can construct the above randomly induced subgraphs of $\mathcal{C}[s]$, $\Gamma_n^{\mathrm{B}}[s]$, 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 $V_{\lambda_u,\lambda_p} \subset \mathbf{C}[s]$. Then $\Gamma_n^{\mathrm{B}}[s]$ is the induced subgraph $\mathcal{C}[s][V_{\lambda_u,\lambda_p}]$ i.e.,

$$\Gamma_n^{\mathrm{B}}[s] = \Gamma_{n_u} \times \Gamma_{n_p} .$$

We have $\Phi_{v_1}^{\Gamma_n^{\mathrm{B}}[s]} \cong \Gamma_{n_p}$ and $\Phi_{v_2}^{\Gamma_n^{\mathrm{B}}[s]} \cong \Gamma_{n_u}$ where we assume $\boldsymbol{\mu}_{n_u,\lambda_u}, \boldsymbol{\mu}_{n_p,\lambda_p}$ to be the underlying probability measures. The situation can be reported by the following diagram:



Theorem 6 has the following analogue for model B:

Theorem 7. *Suppose*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_{n_u} \textit{ is dense and connected } \} = 1 \quad \textit{and} \quad \lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_{n_p} \textit{ is dense and connected } \} = 1 .$$

Then we have

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n^{\mathrm{B}}[s] \textit{ is dense and connected } \} = 1 .$$

Proof. The proof is completely analogous to the proof of theorem 6. ∎

Since $\Phi_x^{\Gamma_n^{\mathrm{B}}[s]} \cong \Gamma_{n_p}$, $\Phi_y^{\Gamma_n^{\mathrm{B}}[s]} \cong \Gamma_{n_u}$ we derive the following criterion for density and connectedness of neutral networks that are randomly induced subgraphs $\Gamma_n^{\mathrm{B}}[s]$.

Corollary 7. *Suppose* $\lambda_u > 1 - \sqrt[\alpha-1]{\alpha^{-1}}$ *and* $\lambda_p > 1 - \sqrt[\beta-1]{\beta^{-1}}$, *then we have*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n^{\mathrm{B}}[s] \textit{ is dense and connected } \} = 1 .$$

Lemma 7. *For the orders of the random graphs* $\Gamma_n^{\mathrm{A}}[s], \Gamma_n^{\mathrm{B}}[s]$ *we have for* $N, N_u, N_p \in \mathbb{N}$*:*

$$\begin{aligned} \boldsymbol{\mu}_n\{\Gamma_n^{\mathrm{A}}[s] = N\} &= B(N, \alpha^{n_u}\beta^{n_p}, [\lambda_u + \lambda_p - \lambda_u\lambda_p]) \\ \boldsymbol{\mu}_n\{\Gamma_n^{\mathrm{B}}[s] = N\} &= \sum_{N_u\ N_p = N} B(N_u, \alpha^{n_u}, \lambda_u)\, B(N_p, \beta^{n_p}, \lambda_p) . \end{aligned}$$

In particular the distributions of the orders become asymptotically Gaussian.

4.3. Shape Space Covering

The combination of a variety of computer simulations [15, 12, 35] provides strong evidence for the existence of a relatively "small" set of the form $B_r(v)$ in a generalized hypercube $\mathcal{Q}_\alpha^n$ (where v is an arbitrary sequence $v \in \mathrm{v}[\mathcal{Q}]_\alpha^n$) that has the following property:
$B_r(v)$ contains sequences whose corresponding secondary structures cover almost all "common" secondary structures.
This statement has been termed *shape space covering conjecture.*

Let $\Gamma_n[s_1], \Gamma_n[s_2]$ be two neutral networks. Then the *minimal Hamming distance* between the $\Gamma_n[s_1]$ and $\Gamma_n[s_2]$ is

$$\mathrm{dist}(\Gamma_n[s_1], \Gamma_n[s_2]) \stackrel{\mathrm{def}}{=\!=} \min\{d_h(v_1, v_2) \,|\, v_1 \in \mathrm{v}[\Gamma_n[s_1]],\, v_2 \in \mathrm{v}[\Gamma_n[s_2]]\}\,.$$

The theory presented above provides a proof (within the limits of the models) for this conjecture in the following form:

Theorem 8. *[Shape Space Covering] Let s_1 and s_2 be two secondary structures and $X = \mathrm{A,B}$. Suppose that for $i = 1, 2$*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n^{\mathrm{X}}[s_i] \textit{ is dense and connected }\} = 1\,.$$

Then
(i) *The minimum distance between the neutral networks $\Gamma_n^{\mathrm{X}}[s_1]$ and $\Gamma_n^{\mathrm{X}}[s_2]$ is a.a.s. at most*

$$\mathrm{dist}(\Gamma_n^{\mathrm{X}}[s_1], \Gamma_n^{\mathrm{X}}[s_2]) \leq 2 \tag{22}$$

(ii) *The expected minimal radius $r_{s_1} \stackrel{\mathrm{def}}{=\!=} \mathbf{E}[\min_{r'}\{r' \,|\, B_{r'}(v) \cap \mathrm{v}[\Gamma_n^{\mathrm{X}}[s_1]] \neq \emptyset\}]$, i.e. the expected Hamming distance from a randomly chosen sequence $v \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ to a neutral network $\Gamma_n^{\mathrm{X}}[s_i]$, is given by*

$$r_{s_i} = [1 - \frac{\beta}{\alpha^2}]\, n_p(s_i) + \theta_n, \quad \text{where } \lim_{n\to\infty} \theta_n = 0\,.$$

Proof. (i) is an obvious consequence of theorem 5 and theorem 6.
(ii) We choose $i = 1$. The expected number of incompatible base pairs with respect to s_1 is $(1 - \frac{\beta}{\alpha^2})\, n_p(s_1)$. Since there are at least $2^{n_p(s_1)}$ different paths connecting v to $\Gamma_n^{\mathrm{X}}[s_1]$ and the probability of not selecting a vertex in $\mathrm{v}[\Gamma(s_1)]$ is a constant less than 1, and the theorem follows. ∎

5. Discussion and Comparison with RNA data

This section presents some evidence that the theory developed in the preceding sections is in fact applicable to RNA secondary structure formation. In particular we shall be consider the dependence of the degree of neutrality of structures on the base pairing alphabet and their influence on the neutral networks. Deviations from random graph theory will be given an interpretation by means of molecular structures.

5.1. Neutral Neighbors in RNA Secondary Structure Networks

The key property that allows to relate random graph theory of neutral networks to combinatory maps obtained by folding RNA sequences into secondary structures is the (mean) fraction of neutral neighbors. According to our approach we distinguish fractions of neutral neighbors derived from single base exchanges in unpaired regions ($\bar{\lambda}_u$) and those fractions derived from base pair exchanges in double helical stacking regions ($\bar{\lambda}_p$). Random samples of sequences (of constant chain length) were chosen and the distribution of neutral neighbors was computed separately for unpaired and paired regions thus yielding the corresponding average values $\bar{\lambda}_u$ and $\bar{\lambda}_p$, respectively. In order to be able to apply the random graph theory of neutral networks to RNA folding these average values have to fulfill the following three properties:

(1) Sequences folding into a given structure s are distributed (approximately) uniformly on the graph of compatible sequences $\mathcal{C}[s]$.

(2) The fraction of neutral neighbors λ_u and λ_p becomes independent of the chain length n for long chains in order to ensure a finite value of $\bar{\lambda}$ for sufficiently long sequences.

(3) The variance of the fraction of neutral neighbors for unpaired and paired regions vanishes for long chains (the *relative* vertex degree δ_v/γ becomes constant for large values of n as predicted by our model.

Inverse folding [22] was used to create large independent samples of sequences folding into a predetermined secondary structure, for example the secondary structure of a tRNA. The distribution of sequences forming such a given (common) structure is (statistically) indistinguishable from a sample of random sequences as far as pair distance statistics is concerned [35].

Table 1. Percolation threshold λ^* and asymptotic values for the fractions of neutral neighbors ($\bar{\lambda}_u$ and $\bar{\lambda}_p$).

α	*Random Graph Model* λ^*	Alphabet	*Experimental (RNA fold)* unpaired	paired
2	0.5	**GC**	0.271	0.436
		AU	0.352	0.495
4	0.370	**GCXK**	0.479	0.509
		GCAU	0.495	
6	0.301	**GCAU**		0.455

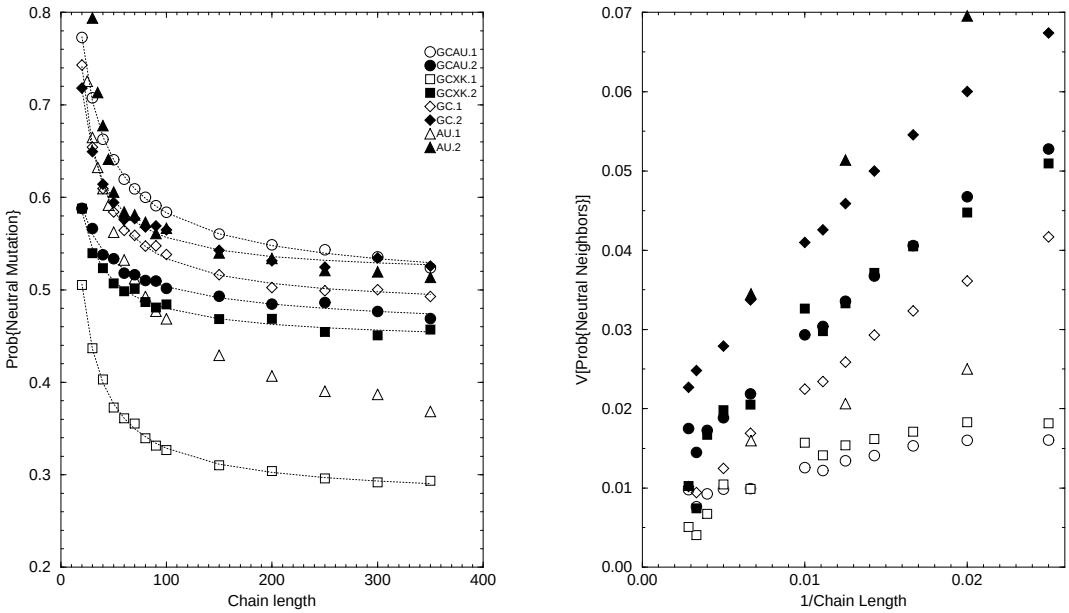


Figure 5: **a)** Frequency of neutral mutations, counted separately for single base exchanges in unpaired regions (open symbols) and base pair exchanges (full symbols) for different alphabets: **GCAU** circles, **GCXK** squares, **GC** diamonds, and **AU** triangles.
b) Variances in frequency for neutral mutations as a function of the inverse chain length.

The dependence of $\bar{\lambda}_u$ and $\bar{\lambda}_p$ on the chain length n of RNA molecules is shown in figure 5a. The curves indicate that the fractions of neutral neighbors approach constant values in the limit of long chain lengths. Asymptotic values are summarized in table 1. Figure 5b presents the variance of the fractions of neutral neighbors as functions of the inverse chain length $(1/n)$.

The variance of the fractions of neutral neighbors for the complete random sample is greater than (or at best equal to) the variance in a single neutral network, because different networks yield different values for the (structure s dependent) fractions $\overline{\lambda_u}(s)$ and $\overline{\lambda_p}(s)$. As every random variable $\hat{Y}$ on a finite set X fulfills

$$\mathbf{V}[\hat{Y}] = \mathbf{E}[\mathbf{V}[\hat{Y}|_{X_i}]] + \mathbf{V}[\mathbf{E}[\hat{Y}|_{X_i}]]$$

Here the values of $\hat{Y}$ are the fractions of neutral neighbors of unpaired and paired bases, respectively, and the partition of sequence space consists of the preimages of the secondary structures. The empirical data in figure 5b are consistent with the vanishing variance condition in the limit of long chains $(\lim n \to \infty)$ as required by the random graph model.

The data shown in table 1 allow for a straightforward molecular interpretation. The essential global molecular quantity that determines percolation behavior is the ratio between the number of sequences and the number of structures. An upper bound for the number of minimum free energy structures (S_n) was derived recently [35]

$$S_n = 1.4848 \times n^{-3/2} (1.8488)^n \ . \tag{23}$$

In figure 6 we compare this upper bound with the number actual numbers of different secondary structures of **GC** only and **AU** only sequences that were determined by folding all possible sequences up to chain length $n = 30$ [18]. (The number of **AUGC** sequences is by far too large to allow for

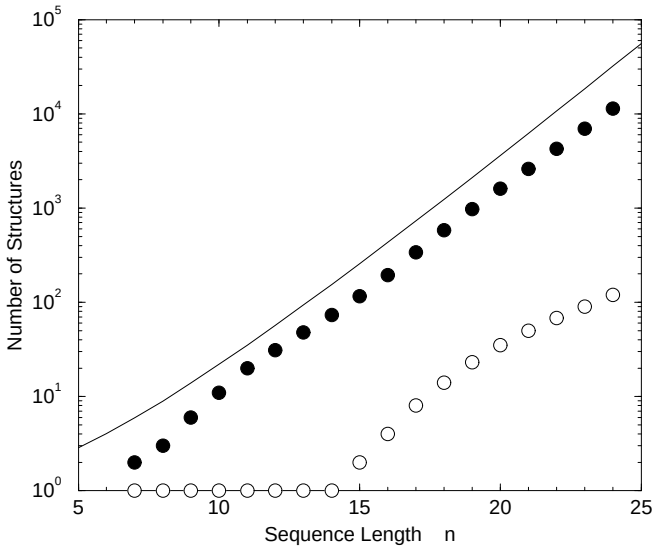


Figure 6: The number of secondary structures that occur as minimum free energy structures for **GC**, •, and **AU**, ∘, respectively. For comparison the upper bound, equ.(23), is displayed as solid line.

complete folding). The sequence to structure ratio is smallest in case of the **GC** only molecules: there are many more sequences in the **AUGC** case and the **AU** only system has a much smaller number of structures. The realistic examples shown in table 1 fall into three classes: **GC** only sequences are below percolation threshold for both paired and unpaired bases, **AU** only sequences are below percolation threshold for unpaired and above percolation threshold for the paired regions, and four letter sequences are generally above the threshold.

5.2. Shape Space Covering

The shape space covering theorem 8 predicts that the expected Hamming distance of a neutral network from a randomly chosen starting point is (up to a constant of order $o(1)$) given by the distance r_0 of the set of compatible sequences $\mathbf{C}[s]$. Hofacker [21] has computed upper bounds for these Hamming distances (r_{upper} in table 2).

Table 2. Shape Space Covering Radius.

	r_{upper}	r_0	r_{upper}	r_0	r_{upper}	r_0
n	**GC**		**AU**		**AUGC**	
50	10.7	8.5	7.0	6.0	9.2	6.5
70	12.5	15.6	11.5	9.3	13.7	10.0
100	22.9	18.6	17.3	14.6	20.5	15.2

The estimates r_{upper} are expected to become worse with increasing alphabet size and increasing chain length because of the increasing size of the search space. Data are taken from reference [21]. The value $r_0 = (1 - \beta/\alpha^2)n_p(s)$ is lower bound on the shape space covering radius.

The data in table 2 show that we have in fact at least an approximate shape space covering for all alphabets investigated so far. It is somewhat surprising to find near covering in the case of the **GC** alphabet, since both λ_p and λ_u are significantly below the threshold values. On the other hand, we expect the covering radius to increase only slowly when λ falls below its critical value. This result was confirmed for **GC** only sequences of chain length $n = 30$ [18]. We find shape space covering but the approach towards convergence is more slowly than predicted by a straightforward probability estimate [34]. All we can say at this stage of our approach is that the random model is consistent with the available data related to shape space covering.

Table 3. Upper Bounds on the Closest Approach of Neutral Networks.

n	**GC**	**AU**	**AUGC**
50	5.6	2.6	2.1
70	9.3	4.6	3.4
100	13.0	7.8	5.6

Data are taken from reference [21].

The first part of theorem 8 predicting dense penetration of sequence space by neutral networks above percolation threshold can be compared with computational data on RNA secondary structures as well. A bound on the distance between the networks of two different structures, the "closest approach" between the two networks, is given by the "closest approach" presented in table 3. The data, produced by Hofacker [21], are expected to become less accurate with increasing chain length n. As expected, the closest approach distances are much larger for the sub critical **GC** landscapes. The numerical method is not sufficiently accurate to decide whether the distance of closest approach in fact obeys theorem 8. Again our theory is consistent with the available numerical data.

5.3. Neutral Networks

The extent of neutral networks has been explored in computer simulations mostly be means of so-called *neutral paths*. Starting from a (randomly chosen) sequence v_0 a path is constructed by iteratively selecting neutral neighbor such that the distance in $\mathcal{Q}_\alpha^n$ from v_0 increases with each step. The simulation stops when no neutral neighbor can be found which increases the distance from v_0. In order to facilitate the interpretation of the data, the length $\mathcal{L}$ of a neutral path is conveniently defined as the Hamming distance (instead of the canonical distance in $\mathcal{C}[s]$) between the starting point and the end point of the path.

The frequency distribution of $\mathcal{L}$ shows a characteristic pattern which depends qualitatively only on the alphabet but not on the length of the sequence [35], see figure 7. More than 20% of all path have length $\mathcal{L} = n$ for **AUGC** or **GCXK** sequences of length $n = 100$.
Neutral paths on the two-letter alphabets terminate almost always at distances smaller than the chain length. This is due to the fact that the number of sequences contained in these distance classes is very small, while for larger alphabets ($\alpha > 2$) these classes contain exponentially many sequences, namely $(\alpha-1)^n\binom{n}{d}$. For the **GC** alphabet one finds that $\mathcal{L}$ is approximately uniformly distributed, i.e., there

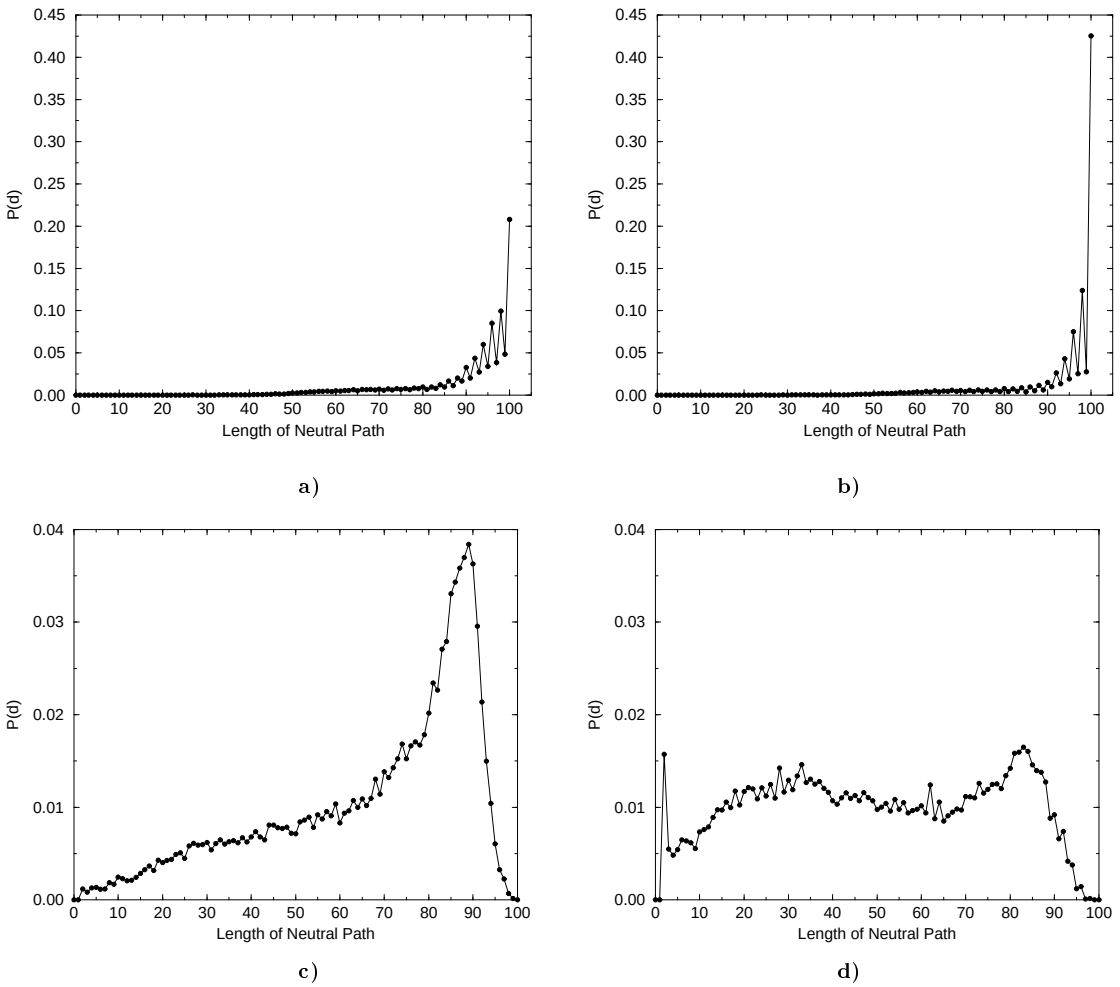


Figure 7: The probabilities of length of neutral paths in sequence space for $n = 100$. For the plots 1500 (randomly chosen) reference sequences were used.
a) AUGC alphabet, **b) GCXK** alphabet, **c) AU** alphabet, **d) GC** alphabet.

are lots of short neutral walks terminating after just a few steps. These data indicate that the neutral networks of the majority of **AUGC** sequences percolate, while for **GC** this not so in most cases. **AU** only sequences provide an interesting intermediate case: they show the characteristic cut-off of path lengths when approaching the chain length n (a consequence of the two-letter alphabet), but they show an increase in the probability of the path lengths almost until n. The numerical data on path lengths are fully consistent with the result that the average neutral network of **AU** only sequences lies above threshold in the paired region (table 1).

These data indicate that neutral networks of the **AUGC** sequences percolate in general, while for **GC** this is not the case in most cases. A more detailed analysis has been performed for the neutral networks of the **AUGC** alphabet [42], see figure 8. It shows very clearly that only the more frequent structures percolate, while neutral networks of rare sequences exhibit significantly shorter neutral paths. Again this is consistent with the predictions of our random graph models. A quantitative comparison is impossible at the moment since the values of λ_p and λ_u for the neutral networks used in figure 8 are not yet known.

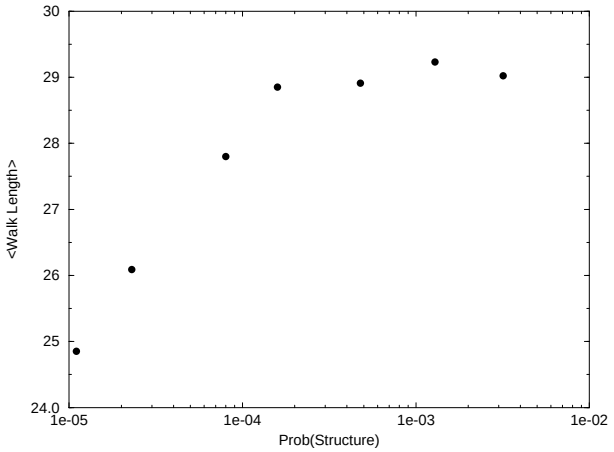


Figure 8: Average length of a neutral path for the **AUGC** alphabet as function of the abundance of the corresponding secondary structure. This is (good) lower bound on the diameter of the neutral network containing the path. Data are for **AUGC**-alphabet and chain length 30.

Connectedness of the neutral networks is a property closely related to *percolation*. A sensible definition of percolation on configuration spaces which comes as close as possible to the conventional definition of percolation on regular, low-dimensional lattices as given in [24, 40]:

Definition 12. *A subgraph $G < \mathcal{C}_n$ percolates if there exist two vertices $v, v' \in \mathrm{v}[G]$ with $d(v, v') = \mathrm{diam}[\mathcal{C}_n]$ that are connected by a path in G.*

Note that percolation neither implies nor is implied by *connectedness*. However, percolation is closely related to density and connectedness (apart from pathological cases).

5.4. The Sequence of Components

Computing the sequence of components of a neutral network, as defined in section 3.3, is a formidable task. It requires the knowledge of all sequences folding into the target structure and the subsequent sorting of these sequences into the components. Since no algorithm for complete inverse folding is known, i. e. that efficiently generates all sequences folding into a target structure, one has to resort to the brute force approach of folding at least the complete set of compatible sequences. It is no wonder therefore, that at present there is only a single source for this kind of data, namely a complete listing of the combinatory map

$$\{\mathbf{G}, \mathbf{C}\}^{30} \longrightarrow \mathcal{S}_{30}$$

for RNA secondary structures, [18].

Here one finds significant deviations between the predictions of the random graph models discussed here and the behavior of the RNA secondary structure. Three classes of neutral networks with values of $\overline{\lambda}(s)$ significantly above the threshold values for density and connectedness can be distinguished:

(1) Connected neutral networks are predicted by theorem 6 and theorem 7. In fact, the five most frequent secondary structures have connected networks, see table 4.

Table 4. Selected Sequences of Components in $\mathbf{GC}_{30}$

Rank	Structure	λ_u	λ_p	Sequence of components†
1	........(((((((((....)))))))))	0.860	0.895	1568485
5	..........(((((((....))))))))	0.614	0.747	1328606
6	(((((((....)))))))..........	0.611	0.742	1314205, [2]
7	........(((((((.......)))))))	0.666	0.748	637048, 603435
10	(((((((......)))))))........	0.652	0.751	622112, 583934
50	........(((((((......))))))).	0.570	0.749	77540, 167420, 162767, 157777
1974	......((..(((((((...)))))))..))	0.562	0.576	118307
1975	.((((((.......)))))).(((...)))	0.367	0.459	33824, 31163, 30751, 22388, [173]
1976	((((..((((....)))).)))).....	0.420	0.312	117782, [514]
1983	........((.((((((.....)))))).))	0.360	0.420	53691, 34137, 17123, 12379, 225, 215, [245]
1984	...((((((.....)))))).....	0.323	0.499	$\lvert f^{-1} \rvert$ = 117971 in 455 components*
3030	..((((((.........)))))).....	0.305	0.336	$\lvert f^{-1} \rvert$ = 88811 in 804 components*
4723	..((((((.........)))))).....	0.286	0.373	$\lvert f^{-1} \rvert$ = 58580 in 649 components*
13135	((((......)))).((((......)))).	0.214	0.246	$\lvert f^{-1} \rvert$ = 13737 in 503 components*

* See figure 9.

† Very small components are not shown in detail here. A number in square brackets gives the total number of sequences in them.
Data are taken from reference [17].

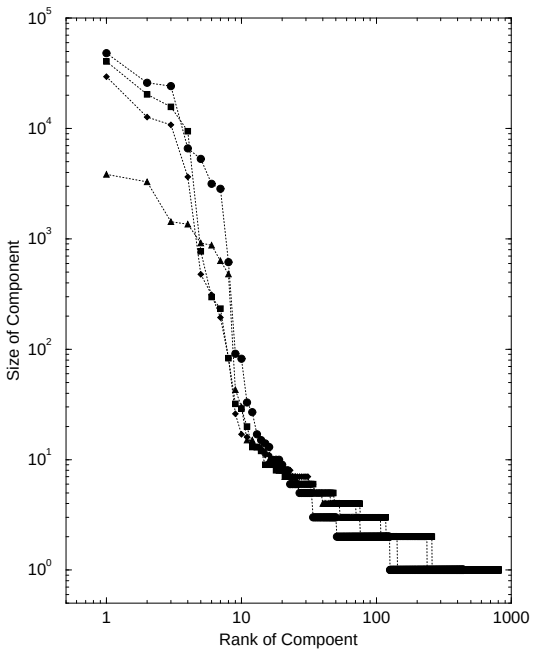
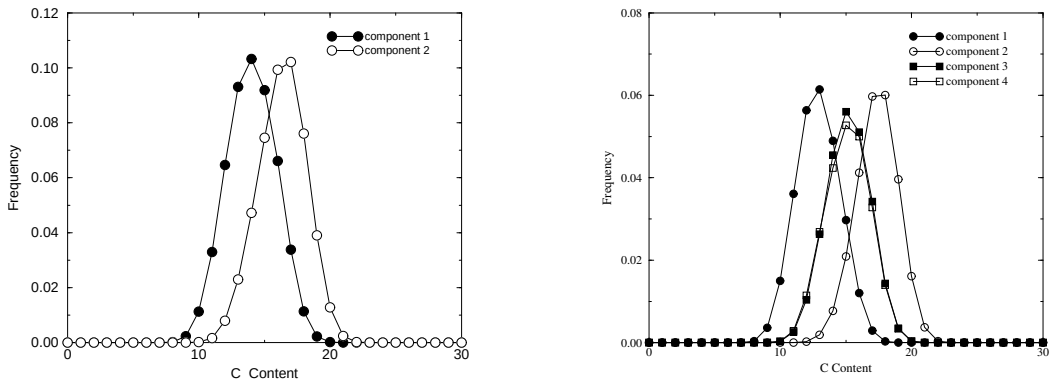


Figure 9: Sequence of Components for a few selected structures, see table 4. •: structure #1984, □: #3030, ◇: #4723, and △: #13135. The numbers refer to the rank of the structure.

(2) Neutral networks that are partitioned into a small number of components with almost equal sizes are forbidden by the random graph approach in the limit $n \to \infty$. Networks with two or four equal sized components are, nevertheless, common in the $\mathbf{GC}_{30}$ reference sample (see table 4). Two explanations are conceivable: (i) finite size effects (sequences with $n = 30$ are rather short indeed) or (ii) an "anisotropy" of structures in sequence space that is caused by the base pairing logic. Detailed inspection of the underlying structures strongly favors the second interpretation

(see below).

(3) Neutral networks consisting of one large component or a small number of large components of almost equal size plus a number of very small components (mostly isolated vertices) which together contain only negligible small fraction of the neutral network. We suspect that the small components are a finite size effect, and that therefore only the large components are relevant.



a) **b)**

Figure 10: Large neutral networks can decompose into a small number of components with almost equal sizes which differ by their $\mathbf{G}/\mathbf{C}$ ratio.
a) Neutral network of the structure with rank 7, see table 4.
b) Neutral network of the structure with rank 50. The decomposition into four components is explained in the text.

Structures with λ_u and λ_p above the threshold values show one, two or four (almost equal sized) components when they have none, one or two structural elements that allow to form additional base pairs (figure 10). These elements are, for example, stacking regions with two dangling ends, hairpin loops with five or more members or sufficiently large bulges, internal loops or multi-loops. Structures with these elements are very likely to form the additional base pair whenever complementary bases are in the appropriate positions. This is more likely to be avoided (and the structure in question is more likely to be formed as minimum free energy structure from the corresponding compatible sequence) when there is an excess of $\mathbf{G}$ or $\mathbf{C}$, $\delta\#\mathbf{G}$ or $\delta\#\mathbf{C}$ respectively, in the base composition of the sequence (the $\mathbf{G}/\mathbf{C}$ ratio being significantly different from one). One structural element that allows to form an additional base pair thus implies a bias towards more $\mathbf{G}$ or more $\mathbf{C}$ in the average composition of the sequences just as it is found with the two components in a two-component neutral network (figure 10a). Two independent structural elements are superimposed in the relative $\mathbf{G}/\mathbf{C}$ content and thus we have

$$\delta\#\mathbf{G}\,\&\,\delta\#\mathbf{G}\ ,\ \delta\#\mathbf{G}\,\&\,\delta\#\mathbf{C}\ ,\ \delta\#\mathbf{C}\,\&\,\delta\#\mathbf{G}\ \text{and}\ \delta\#\mathbf{C}\,\&\,\delta\#\mathbf{C}\ .$$

The first and the last combination show a net $\mathbf{G}$ or $\mathbf{C}$ bias whereas compensation brings the other two cases back into the middle of sequence space. Exactly this distribution of the four components is found by analyzing the data of the reference sample ($\mathbf{GC}_{30}$) [18].

For structures with λ_u and λ_p way below the critical values we find many components and a characteristically decreasing size distribution of these components, see figure 9, as predicted by random graph theory.

5.5. Diffusion on Neutral Networks

Eigen's theory of molecular evolution [7], formulated later as the theory of the *molecular quasispecies* [9, 8], describes the evolution of a population of haploid individuals on the sequence space. Each sequence replicates independently of all other members of the population with a sequence dependent replication rate A_v and a single digit replication accuracy q (implying a probability for correctly replicating the entire sequence given by q^n, n being the sequence length).

The most prominent feature of this model is the existence of an *error threshold*, i.e., a critical mutation rate $p_* = 1 - q_*$ such that replication process becomes essentially random, thereby destroying the sequence information, if $p > p_*$. The value of p_* depends crucially on the structure of the landscape of replication rates A_v. Derrida and Peliti [6] have shown that $p_* = 0$ for the "flat landscape" $A_v = A$. Let N be the population size, and denote by $x_{i,\kappa}$ the frequency of nucleotide κ at sequence position i. The vector $(x_{i,\kappa})$ is called the *barycenter* of the population. The time average $\bar{\Delta}^2(\tau)$ of the mean square displacement of the barycenter,

$$\Delta(t,\tau)^2 \stackrel{\text{def}}{=\!=} \sum_{i,\kappa} \big((x_{i,\kappa}(t+\tau) - x_{i,\kappa}(t) \big)^2$$

increases linearly with τ for sufficiently small time steps τ on the flat landscape. This indicates that the population *diffuses* in sequence space with a diffusion coefficient

$$D \stackrel{\text{def}}{=\!=} \lim_{\tau \to 0} \frac{\partial \bar{\Delta}(\tau)^2}{\partial \tau}.$$

Extensive computer simulations of this model have been performed by Fontana *et al.* [14, 13]. Recently, Huynen *et al.* [23] considered the evolution of a quasispecies on a landscape that is closely related to the combinatory map of the structure formation. They chose the structure of tRNA as a "target" and defined the replication rate constants as a decreasing function of the (tree edit) distance to this target. On this landscape the genotypic error threshold p_* is indistinguishable from 0. Nevertheless, the phenotype, that is, the target structure is conserved for moderate mutation rates. They are lost in the simulations if the p exceeds the *phenotypic error threshold* $p^\dagger$. Below this critical value, the population behave similar to a population on a flat landscape, for instance one can observe diffusion with a diffusion coefficient given by

$$D \approx \lambda' A_{\text{target}} \frac{6np}{3 + 4pN}$$

where λ' is the average fraction of neutral neighbors of a sequence in the population. Spreading on neutral networks is therefore a diffusion like process as postulated by Ronald Fisher, and analyzed in detail by Motoo Kimura [25].

Reidys *et al.* [30] have considered a related system in which the replication rate landscapes is defined by a random subgraph of the hypercube. All individuals on this subgraph have replication rate A and, all other species have a common inferior fitness. A long the lines of [28] it has been shown that there is a critical mutation rate such that the population falls off the network for larger p. It corresponds to the phenotypic error threshold in [23]. Furthermore, the population shows diffusion on the network below the critical mutation rate. These authors also obtain an analytical approximation for the diffusion coefficient which, up to first order in p, agrees with the observations in [23].

It seems hence, that the random subgraph approach to neutral networks is powerful enough to explain the dynamics of a quasispecies on a landscapes g which are completely specified by the sequence structure mapping, i.e., which can be decomposed as

$$g : \mathcal{C}_n \longrightarrow \mathcal{S}_n \longrightarrow \mathbb{R}.$$

6. Conclusions

In this contribution we have shown that

(1) the number of sequences, the number of shapes, and the distribution of preimage sizes alone is insufficient to construct a realistic model of a sequence-structure mapping;

(2) the missing ingredient is the correlation between structures of related sequences which is captured by the frequency λ of neutral neighbors;

(3) random subgraphs of configuration spaces obtained by independent point selections are appropriate models of sequence-structure mappings at least in the case of RNA secondary structures.

(4) these random subgraphs exhibit threshold values for the density and connectedness (both of which are closely related to a percolation phenomenon).

(5) RNA secondary structures require a refinement of the above model taking into account explicitly that unpaired and paired regions of the molecule have very different probabilities for neutral mutations;

(6) these refined models are consistent with most of the computational data obtained from minimum energy folding, including the feasibility of inverse folding, the existence of neutral paths, the shape space covering, neutral evolution on RNA landscapes and the existence of a phenotypic error threshold.

(7) differences between the sequence of components observed in computational studies and the sequence of components predicted from the random subgraph model can be explained at the level of the details of the base pairing logic of RNA.

(8) the sets of compatible sequences of any two secondary structures have a non-empty intersection implying that the neutral networks of any two structures come close to each other.

The random subgraph model provides a tool for fast and yet realistic simulations of RNA-based evolution and adaption to predefined conditions since they give rise to realistic landscapes without requiring the time-consuming task of explicitly computing the sequence-structure mapping.

The results reported here have straightforward implications for evolutionary dynamics in a Darwian scenario, that means evolution by mutation and selection through differential fitness. Evolutionary dynamics takes place on neutral networks and thus combines essentially two features:

(i) Spreading on neutral nets in a diffusion-like process (as postulated by Ronald Fisher, analyzed in detail Motoo Kimura [25], and simulated by computer for RNA structures by Huynen *et al.* [23].

(ii) Transitions from lower to higher fitness levels mediated by the overlapping regions of the two structures which is tantamount to the set of sequences that are compatible with both structures

How can the results derived for RNA secondary structures applied to other problems in molecular biophysics? The primary prerequisite for the use of conventional random graph theory was the representation of RNA secondary structures by simple, undirected, unweighted, and uncolored graphs. In other words, a given nucleotide is either base paired uniquely to another nucleotide, or it is unpaired. Classification of monomers is the basis of the approach pursued here. Similarly model proteins on lattices could be studied by a closely related approach although the advantageous features of a complementary logic would be lost there. Further generalizations to real 3D structures of either RNA or (lattice) proteins are conceivable; they will, however, require edge-colored or even edge-weighted graphs.

Acknowledgments

We thank Walter Grüner for providing unpublished data on the neutral networks of RNA secondary structures. Stimulating discussions with A. Dress, Ivo L. Hofacker, Christian V. Forst and J. Weber are gratefully acknowledged. Special thanks to C. Schaeg for carefully reading the manuscript.

Appendix A: Integer Valued Random Variables

The sieve formula [3] implies a number of results about the convergence in distribution for a sequence of integer valued random variables $(\hat{X}_n)$. The first theorem and its corollary indicate how the distributions of a sequence of random variables $(\hat{X}_n)$ are asymptotically determined by their factorial moments.

Theorem 9. *Let $(\hat{X}_i)_{i\in\mathbb{N}}$ be a sequence of non-negative integer valued random variables such that*

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

and

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

Then we have the following convergence in distribution: $\hat{X}_n \longrightarrow \hat{X}$.

Proof. [3, p. 23] ∎

The following corollary shall be used frequently in the paper:

Corollary 8. *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. If $\forall r \in \mathbb{N}\colon\ \mathbf{E}[\hat{X}_n]_r - \mu^r = 0$ then we have the following convergence in distribution:*

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

where P_μ is the Poisson measure. In other words $\hat{X}_n$ is asymptotically Poisson distributed.

The following classical results is also used in the main text. A proof can be found, e.g., in [10]

Theorem 10. *[Moivre-Laplace] Let X be a binomially distributed random variable, i.e., $\boldsymbol{\mu}\{\, X = k\,\} = \binom{m}{k} p^k \cdot (1-p)^{m-k}$, then*

$$\boldsymbol{\mu}\{\, X = k\,\} \sim \frac{1}{\sqrt{2\pi\, p(1-p)m}} \int_{k-1/2}^{k+1/2} \exp\left(-\frac{(x-pm)^2}{2p(1-p)m}\right)$$

In this contribution we frequently make use of the following

Corollary 9. *If $x \stackrel{\text{def}}{=\!=} \Delta/\sqrt{p(1-p)m} \to \infty$ for $m \to \infty$, then*

$$\boldsymbol{\mu}\{\, X \geq p\,m + \Delta\,\} \;\sim\; \frac{1}{\sqrt{2\cdot\pi x}}\, \exp(-x^2/2)\,.$$

Appendix B: Notation

$\mathcal{C}_n$	Configuration space
$\mathcal{Q}_\alpha^n$	generalized hypercube
n	Parameter of system size, in particular chain length
γ	degree of the regular graph $\mathcal{C}_n$
$\mathcal{C}[s]$	Graph of compatible sequences with respect to s
v$[G]$	Vertex set of the graph G
$\mathbf{C}[s]$	v$[\mathcal{C}[s]]$
$\mathcal{S}_n$	The shape space, in particular the space of RNA secondary structures in n vertices
e$[G]$	Edge set of the graph G
$d[G]$	Average vertex degree of G
$\mid X \mid$	Cardinality of X as a set.
δ_v	Vertex degree in a corresponding graph
$\hat{X}$	X is a random variable
$\mathbf{E}[\hat{X}]$	Expectation value of the random variable $\hat{X}$.
$\mathbf{V}[\hat{X}]$	The variance of $\hat{X}$
$\mathbf{E}[\hat{X}]_r$	The r-the factorial moment of $\hat{X}$.
$\mathcal{X}, \mathcal{X}'$	Components of a random graph Γ
$\partial\mathcal{X}$	The adjacent vertices to a component $\mathcal{X} < \Gamma$.
$\overline{\mathcal{X}}$	v$[\mathcal{X}] \cup \partial\mathcal{X}$
$B_r(v)$	$\{v' \in \mathrm{v}[\mathcal{C}] \mid d(v', v) \leq r\}$, the ball with radius r and center v.
dist(Γ_1, Γ_2)	The minimum Hamming distance between the graphs Γ_1 and Γ_2 considered as subgraphs of $\mathcal{Q}_\alpha^n$.
n_u, n_p	The number of unpaired and paired bases of a certain secondary structure.

References

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

[2] H. Bauer. *Wahrscheinlichkeitstheorie.* de Gruyter Verlag, 1991.

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

[4] S. Bonhoeffer and P. F. Stadler. Errortreshold on complex fitness landscapes. *J. Theor. Biol.*, 164:359–372, 1993.

[5] F. Buckley and F. Harrary. *Distances in Graphs.* Addison-Wesley, Reading, Ma., 1990.

[6] B. Derrida and L. Peliti. Evolution in a flat fitness landscape. *Bull. Math. Biol.*, 53:355–382, 1991.

[7] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 10:465–523, 1971.

[8] M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasispecies. *Adv. Chem. Phys.*, 75:149 – 263, 1989.

[9] M. Eigen and P. Schuster. *The Hypercycle: a principle of natural self-organization.* Springer, Berlin, 1979 (ZBP:234.

[10] W. Feller. *An Introduction to Probability Theory and its Applications*, volume I and II. John Wiley, New York, London, Sydney, 1966.

[11] W. Fontana, T. Griesmacher, W. Schnabl, P. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degredation rate constants of RNA secondary structures. *Monatshefte der Chemie*, 122:795–819, 1991.

[12] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[13] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaption. *Physical Review A*, 40(6):3301–3321, Sep. 1989.

[14] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophysical Chemistry*, 26:123–147, 1987.

[15] 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.

[16] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA*, 83:9373–9377, Dec. 1986.

[17] W. Grüner. *Evolutionary Optimization on RNA Folding Landscapes.* PhD thesis, Inst. of Theoretical Chemistry, Uni. Vienna, Austria, June 1994.

[18] W. Grüner, R. Giegerich, D. Strohmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence-structure maps by exhaustive enumeration. *Monatsh. Chem.*, 1995. submitted.

[19] R. W. Hamming. Error detecting and error correcting codes. *Bell Syst. Tech. J.*, 29:147–160, 1950.

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

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

[22] 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.

[23] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *PNAS*, 1995. submitted.

[24] H. Kesten. *Percolation Theory for Mathematics*. Birkhäuser, Boston, Basel, Berlin, 1982.

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

[26] H. M. Martinez. An RNA folding rule. *Nucl.Acid.Res.*, 12:323–335, 1984.

[27] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[28] M. Nowak and P. Schuster. Error tresholds of replication in finite populations, mutation frequencies and the onset of Muller's ratchet. *Journal of theoretical Biology*, 137:375–395, 1989.

[29] R. Nussinov, G. Piecznik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1):68–82, Jul. 1978.

[30] C. Reidys, C. V. Forst, and P. Schuster. Replication on neutral networks of rna secondary structures. *Bull. Math. Biol.*, 1994. in preparation.

[31] Riordan. *An Introduction to Combinatorial Analysis*. Princton University Press, 1978.

[32] W. Salser. Globin messenger RNA sequences - analysis of base-pairing and evolutionary implications. *Cold Spring Harbour Symp. Quant. Biol.*, 42:985, 1977.

[33] D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B. Lang, and R. Cedergren. Gene comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. USA*, 89:6575–6579, 1992.

[34] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *Biotechnology*, 1995.

[35] 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.

[36] J.-P. Serre. *Linear Representations of Finite Groups*. Springer, 1977.

[37] B. A. Shapiro. An algorithm for comparing multiple RNA secondary stuctures. *CABIOS*, 4(3):387–393, 1988.

[38] B. A. Shapiro and K. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*, 6:309–318, 1990.

[39] P. F. Stadler and R. Happel. Correlation structure of the landscape of the graph-bipartitioning-problem. *J. Phys. A.: Math. Gen.*, 25:3103–3110, 1992.

[40] D. Stauffer. *Introduction to Percolation Theory.* Taylor and Francis, London, 1985.

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

[42] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Robust properties of RNA secondary structure folding algorithms. *In preparation*, 1994.

[43] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 17:167–192, 1988.

[44] M. S. Waterman. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.*, 1:167 – 212, 1978.

[45] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, 1984.

[46] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

Table of Contents

1. Introduction 1

2. Configuration Spaces and Landscapes 4

2.1. Configuration Spaces 4

2.2. Combinatory Maps and Landscapes 5

2.3. Random Maps 6

3. Random Induced Subgraphs 10

3.1. Basic Random Graph Models 10

3.1.1. Basics of Graph and Probability Theory 10

3.1.2. Model and some Basic Facts 12

3.2. Density 14

3.3. Connectedness and the Sequence of Components 16

3.3.1. Definitions and two Auxiliary Lemmata 16

3.3.2. Connectedness in Generalized Hypercubes 18

3.3.3. Connectedness and Giant Components in General Configuration Spaces 23

4. Neutral Networks of RNA Secondary Structures 28

4.1. RNA Secondary Structures and Compatible Sequences 28

4.2. Neutral Networks as Random Induced Subgraphs 30

4.3. Shape Space Covering 33

5. Discussion and Comparison with RNA data 35

5.1. Neutral Neighbors in RNA Secondary Structure Networks 35

5.2. Shape Space Covering 37

5.3. Neutral Networks 38

5.4. The Sequence of Components 40

5.5. Diffusion on Neutral Networks 42

6. Conclusions 44

Acknowledgments 45

Appendix A: Integer Valued Random Variables 46

Appendix B: Notation 47

References 48