NEUTRAL NETWORKS

OF

RNA SECONDARY STRUCTURES

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der mathematischen Fakultät
der Friedrich-Schiller-Universität Jena.
von

Diplom Mathematiker Christian Michael Reidys

geboren am 10.04.1966 in Lippstadt

Gutachter:

1.) Prof.Dr. Fichtner

2.) Prof.Dr. Althöfer

3.) Prof.Dr. Schuster

Tag des Rigorosums: 23.08.1995

Tag der öffentlichen Verteidigung: 06.09.1995

Danksagung

Diese Arbeit entstand von März 1993 bis Dezember 1994 am Institut für Molekulare Biotechnologie (IMB), Jena und am Institut für Theoretische Chemie, Wien. An beiden Instituten waren ausgezeichnete Arbeitsbedingungen gegeben. Zahlreiche Konferenzbesuche ermöglichten zudem einen regen Gedankenaustausch über das Kollegium hinaus.
In dieser Zeit habe ich zahlreiche neue Freunde gefunden und bin, ob in Jena, Wien, oder Santa Fe, immer herzlich aufgenommen worden. Mein besonderer Dank im Zusammenhang mit dieser Arbeit gilt:

Prof. Dr. Peter Schuster für ausgezeichnete Betreuung der vorliegenden Arbeit, deren Ergebnisse nicht ohne eine enge Zusammenarbeit möglich gewesen wären.

Prof. Dr. Fichtner für hervorragende Betreuung bei der mathematischen Formulierung.

Unserem Universalgenie *Dr. habil. Peter Stadler* für die herzliche Aufnahme in die Wiener Arbeitsgruppe und geduldige Hilfestellung bei der Verfassung lesbarer Publikationen.

Prof. Dr. Andreas Dress für viele Anregungen und Diskussionen, insbesondere im Juni 1994 in Oberwolfach.

Dr. Christian Forst, mit dem sich binnen kürzester Zeit über eine produktive Zusammenarbeit hinaus auch eine herzliche Freundschaft entwickelte.

Meinem Freund *Stephan Kopp*, mit dem so manche Publikation nächtens diskutiert wurde.

Prof. Dr. Karl Otto Greulich für seine Unterstützung und freundschaftlichen Rat.

Dr. Ivo L. Hofacker, Dr. Walter Fontana, Dr. Martin Huynen, Jacqueline Weber und Erich Bauer für eine Zusammenarbeit, die einfach Spaß gemacht hat.

Dr. habil. Gottfried Jetschke, stellvertretend für die Arbeitsgruppe der Theoretischen Ökologie, für die gemeinsamen Seminare.

Meinen Eltern danke ich an dieser Stelle sicher nicht nur für die Finanzierung meines Studiums.

Table of Contents

Zusammenfassung

In der von Erdős und Rèyni begründeten Zufallsgraphentheorie werden neue Modelle für Zufallssubgraphen von *Konfigurationsräumen* vorgestellt. Diese Subgraphen formen Wahrscheinlichkeitsräume, in denen die Eigenschaften *Dichte* und *Zusammenhang* der entsprechenden Graphen 0-1-Gesetze erfüllen. Genauer heißt dies, daß in einem gewissen Limes ein *Schwellenwert* existiert, unterhalb dessen Zusammenhang und Dichte für keinen Zufallsgraphen gegeben sind, aber oberhalb dessen alle Zufallsgraphen dicht und zusammenhängend sind. Es kann nachgewiesen werden, daß für jeden positiven Konstruktionsparameter in einem Zufallsgraphen eine einzige riesige Komponente existiert. Diese Resultate nutzen wir für das Studium der *Sequenz-Struktur-Abbildungen*, wobei wir unter "Struktur" RNA-Sekundärstrukturen verstehen. Einzelne Urbilder dieser Abbildung, die *neutralen Netze*, werden als Zufallssubgraphen des Graphen der *kompatiblen Sequenzen* konstruiert. Dichte und Zusammenhang der neutralen Netze, im Graphen der kompatiblen Sequenzen, werden mit Hilfe der zuvor erzielten Resultate analysiert. Wir werden nachweisen, daß in jedem Fall die neutralen Netze im Limes unendlicher Kettenlänge eine einzige große Komponente besitzen. Ferner besitzen je zwei neutrale Netze im Sequenzraum einen geringen Hamming–Abstand und wir können (im Rahmen des Modells) einen Beweis der *shape space covering conjecture* erbringen. Im Anschluß untersuchen wir die Dynamik des Replikationsprozesses einer endlichen Population von RNA-Molekülen (auf limitierten Ressourcen) in einer von einem neutralen Netzwerk induzierten *Landschaft.* Es stellt sich heraus, daß das grundlegende *Fehlerschwellenkonzept* von Eigen *et al.* auf diese Typen von Landschaften erweitert werden kann. Alle Resultate werden konsequent für endliche Populationen formuliert. Weiter wird ein Kriterium für die Lokalisierung der Fehlerschwelle erarbeitet und es erweist sich eine gute Übereinstimmung mit den parallel durchgeführten Gillespie-Simulationen. Die Population auf dem neutralen Netzwerk bewegt sich gemäß einer Diffusionsgleichung und wir können die Verteilung der Paarabstände in Abhängigkeit von der Replikationsgenauigkeit einzelner Positionen bestimmen. Schließlich beschäftigen wir uns mit der algebraischen Darstellung von RNA Sekundärstrukturen. Wir fassen die Biopolymerstrukturen als *Kontakt-Strukturen* auf und betten diese in Involutionen bzw. Untergruppen der S_n ein. Aus diesen Darstellungen ergeben sich dann verschiedene Metriken, mit deren Hilfe sich das Konzept der *Quasispezies von Sekundärstrukturen* formulieren läßt.

1. Introduction

1.1. Theoretical Biology

Theoretical biology is one of the scientific fields that has experienced an enormous development in the last decades. Computers have become an important scientific tool and have produced a wealth of data based on simulations. Except for the field of population genetics [16, 67] we are, however, far away from having a sound mathematical foundation of theoretical biology comparable to that of physics or chemistry. To cope with this lack of mathematical theories is precisely what mathematical biology is committed to.

Theoretical biology – even at the molecular level only – attempts to describe and analyze phenomena of particular complexity. The important contribution that mathematics is able to make is that of providing underlying models based on adequate and inherently drastic reductions. Mathematical modeling has often proved to highlight the essential features of complex processes. It turns out that many phenomena can be described by surprisingly simple rules. In this context we mention, for example

- the theory of *cellular automata* founded by John. v. Neumann
- the concept of *molecular quasispecies* of Manfred Eigen, John McCaskill and Peter Schuster [12].
- the concept of evolutionary stable states of J. Maynard Smith and G. R. Price [45, 44].

All three concepts are famous examples of how mathematical theory and biosciences interact synergetically.

Leaving those fundamental results we observe that for most other areas of theoretical biology, however, there are well defined abstract models but no developed mathematical theory. One example are the *Random Boolean networks*, introduced by Stuart Kauffmann, which have been studied extensively by computer simulations. [35, 34, 33]. At present there is no mathematical theory of those networks which would give us information on, say, the distribution of *cycle lengths*, or the number of *basins of attraction*. A first attempt in this direction is the work of Jim Lynch [41, 42] who recently proved a beautiful theorem[1] on the chaotic behavior of random Boolean networks with exactly two inputs.

Another interesting example is the field of *artificial life* [36, 53], or "Alife", founded by Chris Langton [40]. "Alife" is an example of a field without a well defined methodology. It shows beautifully how purely theoretical biological research can be. Instead of restricting their attention to the description of existing biological processes, theoreticians began to study completely artificial scenarios in the hope of being able to distinguish the generic features of "life" from the historical contingencies of the evolution on our planet. Here again, the research is done almost exclusively in form of computer experiments and simulations. As yet, "Alife" is completely lacking a unifying mathematical description.

1.2. Sequence to Structure Maps

Conventional biophysics is concerned with structure predictions of biopolymers that relate 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*. The sequence space is a metric space of all sequences where the metric is given by the Hamming distance [25] (counting the number of positions in which two aligned sequences differ). It has a natural graph structure by

- the vertex set given by all tuples $a = (a_1, ..., a_n)$ of length n where $a_i \in \mathcal{A}$ and $\mathcal{A}$ is a finite alphabet
- the edge set consisting of all pairs $\{a, a'\}$ such that for exactly one index holds $a_i \neq a'_i$.

The shape space is a metric space whose points are abstract structures. In general such a mapping will not be one-to-one: many sequences will be mapped onto the same structure. The degree of this redundancy will strongly depend on the notion of structure. In X-ray crystallography, structure is tantamount to a set of atomic coordinates and at sufficiently high resolution structures

[1] His proof was formulated in the language of random graph theory [14, 3].

are unique in the sense that structures from different sequences will never coincide. Molecular biologists, however, commonly apply another, more coarse-grained notion of structure when they say intuitively that two proteins have the same structure. The appropriate notion of structure is clearly context dependent and therefore 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 (see figure 1). The figure also illustrates the relation of secondary and tertiary structure–only few further contacts induce the "L"-shape of the tRNA. In addition, biochemists and molecular biologists have successfully used RNA secondary structures for molecular interpretations of RNA function.

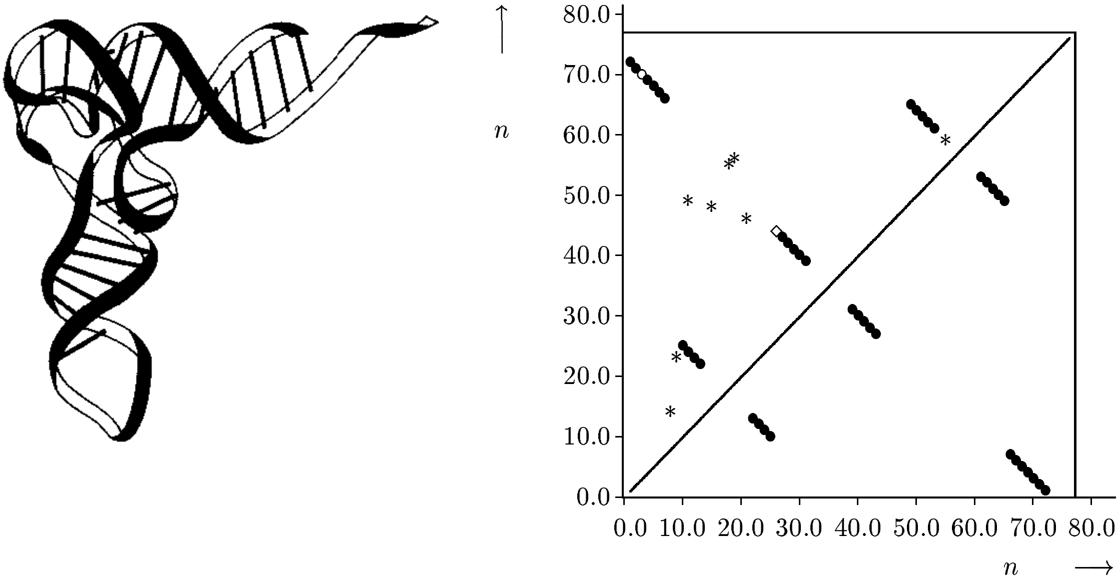


Figure 1: **a)** Three dimensional structure of phenylalanin tRNA from yeast.
b) Contact map of tRNA-phe.
The upper triangle shows the contact of the three-dimensional structure: ● Watson-Crick base pairs, ○ GU base pairs, ◇ other non-Watson-Crick base pairs in double helical regions, and * tertiary contacts between bases.
The lower triangle shows the secondary structure consisting of both Watson-Crick and GU base pairs, ●.

In this thesis we are mostly 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 [28]. Straightforward estimates on the numbers of possible secondary structures for a given chain length n show a high degree of redundancy: there are many more sequences than structures and hence many sequences have to fold into the same structure.

The relation between RNA sequences and secondary structures is understood as a (non invertible) mapping from sequence space into shape space [21]. 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 can be interconverted by a small number of mutations. The other metric space 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 [21]. 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 otherwise inaccessible. (An illustrative example is the idea of *shape space covering* described in [54].) An understanding of sequence-structure relations, however, is of central importance in biophysics of biopolymer structure. The mapping inherits essential features of *fitness landscapes* and is fundamental for conventional biotechnology since it represents the basis of rational design of biopolymers. Further 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, 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 [61].

In contrast to counting compatible sequences being able to fold 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 [24], the other using some abstract model for sequence to structure mapping and performing rigorous mathematical analysis. In this thesis 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 interconverted 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].

1.3. Basic Questions of Evolutionary Optimization

The first successful theory of biological evolution was presented last century by Charles Darwin (1859) in his famous book *The Origin of Species.* It is based on two fundamental principles, genetic variability caused by *mutation* and *natural selection.* The first principle leads to *diversity* and the second one to the concept *survival of the fittest*, where fitness is an inherited characteristic property of an individual and can basically be identified with its *reproduction rate.* In particular in his book Darwin presents also the most essential features of *neutral evolution.*

In extension of Darwin's theory of evolution the role of stochastic processes has been stated. Wright [68, 69] saw the importance of the genetic drift in evolution in improving the "evolutionary search capacity" of the whole population. He saw genetic drift merely as a process that could improve evolutionary search whereas Kimura proposed that the majority of changes that are observed in evolution at the molecular level were the results of random drift of genotypes [38, 39]. Paraphrasing the situation the "selectionist" considers the differences in fitness values to be responsible for the fixation of new genotypes whereas the "neutralist" assumes that most mutants are neutral and the fixation of new genotypes is the outcome of a stochastic process. The *neutral theory* of Kimura does not assume that selection plays no role but denies that any appreciable fraction of observable molecular change is caused by selective forces: Mutations, in this view, are either a disadvantage or at best neutral in present day organisms. A "negative selection" plays a major role in the neutral evolution that is deleterious mutants die out caused by their lower fitness.

Over the last few decades, however, there has been a shift of emphasis in the study of evolution. Instead of focusing on the differences in the selective value of mutants and on population genetics, interest has moved to evolution through natural selection as an *optimization problem* whose fundamental three ingredients are

- the configuration space, i.e., the graph formed by all possible *genotypes*,
- the set of *elementary moves* by which the search for better shapes is performed and
- the structure of the fitness landscape itself.

Fisher [16, 29] stated a fundamental theorem on optimization processes by specifying macroscopic parameters (e. g. the mean-fitness) that measure the optimization process. A further aspect in theory of optimization is to relate and analyze the above three ingredients. Apparently, in evolution the move sets are rather simple and inherently random. Here we have *point mutations, insertions, deletions* and possibly *recombinations*. Consequently in evolutionary theory much research is concentrated on the analysis of the landscapes in which evolutionary adaption takes place[2]. Evolutionary dynamics studies basically how the search for the best configurations is organized for a given landscape and configuration space. In this context a fundamental result is the concept of the molecular quasispecies introduced by Eigen and coworkers [12].

Let us return for a short moment to Darwin and have a look at his minimum requirements for adaption:

- a population of objects that are capable of replication,
- occasional variations which are inheritable, and
- restricted proliferation which is constrained by limited resources.

We first introduce a new type of landscape that is based on the concept of *neutral networks* associated to RNA secondary structures. For the moment it suffices to consider neutral networks simply as certain subsets of sequences. The main idea is then to assign to each sequence contained in

[2] This question is closely related to the study of sequence to structure maps as pointed out in the previous section

the neutral network a superior fitness whereas all other sequences have an inferior fitness. Thereby we obtain the landscape. Those landscapes combine in a natural way both the selectionists' and the neutralists' view of biological evolution namely Darwin's *survival of the fittest* and Kimura's *neutral random drift*. (In literature there are convincing evidences that RNA landscapes are as simple as they can be for evolutionary adaption [55].) Assuming this landscape to be given we proceed by studying the dynamics along the lines of Eigen and coworkers [10, 12, 52, 56, 60].

On the one hand we investigate the dynamics from the point of view of a selectionist: For a finite population of strings that replicate on the neutral networks we consider their number of master-strings i.e. those that are located on the neutral network. We can show that, as in Eigen's mean field approach the *single peak landscape*, there exists a critical mutation rate above which the populations drift randomly through sequence space.

On the other hand we investigate neutral evolution by computing the spatial distribution of the fraction of masters. We can prove that the master-fraction diffuses on the neutral network and we evaluate its *diffusion-coefficient*.

The dynamics of the replication-deletion process on a neutral network gives further theoretical insight into the search for better shapes is organized. This model is on the one hand sufficiently simple to deduce analytical expressions for basic parameters and allows on the other hand to investigate the influence of the structure of the neutral network on the dynamics.

1.4. Organization of this Thesis

This thesis is devoted to the following two fields:

- the mathematical modeling of sequence to structure maps in RNA and in particular the analysis of neutral networks of RNA secondary structures and
- the dynamics of finite populations replicating erroneously on neutral networks.

In chapter 2 we introduce so called *configuration spaces*. We show that pure random maps contradict the essential features of known sequence to structure maps in RNA.

In chapter 3 we present our random graph models and among some basic results we prove the existence of threshold values in the probability spaces formed by random subgraphs of a configuration space. Our approach is formulated in the language of random graph theory developed by

Paul Erdös, Alfred Réyni, and Béla Bollobás [3]. The mathematical core consists in proving the existence of the *threshold value* for the connectivity property of the above random graphs. The proof of this theorem is completely constructive and gives also further insight into k-connectivity properties of the graphs. Having introduced the theory of random subgraphs of configuration spaces we proceed in chapter 4 by applying it to the problem of the sequence to structure mapping in RNA. This is done by constructing single preimages of RNA secondary structures, called *neutral networks*, as random graphs. We can derive a sufficient condition for density and connectivity properties of neutral networks which have already been proven to be valid for a chain length up to 30 [24]. We furthermore prove the *shape space covering conjecture* of Schuster [54] within the random graph model and present a method for obtaining the complete sequence to structure map recursively from the random graph approach.

Chapter 5 is dealing with evolutionary dynamics on neutral networks constructed by the random graph models as described in chapter 4. We consider a finite population of erroneously replicating strings (of constant length) in a landscape induced by a single neutral network. Here we combine the concept of neutral networks and the molecular quasispecies of Eigen and coworkers [10]. We can extend the error-threshold concept to single shape-landscapes by applying a stochastic ansatz analogous to Nowak and Schuster [47]. We investigate some aspects of neutral evolution by studying the distribution of the fraction of the population that is located on the neutral network. This is done by studying *random walks* on neutral networks, extending the work of Derrida & Peliti [9]. The random walks together with sequence genealogies [7] allow to compute the distribution of pair distances in the population. Towards a theoretical understanding of evolutionary optimization we prove that the population *diffuses* on the neutral network.

In chapter 7 we discuss two algebraic representations of RNA secondary structures. The first one interprets a structure as an involution in a corresponding permutation group S_n and the second one maps a secondary structure to a subgroup of the S_n. Both approaches lead to *metrics* on RNA secondary structures.

Finally we present in chapter 8 a detailed discussion of the results derived so far and chapter 9 contains among conclusions an outlook on future projects.

2. Configuration Spaces and Random Maps

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. Mutation, thus, acts on the nucl of variation is simply the limited accuracy of replication. Replication errors or mutations produce RNA sequences which differ from the parental template sequence. Mutation, thus, acts on the nucleotides of DNA (or RNA in case of viroids and 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 point mutations.

ACGAUGGGUUACC|G|AGGCAAGUCGUAG
Point mutation ↓
ACGAUGGGUUACC|A|AGGCAAGUCGUAG

ACGAUG|GGUUACCG|AGGCAAGUCGUAG
Insertion ↓
ACGAUG|GGUUACCG|GGUUACCG|AGGCAAGUCGUAG

ACGAUGGG|UUACCGAGGC|AAGUCGUAG
Deletion ↓
ACGAUGGG|AAGUCGUAG

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.

Other types of mutations occur 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 point mutations and recombinations conserve their number. They lead to a permutation of the genome, see [51].

Mutations can be viewed as "moves" in an abstract space of *configurations.* This suggests a natural "geometrical" arrangement of the configurations (be they polynucleotides, arrangements of genes on a mitochondrial genome or something else): configurations that can be interconverted by a single move (mutation) may be viewed as neighbors. Consequently, the smallest number of moves which is necessary to interconvert two 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]. In evolution the existence of phylogenies guarantees that the mutation operators lead to a connected graph: every configuration can be reached from any other configuration by a sequence of mutations.

2.2. Random Maps

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

$$\mathbf{Map}(X, Y) := \{f \mid f : X \longrightarrow Y\}.$$

Using the *uniform 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

$$\hat{Z}_k(f) := |\{z \in Y \mid |f^{-1}(z)| = k\}|,$$

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

$$\mathbf{E}[\hat{Z}_k] := \frac{1}{y^x} \sum_{\ell=0}^{y} \ell\,|\,\{f \,|\, \hat{Z}_k(f) = \ell\}\,|\,.$$

Claim: *For $k \in I\!N$, $\mathbf{E}[\hat{Z}_k]$ 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 [49]

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

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

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

and obtain

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

Then we 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

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

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. Using the recursion formula for $|\,\mathrm{dis}(x_k,m,x)\,|$ it follows

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

and by comparison of coefficients:

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

This implies for $\mathbf{V}[\hat{Z}_k] := \mathbf{E}[\hat{Z}_k^2] - \mathbf{E}[\hat{Z}_k]^2$

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

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 observations, made for RNA sequence-to-secondary structure maps. For this purpose let $\mathcal{Q}_\alpha^n$ be the (Hamming) graph of all sequences over the alphabet $\mathcal{A}$ of length α with chain length n. Then we set

$$\wp(s) := \frac{|f_n^{-1}(s)|}{|\mathcal{Q}_\alpha^n|}$$

for the *frequency* of s. Then 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 $\psi(\imath(s)) := \wp(s)$. The rank order function is well known for the case of folding RNA into its secondary structure–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 [54, 61]. The above parameters 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 $\int_0^\infty \psi^2(\imath)d\imath = [(c-1)^2/(2c-1)]b$.

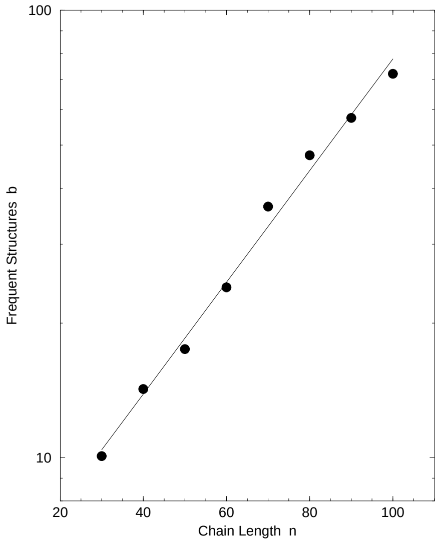


Figure 3: Dependence of the number of "frequent" structures b on the chain length n for RNA secondary structures (biophysical alphabet). These data have been obtained for so-called loop-structures, a coarse grained version of the full secondary structure graphs. For details see [18, 54, 28].

Extending the above result on $\mathbf{E}[\hat{Z}_k]$ by introducing a *a priori* probability $\wp(s)$ for mapping a x to $y \in Y$ we obtain an inhomogeneous spectrum of preimage sizes. We shall now show that even this more general 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 $\mathcal{Q}^n_\alpha$ and Y the set of RNA secondary structures $\mathcal{S}_n$. We consider the average number of neutral neighbors with respect to the structure s for a mapping that assigns each $x \in X$ with the probability $\wp(s)$ to s. Clearly this number is a random variable that is expected to have the mean $(\alpha - 1)\, n \wp(s)$. But 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 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 $\wp(s_i)$ the expected number of "neutral neighbors" tends to zero in the limit of infinite chain length $(n \to \infty)$ 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" [54] i.e. extended networks in sequence space that consist of sequences folding into a fixed secondary structure. For this reason we shall have to take into account the "correla" structures obtained from nearby configurations. We shall proceed by introducing general models for random induced subgraphs of configuration spaces.

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 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 $e = \{v, v'\}$.
- G is *γ-regular* if for each vertex $v \in \mathrm{v}[G]$ holds $\delta_v = \gamma$.
- The *order* of a graph G, $|\,G\,|$ is the cardinality of its vertex set i.e. $|\,\mathrm{v}[G]\,|$.
- A *path* π *in* G is a tuple of the form $(v = v_1, e_1, v_2, ...e_{m-1}, v_m = v')$ 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

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

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

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

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

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

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

Notation. In the sequel we write ∂ instead of ∂_G.

Definition 1. *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 $|\,\mathcal{C}_n\,|\,\gamma_n^{-\ell} \nearrow \infty$.*

(P) *For $v, v' \in \mathrm{v}[\mathcal{C}_n]$ with $v \neq v'$ 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*

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

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

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

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

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

$$\mathrm{v}[\mathcal{Q}_\alpha^n] := \{(x_1,...,x_n) \,|\, x_i \in \mathcal{A},\ 1 \le i \le n\,\}$$

$$\mathrm{e}[\mathcal{Q}_\alpha^n] := \{\{(x_1,...,x_n),(x_1',...,x_n')\} \mid \text{ for exactly one index } 1 \le i \le n \text{ holds } x_i \ne x_i'\,\}$$

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

$$v = (x_1,...,x_n) \quad \text{and} \quad v' = (x_1',...,x_k',x_{k+1},...,x_n)\,.$$

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

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

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 $\mathbf{P}_{\mathcal{Q}_\alpha^n}^{v,v',0}$ to be the following set of paths

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

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

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

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 $|\, \partial\{v\} \cap \partial\{v'\} \,| \leq \alpha$ where $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$. 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 $\mathbf{P}_{\mathcal{Q}_\alpha^n}^{v_1^{(i)}, v_1'^{(i)}, 0}$ $1 \leq i \leq \ell$ have all equal length $k+2$ and fulfill (R) (iv):

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

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

(1) the family of Boolean hypercubes $(\mathcal{Q}_2^n)_{n\in\mathbb{N}}$;

(2) the family of generalized hypercubes $(\mathcal{Q}_\alpha^n)_{n\in\mathbb{N}}$;

(3) the family of the canonical configuration spaces for the graph bipartitioning problem [58].

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 $\mathbf{E}[\hat{X}] = \sum_x x\, f(x)$ 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 furthermore introduce the r–th *factorial moment* of a positive, integer valued random variable $\hat{Y}$ by:

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

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

3.1.2. Models

Let us begin by considering the set of all subgraphs of a finite graph H i.e. $\mathcal{G}(H)$. For each pair of subgraphs $G, G' < H$ the relation $G \sim G' \iff \mathrm{e}[G] = \mathrm{e}[G']$ is an equivalence relation. In each equivalence class $[G] := \{G'' < H \,|\, G \sim G''\}$ with $\mathrm{e}[G] > 0$ there is a unique element G^* such that $|\,\mathrm{v}[G^*]\,| = \min\{|\,\mathrm{v}[G'']\,| < H \,|\, G \sim G''\}$, given by

$$G^* := (\{\tilde{\iota}(\mathrm{e}[G]) \cup \tilde{\tau}(\mathrm{e}[G])\}, \mathrm{e}[G]) \,. \tag{4}$$

For $0 \le p \le 1$ we obtain setting

$$\boldsymbol{\mu}_p\{[G]\} := p^{|\,\mathrm{e}[G]\,|}(1-p)^{|\,\mathrm{e}[H]\,| - |\,\mathrm{e}[G]\,|}$$

a probability measure on the set of equivalence classes $\mathcal{G}(H)/\sim$ since $\sum_{[G]} \boldsymbol{\mu}_p\{[G]\} = 1$ and $\big([\mathcal{G}(H)/\sim], \mathcal{P}(\mathcal{G}(H)), \boldsymbol{\mu}_p\big)$ becomes a probability space. We shall write $\boldsymbol{\mu}_p\{G^*\} := \boldsymbol{\mu}_p\{[G]\}$.

Remark. We can construct the subgraph G^* for $0 < p \le 1$ as follows: we select each edge $e \in \mathrm{e}[H]$ with the probability p and thereby obtain the set $Y \subset \mathrm{e}[H]$. In order to construct the graph G^* we define $\mathrm{v}[G^*] := \{v \in \mathrm{v}[H] \,|\, \exists e \in Y : v \text{ incident to } e\}$ and set $\mathrm{e}[G^*] := Y$.

Model I *Let H be a finite graph and G^* with $\mathrm{v}[G^*] \neq \emptyset$ be given by equation (4). Then $H[\mathrm{v}[G^*]]$ is an induced subgraph of H and we have a one-to-one correspondence between $\mathrm{v}[G^*]$ and $H[\mathrm{v}[G^*]]$. Let $\mathcal{G}^{\mathrm{I}}(H)$ be the set of all induced subgraphs Γ^{I} of the finite graph H for which there exists a $G^* < H$ with $\mathrm{v}[G^*] \neq \emptyset$ and $\Gamma^{\mathrm{I}} = H[\mathrm{v}[G^*]]$. For $\Gamma^{\mathrm{I}} \in \mathcal{G}^{\mathrm{I}}(H)$ we set*

$$\boldsymbol{\mu}^{\mathrm{I}}\{\Gamma^{\mathrm{I}}\} := \frac{1}{1-(1-p)^{|\,\mathrm{e}[H]\,|}} \sum_{\{G^* : \Gamma^{\mathrm{I}} = H[\mathrm{v}[G^*]]\}} \boldsymbol{\mu}_p\{G^*\} \ .$$

Plainly $\sum_{\Gamma^{\mathrm{I}} = H[\mathrm{v}[G^*]]} \boldsymbol{\mu}^{\mathrm{I}}\{\Gamma^{\mathrm{I}}\} = 1$ and we have the probability space

$$\Omega^{\mathrm{I}} := \big(\mathcal{G}^{\mathrm{I}}(H), \mathcal{P}(\mathcal{G}^{\mathrm{I}}(H)), \boldsymbol{\mu}^{\mathrm{I}}\big) \ .$$

Next we introduce a second model that has “better” independence properties and builds the basis for the mathematical modeling of “neutral networks” of RNA secondary structures as random graphs.

Model II *Let H be a finite graph. Each subset $X \subset \mathrm{v}[H]$ induces the subgraph $H[X]$ and more precisely 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 $\mathcal{G}^{\mathrm{II}}(H)$. Further we suppose $0 \le \lambda \le 1$ to be given and set for $\Gamma \in \mathcal{G}^{\mathrm{II}}(H)$*

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

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

$$\Omega^{\mathrm{II}} := \left(\,\mathcal{G}^{\mathrm{II}}(H), \mathcal{P}(\mathcal{G}^{\mathrm{II}}(H)), \boldsymbol{\mu}_\lambda\,\right)\,.$$

Remark. We can construct each $\Gamma \in \mathcal{G}^{\mathrm{II}}(H)$ by selecting each vertex $v \in H$ with the independent probability $0 < \lambda \le 1$. 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 1. This leads to a family of corresponding probability spaces $\Omega_n^{\mathrm{I}}, \Omega_n^{\mathrm{II}}$. If we want to emphasize that we are working with random subgraphs of a graph $\mathcal{C}_n$ (according to model I or model II) we shall write $\Gamma_n^{\mathrm{I}}, \Gamma_n^{\mathrm{II}}$. Accordingly, we shall refer to the underlying probability measures as $\boldsymbol{\mu}^{n,\mathrm{I}}$ for model I and $\boldsymbol{\mu}_{n,\lambda}$ for model II.

Remark. For model I we can further establish a connection between the basic parameter p and the probability λ of selecting a vertex $v \in \mathrm{v}[H]$ to be contained in a random graph.
For a sequence of configuration spaces $(\mathcal{C}_n)$ we assume $p_n\gamma_n = c$, with constant $c > 0$. The probability that a $v \in \mathrm{v}[H]$ is selected is then

$$\boldsymbol{\mu}_p\{v \in \mathrm{v}[G^*]\} = 1-(1-p)^{\gamma_n} \quad \text{hence} \quad \lim_{n\to\infty} \boldsymbol{\mu}_p\{v \in \mathrm{v}[G^*]\} = 1-e^{-c}\,.$$

Consequently to each $p_n = c/\gamma_n$ there corresponds a

$$\lambda^{\mathrm{I}}(p_n) := 1-(1-p_n\gamma_n/\gamma_n)^{\gamma_n}\,. \tag{5}$$

In order to compare both models it is sometimes convenient to use λ^{I} as the underlying parameter.

3.2. Orders and Vertex Degrees

Suppose a sequence of configuration spaces $(\mathcal{C}_n)$ is given. In this section we state some simple properties of the random graphs $\Gamma_n^{\mathrm{I}}, \Gamma_n^{\mathrm{II}}$. We observe that vertices $v, v' \in \mathrm{v}[\mathcal{C}_n]$ are chosen asymptotically independently if $p_n\gamma_n$ is a positive constant c independent of n. For this purpose we introduce the random variables $\hat{X}_v, \hat{X}_{v'}$ given by

$$\hat{X}_v := \begin{cases} 1 & \text{if } v \text{ is selected} \\ 0 & \text{otherwise.} \end{cases}$$

The induced σ-algebra of a random variable $\hat{X}_v$ for $v \in \mathrm{v}[\mathcal{C}_n]$ is $\{\emptyset, \{\hat{X}_v = 1 \vee 0\}, \{\hat{X}_v = 1\}, \{\hat{X}_v = 0\}\}$ and by symmetry it remains to show

$$\begin{aligned}
\boldsymbol{\mu}\{\,\{\hat{X}_v = 1\} \cap \{\hat{X}_{v'} = 1\}\,\} &= \boldsymbol{\mu}\{\,\{\hat{X}_v = 1\}\,\}\,\boldsymbol{\mu}\{\,\{\hat{X}_{v'} = 1\}\,\} \\
\boldsymbol{\mu}\{\,\{\hat{X}_v = 1\} \cap \{\hat{X}_{v'} = 0\}\,\} &= \boldsymbol{\mu}\{\,\{\hat{X}_v = 1\}\,\}\,\boldsymbol{\mu}\{\,\{\hat{X}_{v'} = 0\}\,\} \\
\boldsymbol{\mu}\{\,\{\hat{X}_v = 0\} \cap \{\hat{X}_{v'} = 0\}\,\} &= \boldsymbol{\mu}\{\,\{\hat{X}_v = 0\}\,\}\,\boldsymbol{\mu}\{\,\{\hat{X}_{v'} = 0\}\,\}\,.
\end{aligned}$$

We have $\boldsymbol{\mu}\{\,\{\hat{X}_v = 0\} \cap \{\hat{X}_{v'} = 0\}\}\,\} = (1-p)^{2\gamma_n - \varphi_n}$ where $\lim_{n\to\infty}(\varphi_n\,\gamma_n^{-1}) = 0$ (see def 1) and $\boldsymbol{\mu}\{\,\{\hat{X}_v = 0\}\,\}\,\boldsymbol{\mu}\{\,\{\hat{X}_{v'} = 0\}\}\,\} = (1-p)^{2\gamma_n}$. Then

$$\lim_{n\to\infty} (1-p)^{-\varphi_n} = e^{-c\,\lim_{n\to\infty}(\varphi_n\,\gamma_n^{-1})} = 1$$

implies $\lim_{n\to\infty} \boldsymbol{\mu}\{\,\{\hat{X}_v = 0\} \cap \{\hat{X}_{v'} = 0\}\,\} = \lim_{n\to\infty} \boldsymbol{\mu}\{\,\{\hat{X}_v = 0\}\,\}\,\boldsymbol{\mu}\{\,\{\hat{X}_{v'} = 0\}\,\}$. Along these lines one immediately verifies the other equalities. Therefore for each finite family $(\hat{X}_{v_i})_{1\le i\le r}$ the random variables are asymptotical independent i.e. $(\hat{X}_v)_{v\in \mathrm{v}[\mathcal{C}_n]}$ is a family of asymptotical independent random variables. Further we shall make use of the following argument in the main body of the thesis. Let $\hat{X}_n$ be an (integer valued) random variable such that $\lim_{n\to\infty} \mathbf{E}[\hat{X}_n] = \infty$ and $\mathbf{V}[\hat{X}_n] = \theta_n\,\mathbf{E}[\hat{X}_n]^2$ where $\lim_{n\to\infty}\theta_n = 0$. Then we obtain from the Markovian inequality

$$\forall \epsilon > 0: \qquad \lim_{n\to\infty} \boldsymbol{\mu}_n\{\,|\,\hat{X}_n/\mathbf{E}[\hat{X}_n] - 1\,| > \epsilon\,\} = 0\,,$$

i.e. $(\hat{X}_n/\mathbf{E}[\hat{X}_n]\,)_n$ converges stochastically to the constant random variable $\hat{1}$. This implies in particular convergence in distribution.

We next introduce

$$\mathcal{B}_n(k,p) := \frac{1}{\sqrt{2\pi\,p(1-p)n}} \int_{k-1/2}^{k+1/2} \exp\left(-\frac{(x-pn)^2}{2p(1-p)n}\,dx\right)\,,$$

that is the discretized version of the Gaussian distribution with mean $p\,n$ and standard deviation $\sqrt{p\,(1-p)\,n}$. Let $\hat{\omega}_n^{\mathrm{j}} : \Omega_n^{\mathrm{j}} \to \mathbb{R}$ $j = \mathrm{I,II}$ the following random variables: $\hat{\omega}_n^{\mathrm{j}}(\Gamma_n^{\mathrm{j}}) := |\,\Gamma_n^{\mathrm{j}}\,|$ $j = \mathrm{I,II}$ i.e. the *order* of Γ_n^{j}.

Lemma 1. *Let $(\mathcal{C}_n)_n$ be a sequence of configuration spaces such that $c := p_n\,\gamma_n$ is a positive constant. Then for the random graphs Γ_n^{I} holds*

$$\lim_{n\to\infty} \boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{\omega}_n^{\mathrm{I}} = k\,\} \;=\; \lim_{n\to\infty} \mathcal{B}_n\left(\frac{c\;k}{2(1-e^{-c})}, c\,\gamma_n\right)$$

and in particular $\lim_{n\to\infty}\mathbf{E}[\hat{\omega}_n^{\mathrm{I}}]\,/\,\lim_{n\to\infty}(1-e^{-c})\,|\,\mathcal{C}_n\,| = 1$.
For random graphs $\Gamma_n^{\mathrm{II}} \in \Omega^{\mathrm{II}}$ we have

$$\boldsymbol{\mu}_{n,\lambda}\{\hat{\omega}_n^{\mathrm{II}} = k\,\} \;=\; B(k, |\,\mathcal{C}_n\,|, \lambda)\ ,$$

$\lim_{n\to\infty}\mathbf{E}[\hat{\omega}_n^{\mathrm{II}}]\,/\,\lim_{n\to\infty}\lambda\,|\,\mathcal{C}_n\,| = 1$ *and* $\lim_{n\to\infty}\boldsymbol{\mu}_{n,\lambda}\{|\,\Gamma_n^{\mathrm{II}}\,| = k\} = \lim_{n\to\infty}\mathcal{B}_n(k,\lambda)$.

Proof. Model I: First we observe $\boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{\omega}_n = k\} = \boldsymbol{\mu}_{p_n}\{G^*\,|\,|\,G^*\,| = k\}$; thus it suffices to determine the distribution of orders of the graphs G^* (equ. 4).
For $v \in \mathrm{v}[G^*]$, let δ_v^* be its vertex degree. Next we have $\sum_{j=1}^{\gamma_n} j\,\boldsymbol{\mu}\{\,\delta_v^* = j\,\} \sim p_n\,\gamma_n$ implying

$$\lim_{n\to\infty}\mathbf{E}[\delta_v^*] \;=\; \frac{c}{1-e^{-c}}\,. \tag{6}$$

$\hat{X}_v$ and $\hat{X}_{v'}$ are asymptotically independent random variables and therefore the vertex degrees $\delta_v = \delta_v^*$ and $\delta_{v'} = \delta_{v'}^*$ are pairwise asymptotically uncorrelated (and equally distributed) random variables.

$$\mathbf{Cov}(\hat{\delta}_v\,\hat{\delta}_{v'}) = \mathbf{E}[\hat{\delta}_v - \mathbf{E}[\hat{\delta}_v]\,]\;\mathbf{E}[\hat{\delta}_{v'} - \mathbf{E}[\hat{\delta}_{v'}]\,] = \mathbf{E}[\hat{\delta}_v\,\hat{\delta}_{v'}\,] - \mathbf{E}[\hat{\delta}_v]\;\mathbf{E}[\hat{\delta}_{v'}] = \theta_n;\quad \lim_{n\to\infty}\theta_n = 0.$$

The variance of $\hat{X}(G^*) := \frac{1}{|\,G^*\,|}\sum_{v\in\mathrm{v}[G_p]}\hat{\delta}_v$ fulfills

$$\lim_{n\to\infty}\mathbf{V}[\hat{X}] \;=\; \lim_{n\to\infty}\frac{1}{|\,G^*\,|}\mathbf{V}[\hat{\delta}] \;+\; \lim_{n\to\infty}\frac{1}{|\,G^*\,|^2}(|\,G^*\,|\,\gamma_n\,\theta_n)$$

and we inspect for arbitrary natural number k:

$$\lim_{n\to\infty}\boldsymbol{\mu}_{p_n}\{|\,\mathrm{v}[G^*]\,|^{-1}\,\gamma_n\,\theta_n \le 1/k\} = 1\,.$$

Using the basic relation $\sum_{v\in\mathrm{v}[G^*]}\hat{\delta}_v = 2\,\mathrm{e}[G^*]$ we obtain $\lim_{n\to\infty}|\,\mathrm{v}[G^*]\,|\,\mathbf{E}[\hat{\delta}_v]\,/\,\lim_{n\to\infty}2\,|\,\mathrm{e}[G^*]\,| = 1$ whence

$$\lim_{n\to\infty}\boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{\omega}_n = k\} \;=\; \lim_{n\to\infty}\boldsymbol{\mu}_{p_n}\{|\,\mathrm{e}[G^*]\,| = k\,\frac{c}{2\lambda}\} \;=\; \lim_{n\to\infty}\mathcal{B}_n(k\,\frac{c}{2\lambda}, p_n)\,.$$

Model II: The proof for model II is a simple application of the Moivre–Laplace Theorem, see appendix A. ∎

Lemma 2. *Let* $(\mathcal{C}_n)_n$ *be a sequence of configuration spaces such that* $p_n\,\gamma_n = c$ *is a positive constant and let* $v, v' \in \mathrm{v}[\mathcal{C}_n]$ *with* $d(v,v') = 1$*. Then for* $v \in \mathrm{v}[\Gamma_n^{\mathrm{I}}]$ *the vertex degrees* δ_v *fulfilling* $\liminf \delta_v \,|\,\partial\{v\} \cap \partial\{v'\}\,|^{-1} = \infty$ *are asymptotically Gaussian distributed:*

$$\text{Model I:} \quad \lim_{n\to\infty} \boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{\delta}_v = \ell\} = \lim_{n\to\infty} \mathcal{B}_n\left(\ell\,\frac{c}{2(1-e^{-c})}, p_n\right) .$$

For the vertex degrees $\hat{\delta}_v$ *of a random graph* Γ_n^{II} *we obtain*

$$\text{Model II:} \quad \boldsymbol{\mu}_{n,\lambda}\{\hat{\delta}_v = \ell\} = B(\ell, \gamma_n, \lambda)$$

and in particular $\lim_{n\to\infty} \boldsymbol{\mu}_{n,\lambda}\{\hat{\delta}_v = \ell\} = \lim_{n\to\infty} \mathcal{B}_n(\ell,\lambda)$*.*

Proof. Model I: For a vertex $v \in \mathrm{v}[\mathcal{C}_n]$ we consider

$$Y_v := \{e \in \mathrm{e}[\mathcal{C}_n] \,|\, \tilde{\imath}(e) \in \partial\{v\} \,\vee\, \tilde{\tau}(e) \in \partial\{v\}\,\} .$$

We select each edge of Y_v with the probability $p_n = c/\gamma_n$. For a vertex $v' \in \partial\{v\}$ incident to a selected edge the number of all incident selected edges is a random variable with expectation value $c/[1-(1-p_n)^{\gamma_n}]$. As shown in lemma 1 the random variables corresponding to pairs of vertices $v', v'' \in \partial\{v\}$ are pairwise asymptotically uncorrelated. By assumption on

$$V_v := \{v' \in \partial\{v\} \mid v' \text{ is incident to a selected edge}\,\}$$

we have $\lim_{n\to\infty} |\,V_v\,|\,|\,\partial\{v\} \cap \partial\{v'\}\,|^{-1} = \infty$. We consider the induced subgraph of V_v in G^* and obtain in complete analogy to the argument in lemma 1 that the average number of incident edges becomes asymptotical constant for $G^*[V_v]$. Then $\lim_{n\to\infty} |\,V_v\,|\,\frac{c}{[1-(1-p_n)^{\gamma_n}]} / \lim_{n\to\infty} 2\,|\,Y_v\,| = 1$ implies

$$\lim_{n\to\infty} \boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{\delta} = \ell\} = \lim_{n\to\infty} \boldsymbol{\mu}_{p_n}\{|\,\mathrm{e}[Y_v]\,| = \ell\,\frac{c}{2\lambda}\} = \lim_{n\to\infty} \mathcal{B}_n(\ell\,\frac{c}{2\lambda}, p_n)$$

and the first statement follows.

Model II: The proof is obvious. ∎

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

$$\hat{X}_{n,k}^{\mathrm{j}} := |\,\{v \in \mathrm{v}[\Gamma_n^{\mathrm{j}}] \,|\, \hat{\delta}_v = k\}\,| \qquad \mathrm{j} = \mathrm{I},\mathrm{II}, \tag{7}$$

counting the number of vertices with degree k in a random graph Γ_n^{j}.

Proposition 1. *Let $(\mathcal{C}_n)_n$ be a sequence of configuration spaces such that $c := p_n\,\gamma_n$ is a positive constant and $v, v' \in \mathrm{v}[\mathcal{C}_n]$ with $d(v,v') = 1$. For $k = k_n$ such that* $\liminf k\,|\,\partial\{v\} \cap \partial\{v'\}\,|^{-1} = \infty$ *and* $\liminf \mathcal{B}_n(k\,\frac{c}{2(1-e^{-c})},p_n)\,\mathbf{E}[\hat{\omega}_n^{\mathrm{I}}] = \infty$ *we obtain*

$$\lim_{n\to\infty}\boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{X}^{\mathrm{I}}_{n,k} = \mathcal{B}_n(k\,\frac{c}{2(1-e^{-c})},p_n)\,\lambda\,|\,\mathcal{C}_n\,|\,\} = 1\,.$$

For the random graphs $\Gamma_n^{\mathrm{II}} < \mathcal{C}_n$ and arbitrary k holds

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

Proof. The previous lemma provides under the above assumption on k for model I: $\lim_{n\to\infty}\boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{\delta}_v = k\} = \lim_{n\to\infty}\mathcal{B}_n(k\,\frac{c}{2(1-e^{-c})},p_n)$ and a corresponding expression for model II. We shall prove the statement only for model I remarking that the proof for model II is completely analogous and shall omit the index "I" for the corresponding random variables.
We first observe $\mathbf{E}[\hat{X}_{n,k}] = \sum_{\ell=1}^{|\,\mathcal{C}_n\,|}\boldsymbol{\mu}\{\,\hat{\delta}_v = k\,\}\,\ell\,\boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{\omega}_n = \ell\}$ whence

$$\lim_{n\to\infty}\mathbf{E}[\hat{X}_{n,k}]\,\left[\lim_{n\to\infty}\mathcal{B}_n(k\,\frac{c}{2(1-e^{-c})},p_n)\,\mathbf{E}[\hat{\omega}_n]\right]^{-1} = 1\,.$$

We next compute the variance of $\hat{X}_{n,k}$ by evaluating the second factorial moment $\mathbf{E}[\hat{X}_k]_2$. Property (P) of definition 1 implies that the probability for selecting a pair of vertices with degree k is asymptotically given by $\mathcal{B}_n(k\,\frac{c}{2(1-e^{-c})},p_n)^2$, therefore

$$\lim_{n\to\infty}\mathbf{E}[\hat{X}_{n,k}]_2\,\left[\lim_{n\to\infty}\sum_{\ell=2}^{|\,\mathcal{C}_n\,|}\ell\,(\ell-1)\,\mathcal{B}_n(k\,\frac{c}{2(1-e^{-c})},p_n)^2\,\boldsymbol{\mu}^{n,\mathrm{I}}\{\hat{\omega}_n = \ell\}\right]^{-1} = 1\,,$$

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 $\lim_{n\to\infty}\mathbf{E}[\hat{X}_{n,k}]_2/\lim_{n\to\infty}(\mathcal{B}_n(k\,\frac{c}{2(1-e^{-c})},p_n)\,\mathbf{E}[\hat{\omega}_n]\,)^2 = 1$ and since $\mathbf{E}[\hat{X}^2_{n,k}] = \mathbf{E}[\hat{X}_{n,k}]_2 + \mathbf{E}[\hat{X}_{n,k}]$ we end up with

$$\lim_{n\to\infty}\mathbf{V}[\hat{X}_{n,k}]\,/\,\lim_{n\to\infty}(\mathbf{E}[\hat{X}_{n,k}] + [\theta_n\,\mathbf{E}[\hat{X}_{n,k}]\,]^2) = 1 \quad \text{where } \lim_{n\to\infty}\theta_n = 0.$$

By assumption we have $\liminf \mathbf{E}[\hat{X}_{n,k}] = \infty$ whence for $\epsilon > 0$ holds

$$\lim_{n\to\infty}\boldsymbol{\mu}^{n,\mathrm{I}}\{\,|\hat{X}_{n,k}/\mathbf{E}[\hat{X}_{n,k}] - 1\,| > \epsilon\} = 0\,,$$

thus the proposition. ∎

3.3. Density

In this section we shall suppose that a family of configuration spaces $(\mathcal{C}_n)$ is given. We restrict ourselves to the sequence of probability spaces (Ω_n^{II}). In other words for each graph $\mathcal{C}_n$ we consider 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 and write for short $\Omega_n = \Omega_n^{\mathrm{II}}$ and $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n,\lambda}$.

Definition 2. *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 in this section the *density property* of random graphs $\Gamma_n < \mathcal{C}_n$ where $0 < \lambda < 1$ and establish the existence of a "critical" λ-value, λ^* that has the following property:

- for $\lambda < \lambda^*$ a.a.s. no random graph Γ_n is dense and
- for $\lambda > \lambda^*$ a.a.s. every random graph Γ_n is dense.

We shall call λ^* the *threshold value for the density property.* For this purpose we consider the random variable

$$\hat{Z}_n(\Gamma_n) := |\,\{v \in \mathrm{v}[\mathcal{C}_n] \mid v \notin \overline{\mathrm{v}[\Gamma_n]}\}\,| \tag{8}$$

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

Lemma 3. *Let $(\mathcal{C}_n)_n$ be a family of configuration spaces. Suppose that*

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

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

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\hat{Z}_n = \ell\} = \frac{\mu^\ell}{\ell!}\,e^{-\mu}\ . \tag{9}$$

In particular we have

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

Finally, 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 11 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 independent probability λ and r is fixed we obtain

$$(1-\lambda)^{\gamma_n\,r} \sum_{\ell=0}^{|\,\mathcal{C}_n\,|} [[|\,\mathcal{C}_n\,| - |\,\Gamma_n\,|]_r]\,\boldsymbol{\mu}_n\{|\,\Gamma_n\,| = \ell\} \,\leq \mathbf{E}[\hat{Z}_n]_{\,r} \,\leq$$

$$\sum_{\ell=0}^{|\,\mathcal{C}_n\,|} \left[[|\,\mathcal{C}_n\,| - |\,\Gamma_n\,|]_r\,(1-\lambda)^{\gamma_n\,r} + \left\{(|\,\mathcal{C}_n\,| - |\,\Gamma_n\,|)^{r-1}\,r\,\gamma_n\,\right\}\,(1-\lambda)^{(\gamma_n - r)\,r}\right]\,\boldsymbol{\mu}_n\{|\,\Gamma_n\,| = \ell\}\,.$$

According to lemma 2 we have $\mathbf{E}[\hat{\omega}_n] \,= \lambda\,|\,\mathcal{C}_n\,|$ and $\mathbf{V}[\hat{\omega}_n] \,= \lambda\,(1-\lambda)\,|\,\mathcal{C}_n\,|$ whence $\lim_{n\to\infty} \boldsymbol{\mu}_n\{|\,\hat{\omega}_n/\lambda|\,\mathcal{C}_n\,| - 1\,| \,> \epsilon\,\} = 0$ and the above inequality reads

$$\lim_{n\to\infty}\left[1 - \frac{r^2}{(1-\lambda)\,|\,\mathcal{C}_n\,|}\right]\,[(1-\lambda)\,|\,\mathcal{C}_n\,|\,(1-\lambda)^{\gamma_n}]^r \leq \lim_{n\to\infty} \mathbf{E}[\hat{Z}_n]_{\,r}$$

$$\leq \lim_{n\to\infty}\,[(1-\lambda)\,|\,\mathcal{C}_n\,|\,(1-\lambda)^{\gamma_n}]^r\,\left[1 + |\,\mathcal{C}_n\,|^{-1}\,r\,\gamma_n(1-\lambda)^{-r^2}\right]\,,$$

using $[(1-\lambda)\,|\,\mathcal{C}_n\,|]^r - \,r\,[(1-\lambda)\,|\,\mathcal{C}_n\,|]^{r-1}\,r \leq [(1-\lambda)\,|\,\mathcal{C}_n\,| - r]^r \leq [(1-\lambda)\,|\,\mathcal{C}_n\,|]_r$. This proves the first statement:

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

For $\mu = \infty$ the above argument shows that

$$\mathbf{E}[\hat{Z}_n] \,\nearrow \infty \quad \text{and} \quad \lim_{n\to\infty} \mathbf{E}[\hat{Z}_n]_{\,2}/\lim_{n\to\infty} \mathbf{E}[\hat{Z}_n]^{\,2} = 1\,.$$

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 $\lim_{n\to\infty} \mathbf{V}[\hat{Z}_n]\,/\lim_{n\to\infty} \theta_n\,\mathbf{E}[\hat{Z}_n]^{\,2} = 1$ where $\lim_{n\to\infty} \theta_n \,= 0$. 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 such that* $\lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$ *exists and*

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

Let $\lambda^* := \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$ *then for* $\lambda > \lambda^*$ *holds*

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

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

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

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 3 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 3 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 3 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. $\mathcal{C}_n = \mathcal{Q}_\alpha^n$, the formula for the density threshold reads

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

3.4. Connectivity and the Sequence of Components

3.4.1. *Definitions and two Auxiliary Lemmata*

We assume in this section that for each $\mathcal{C}_n$ the limes $\lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$ exists and fulfills $0 < \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$. Further we set $\lambda^* := \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$. 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. 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 $V = \emptyset$, $G[\emptyset]$ 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 3. *Given a graph G, the sequence of components of G is the ordered tuple $(|\mathcal{X}_i|)_{1\le i\le|G|}$, where each $\mathcal{X}_i$ is a component of G and $|\mathcal{X}_i| \ge |\mathcal{X}_{i+1}|$. We call a component $\mathcal{X} < G$ a giant component if and only if $|\mathcal{X}| \ge 2/3\,|G|$.*

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 that have 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.

Lemma 4. *Suppose $k \in \mathbb{N}$, $v \in \mathrm{v}[\mathcal{C}_n]$ and $0 \le \lambda \le 1$. Then we have for $\lambda > \lambda^*$*

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

and for $\lambda < \lambda^*$:

$$\forall\, \ell \in \mathbb{N}: \quad \lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \textit{ contains at least } \ell \textit{ components with } 1 \le |\mathcal{X}| \le \gamma_n\} = 1\,.$$

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

$$\hat{I}_n(\Gamma_n) := |\{v \in \mathrm{v}[\Gamma_n] \mid \partial\{v\} \cap \mathrm{v}[\Gamma_n] = \emptyset\}|\,.$$

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

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

Further we inspect for $\lambda < \lambda^*$: $\lim_{n\to\infty} \mathbf{E}[\hat{I}_n]_2 / \lim_{n\to\infty} \mathbf{E}[\hat{I}_n]^2 = 1$ and thus $\lim_{n\to\infty} \mathbf{V}[\hat{I}_n] \,/\, \lim_{n\to\infty} \theta_n\, \mathbf{E}[\hat{I}_n]^2 = 1$ where $\lim_{n\to\infty} \theta_n = 0$. Therefore we end up with $\lim_{n\to\infty} \boldsymbol{\mu}_n\{|\hat{I}_n/\mathbf{E}[\hat{I}_n] - 1| > \epsilon\} = 0$ and obtain for $\lambda < \lambda^*$ and $\ell \in \mathbb{N}$

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

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

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

and there exists a $b \in \mathbb{R}_+$ such that $|\{X \mid \mathcal{C}_n[X] \text{ is connected}\}| \le e^{b\ln(\gamma_n)\,|\mathcal{C}_n[X]|}$.

Suppose now $|X| \le \gamma_n$ and that $\mathcal{C}_n[X]$ 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 for a component 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 its $\mathcal{C}_n$-boundary ∂X. Suppose $|X| \le \gamma_n$ then we inspect using property (Q) of definition 1

$$\exists\, a \in R_+ : \qquad a\gamma_n|X| \le |\partial X|\,.$$

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

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

and the proof of the lemma is complete. ∎

We proceed investigating the connectivity property of the random graphs $\Gamma_n < \mathcal{Q}^n_\alpha$. Basically we shall establish that the parameter $\lambda^* = \lim_{n\to\infty} 1 - |\mathcal{C}_n|^{-\frac{1}{\gamma_n}}$ is not only a "critical" parameter for the density but also a "critical" parameter for the connectivity property. In other words we shall show that λ^* is also a threshold value for connectivity.

According to lemma 4 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 be proven that $\lambda > \lambda^*$ implies

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

The first step is

Lemma 5. *Let* $\Gamma_n < \mathcal{C}_n$ *be a random graph* $\Gamma_n \in \Omega_n$ *and* $\lim_{n\to\infty} 1 - |\mathcal{C}_n|^{-1\frac{1}{\gamma_n}} < \lambda$. *Then*

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

Proof. For $k \in \mathbb{N}$ we consider again the random variable $\hat{X}_{n,k}$ on Ω_n the states of which are the numbers of vertices $v \in \mathrm{v}[\Gamma_n]$ with $\delta_v = k$. An upper bound for $\mathbf{E}[\hat{X}_{n,k}]$ 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. We further inspect that λ^* solves for x:

$$\forall k \in \mathbb{N}: \quad \lim_{n\to\infty}\left[|\,\mathcal{C}_n\,|\binom{\gamma_n}{k}\,x^k\,(1-x)^{\gamma_n-k}\right] = a_k \quad a_k \in \mathbb{R}_+\,.$$

For $\lambda > \lambda^*$ the above equation implies

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

proving

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

and the lemma follows. ∎

Using precisely the same argument we further obtain

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

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

In order to make the following discussion more transparent, we first analyze connectedness for a special class of configuration spaces namely the class of general hypercubes (Hamming graphs) (see p. 16).

3.4.2. Connectedness in Generalized Hypercubes

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

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

Assuming that $\boldsymbol{\mu}_{n,\lambda}$ and $\mathcal{Q}_\alpha^n$ are given we speak of *random graphs* Γ_n and (as usual) we write for short $\Omega_n = \Omega_n^{\mathrm{II}}$ and $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n,\lambda}$.

We recall that a generalized hypercube $\mathcal{Q}_\alpha^n$ is a $\gamma_n := [(\alpha - 1)\, n]$-regular graph and observe

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

(see the example p. 16).

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

$$B_r(v) := \{v' \,|\, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n] \wedge d(v,v') = r\}\ .$$

First we estimate loosely speaking "how many independent paths exists in $\mathcal{Q}_\alpha^n$ that connect the balls $B_2(v), B_2(v')$ where v, v' are assumed to have finite distance k".

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

$$v = (x_1, x_2, \ldots, x_n) \quad \text{and} \quad v' = (x_1', x_2', \ldots, x_k', x_{k+1}, ..., x_n)\,. \tag{10}$$

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

$$f_j(v_2) := (x_1, ..., x_j, x_{j+1}', ..., x_k', x_{k+1}, ..., x_{r-1}, \hat{x}_r, x_{r+1}, ..., x_{s-1}, \hat{x}_s, x_{s+1}, ...)\,. \tag{11}$$

Clearly $f_k(v_2) = v_2$, $f_0(v_2) \in B_2(v') \cap B_{2+k}(v)$ and each family $(f_j(v_2))_{0 \le j \le k}$ can be identified with a path of length k in $\mathcal{Q}_\alpha^n$, $\pi(v_2)$. 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

$$\begin{array}{rcc} \psi_{n,k}^{v,v'} : B_2(v) \cap B_{2+k}(v') & \longrightarrow & \Pi(\mathcal{C}_n) \\ v_2 & \mapsto & \psi_{n,k}^{v,v'} := \pi(v_2)\,. \end{array}$$

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

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

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 an order of n^2 pairwise disjoint paths in Γ_n "connecting" $B_2(v)$ and $B_2(v')$.

Lemma 6. *Let $\mathcal{Q}_\alpha^n$ be a generalized hypercube and $\Gamma_n < \mathcal{Q}_\alpha^n$ a random graph with underlying probability measure $\boldsymbol{\mu}_n = \boldsymbol{\mu}_{n,\lambda}$ and $0 < \lambda \leq 1$. Then for arbitrary $\chi \in I\!R_+$ holds*

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

Proof. We consider the random variable $\hat{Z}_{n,k}^{v,v'}$. For each $v_2 \in B_2(v) \cap B_{2+k}(v')$ there corresponds the path $\pi(v_2)$ obtained according to equation (11) in $\mathcal{Q}_\alpha^n$. 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 $\hat{Z}_{n,k}^{v,v'}$, counting the number of those paths is binomial distributed with expectation value

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

By applying corollary 11 in the appendix we obtain

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

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

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

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

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

proving the lemma. ∎

We now proceed by showing 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 $\mathcal{Q}_\alpha^n$ which we introduce now:

For $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ with $d(v,v') = k$ we assume v, v' to be given by equation (10). For $v_1 \in \partial\{v\} \cap B_{1+k}(v')$ we set

$$g_j(v_1) := (x_1, ..., x_j, x'_{j+1}, ..., x'_k, x_{k+1}, .., \hat{x}_r, ..., x_n) \qquad 0 \leq j \leq k \;\; \hat{x}_r \neq x_r \tag{12}$$

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

$$\begin{array}{rcl} \partial\{v\} \cap B_{k+1}(v') & \longrightarrow & \Pi(\mathcal{Q}_\alpha^n) \\ v_1 & \mapsto & \pi(v_1) := (g_k(v_1), g_{k-1}(v_1), ..., g_1(v_1), g_0(v_1)) \,. \end{array}$$

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

$$M_{n,k}^{v,v'}(\Gamma_n) := \{\pi(v_1) \,|\, \pi(v_1) \in \Pi(\Gamma_n)\,\} \,.$$

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

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

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

We now introduce the random variable

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

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

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

Remark. The following lemma is not simply implied by corollar 11 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 7. *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*

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

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

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n | \forall\, v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n],\, \exists\, v_1^{(1)}, ..., v_1^{(\ell)} \in \partial\{v\} \cap \mathrm{v}[\Gamma_n] \;\wedge\; \exists\, v_1'^{(1)}, ..., v_1'^{(\ell)} \in \partial\{v'\} \cap \mathrm{v}[\Gamma_n]\,\} = 1\,. \tag{13}$$

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 can apply corollary 11 (see appendix A) and obtain

$$\forall b' \in \mathbb{R}_+ \,\exists\, b \in \mathbb{R}_+ : \qquad \lim_{n\to\infty} \boldsymbol{\mu}_n\{\,|\,\hat{Y}_{n,k}^{w,w'} - \mathbf{E}[\hat{Y}_{n,k}^{w,w'}]\,|\geq b'\,n\} \leq e^{-b\,n}\,. \tag{14}$$

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 (13) a.a.s. for each pair $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ and arbitrary $\ell \in \mathbb{N}$ there are a. s. 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 $v_1'^{(i)} \in B_1(v') \cap \mathrm{v}[\Gamma_n]$. I. e. there exists a.a.s. nontrivial random variables $\hat{Y}_{n,k}^{v_1^{(i)}, v_1'^{(i)}}$, for $1 \leq i \leq \ell$.

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 $\partial\{v\} \nearrow \infty$ and we can assume 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

$$|\,\partial\{v\} \cap \partial\{w\}\,| \leq \alpha \tag{15}$$

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 $v_1'^{(i)} = (x_1'^{(i)}, ..., x_k'^{(i)}, x_{k+1}'^{(i)}, x_{k+2}'^{(i)}, x_{k+3}^{(i)}, ..., x_n^{(i)})$. Then to each pair $(v_1^{(i)}, v_1'^{(i)})$ there corresponds the path

$$(g_j(v_1^{(i)})_{0\leq j\leq k+2}), \quad g_j(v_1^{(i)}) := (x_1^{(i)}, ..., x_j^{(i)}, x_{j+1}'^{(i)}, ..., x_n^{(i)})\,. \tag{16}$$

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 (14) 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 a.a.s. 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. ▮

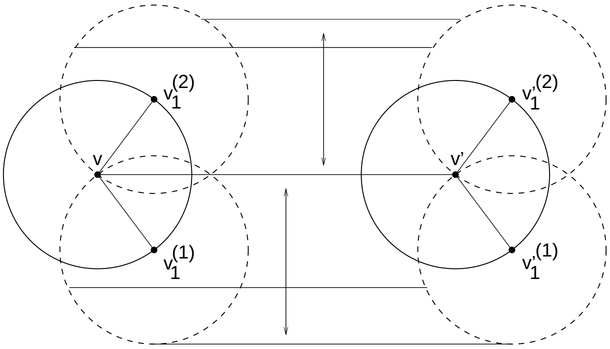


Figure 4: An illustration for the proof of lemma 7. For given $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ each pair of vertices $(v_1^{(i)}, v_1'^{(i)})$ leads to "sufficiently many" independent pairwise disjoint paths in $\Pi(\mathcal{Q}_\alpha^n)$.

Corollary 2. *Let k be a fixed natural number and $\hat{\Lambda}_{n,k}$ the random variable:*

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

Then under the assumptions of lemma 7 holds $\lim_{n\to\infty} \boldsymbol{\mu}_n\{\,\Gamma_n \,|\, \hat{\Lambda}_{n,k} = 1\,\} = 1$.

Proof. According to equation (13) 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 in the a.a.s. for each pair $v, v \in \mathrm{v}[\Gamma_n]$ there exist $v_1 \in \partial\{v\} \cap \mathrm{v}[\Gamma_n], v_1' \in \partial\{v'\} \cap \mathrm{v}[\Gamma]$ with the property

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

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. ∎

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 $\Gamma_n < \mathcal{Q}_\alpha^n$ random induced subgraphs. Then*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is connected}\} = \begin{cases} 1 & \text{for} \quad \lambda > 1 - \sqrt[\alpha-1]{\alpha^{-1}} \\ 0 & \text{for} \quad \lambda < 1 - \sqrt[\alpha-1]{\alpha^{-1}}\,. \end{cases} \tag{17}$$

I. e. λ^* is a threshold value for the connectivity property.

Proof. The existence of components $\mathcal{X}$ with $|\,\mathcal{X}\,| \le \gamma_n$ with has been investigated in lemma 4, 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 connectivity of $\mathcal{Q}_\alpha^n$ guarantees that there exists a path (in $\Pi(\mathcal{Q}_\alpha^n)$) say $(v_i)_{0\le i\le d(v,v')}$ 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 $0 \le i \le d(v, v')$ and introduce

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

Claim:

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

We only have to observe that $\hat{Z}_{n,j}^{v,v'}$ is binomially distributed and $\mathbf{E}[\hat{Z}_{n,j}^{v,v'}] = \binom{n}{2}\,(\alpha-1)^2\,\lambda$. Then we apply corollary 11 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 $B_2(v_i)$, $0 \le i \le d(v,v')$ we simultaneously find vertices of a random graph with probability one. 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 a.a.s. pairwise finite distance (in $\mathcal{Q}_\alpha^n$) and are according to corollary 2 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 $(v_j')_{0\le j\le d(v,v')}$ 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 $\mathrm{v}[A_p] := \mathrm{v}[\mathcal{Q}_2^n]$ 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 [26, 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.

3.4.3. Connectivity and Giant Components in 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 that $\lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$ exists and $0 < \lambda^* = \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$.

For convenience of the reader we first recall property (R) from definition 1:

(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

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

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

Remark. We remark that according to property (R) (ii) that each path $\pi \in \mathbf{P}_{\mathcal{C}_n}^{v,v',m}$ leads to a paths in $\mathcal{C}_n$ in which v and v' occur.

We introduce, generalizing our proceeding in the special case of generalized hypercubes, for $v, v' \in \mathrm{v}[\mathcal{C}_n]$ with $d(v,v') = k$, $k \in \mathbb{N}$ the random variable

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

Note that there are two fact implying that $\hat{Y}_{n,k}^{v,v'} \equiv 0$, namely (i) that $v \,\vee\, v' \not\in \mathrm{v}[\Gamma_n]$ or (ii) that there exists no path $\pi \in \Pi(\Gamma_n)$ in which both vertices occur.

Property (R) (iii) guarantees that for $v, v' \in \mathrm{v}[\Gamma_n]$ $\hat{Y}_n^{v,v'}$ is in the limit Gaussian (using the Moivre Laplace theorem) [2] since each path $\pi \in \mathbf{P}_{\mathcal{C}_n}^{v,v',m}$ 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 7 we state:

Lemma 8. *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 $\lambda > \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}$. Then we have*

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

Proof. The proof is completely analogous to that of lemma 7. We first observe that corollary 1 implies for arbitrary $\ell \in \mathbb{N}$

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n | \forall v, v' \in \mathrm{v}[\mathcal{C}_n] : \exists v_1^{(1)}, ..., v_1^{(\ell)} \in \partial\{v\} \cap \mathrm{v}[\Gamma_n] \wedge \exists\, v_1'^{(1)}, ..., v_1'^{(\ell)} \in \partial\{v'\} \cap \mathrm{v}[\Gamma_n]\,\} = 1\,. \tag{18}$$

Suppose $w, w' \in \mathrm{v}[\Gamma]$ then $\hat{Y}_{n,k}^{w,w'}$ is asymptotically Gaussian and we have according to corollary 11:

$$\exists\, b \in \mathbb{R}_+ :\ \lim_{n\to\infty} \boldsymbol{\mu}_n\{\, \hat{Y}_{n,k}^{w,w'} = 0\,\} \le e^{-b\,\gamma_n}\,. \tag{19}$$

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 $v_1^{(i)} \in \partial\{v\} \cap \mathrm{v}[\Gamma_n], v_1'^{(i)} \in \partial\{v'\} \cap \mathrm{v}[\Gamma_n]$, $1 \le i \le \ell$ fulfill

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

Then we claim that there exists a $b \in R_+$ such that

$$\lim_{n\to\infty} P \le \lim_{n\to\infty} e^{-\ell\, b\, \gamma_n}\,.$$

Equation (19) implies that a.a.s. for each pair $v, v' \in \mathrm{v}[\mathcal{C}_n]$ there exist any finite number of vertices of $\mathrm{v}[\Gamma_n]$ that are adjacent to v and v' respectively. In other words there exist sets Φ_v and $\Phi_{v'}$ that fulfill condition (R) (iv).
According to (R) (i)-(iv) there exist ℓ pairs of vertices $(v_1^{(i)}, v_1'^{(i)})$ such that:

- the random variables $\hat{Y}^{v_1^{(i)}, v_1'^{(i)}}_{n, d(v_1^{(i)}, v_1'^{(i)})}$ are all nontrivial.
- the pairs $(v_1^{(1)}, v_1'^{(1)}), ..., (v_1^{(\ell)}, v_1'^{(\ell)})$ induce paths

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

  fulfilling (R) (iii) and in particular $k - 2 \le \ell(\pi) \le k + 2 + m$.
- for each pair $(v_1^{(i)}, v_1'^{(i)})$ we have

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

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

Form the above properties follows that to each pair $(v_1^{(i)}, v_1'^{(i)})$ there corresponds asymptotically at least $\gamma_n - 2\,\ell\,\theta_n\,\gamma_n$, $\lim_{n\to\infty}\theta_n = 0$, paths that are (see (R) (iii)) pairwise disjoint. Then the probability that none of these is a path in Γ_n has the upper bound

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

since $d(v_1^{(i)}, v_1'^{(i)}) \leq k+2$. We therefore obtain

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

The fact $\lim_{n\to\infty} 0 < 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$ and equation (18) guarantees that we can choose ℓ such that

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

From the above equation we obtain that a.a.s. at least one pair $(v_1^{(i)}, v_1'^{(i)})$ fulfills

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

and the proof of the lemma is complete. ∎

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

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

Then under the assumptions of lemma 8 we have $\lim_{n\to\infty}\boldsymbol{\mu}_n\{\,\Gamma_n \,|\, \hat{\Lambda}_{n,k} = 1\,\} = 1$.

Proof. According to equation (18) 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 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 random induced subgraphs Γ_n.

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

$$U_n(\Gamma_n) := \{v \in \mathrm{v}[\Gamma_n] \,|\, \delta_v \ge \ln(\gamma_n)\,\}\,. \tag{20}$$

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. *Let $(\mathcal{C}_n)$ be a sequence of configuration spaces such that $0 < \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$. Then*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is connected}\} = \begin{cases} 1 & \text{for} \quad \lambda > \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} \\ 0 & \text{for} \quad \lambda < \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}}\,. \end{cases} \tag{21}$$

Therefore λ^* is a threshold value for the connectivity property.

Proof. We essentially reorganize the proof of theorem 2. The case of small components follows from lemma 4 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 connectivity 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 $(v_j)_{0\le j\le d(v,v')}$ we consider the sets $B_2(v_j) \subset \mathrm{v}[\mathcal{C}_n]$ for $1 \le j \le d(v, v')$ and show in complete analogy to the proof of theorem 2 that

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \,|\, \forall\, 0 \le j \le d(v, v'),\, B_2(v_j) \cap \mathrm{v}[\Gamma_n] \ne \emptyset\,\} = 1\,. \tag{22}$$

Clearly, the probability that for one index $0 \le j \le d(v, v')$ 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 1 we have $\lim_{n\to\infty} |\,B_2(v_j)\,| = c\,\gamma_n^2$, $c \in \mathbb{R}_+$. Since $d(v, v') \le |\,\mathcal{C}_n\,|$ 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 (22) 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 a. s. to given $k \in \mathbb{N}$ there exists 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* $0 < \lim_{n\to\infty} 1 - |\,\mathcal{C}_n\,|^{-\frac{1}{\gamma_n}} < 1$ *and* $0 < \lambda \le 1$. *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 $\lambda > \lambda^*$ the statement is implied by theorem 3 i. e. :

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

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

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

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

$$\forall i \in \mathbb{N}: \quad \lim_{n\to\infty} \gamma_n^i\,\xi_n = 0\,. \tag{23}$$

We shall show that $\Gamma_n[U_n(\Gamma_n)]$ (see equation (20)) forms a giant component in Γ_n. For this purpose we first consider the case $v, v' \in U_n(\Gamma_n)$ and $\lim_{n\to\infty} d(v,v') < k$, $k \in \mathbb{N}$.

We can apply corollary 4 and obtain that a.a.s. in the random graphs Γ_n pairs of vertices $v, v' \in U_n(\Gamma_n)$ having finite distances occur in a path in Γ_n.

For arbitrary $v, v' \in U_n(\Gamma_n)$ the connectivity of $\mathcal{C}_n$ implies that there exists a path $(v_j)_{0\le j\le d(v,v')}$ in $\mathcal{C}_n$ in which v and v' occur. It remains to show that

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

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 $0 \le j \le d(v,v')$ 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 (23) and property (Q) of definition 1 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 induced subgraph of

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

in $\mathcal{C}_n$ i. e. $\mathcal{C}_n[U_n]$ is a giant component that fulfills $\mathcal{C}_n[U] < \mathcal{X}_1$ and the theorem follows. ∎

4. Neutral Networks of RNA Secondary Structures

4.1. RNA Secondary Structures and Compatible Sequences

In this chapter we assume a generalized hypercube $\mathcal{Q}_\alpha^n$ to be fixed. The elements of its vertex set $\mathrm{v}[\mathcal{Q}_\alpha^n]$ can be interpreted as *RNA molecules* or *sequences* of length n. 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 possibilities of present-day computers, secondary structures, which are defined as the list of base pairs in the molecules, are readily accessible. A large body of computational data has been published [17, 21, 18, 54, 4, 61] on this example of a sequence-structure mapping, allowing for a check of our theory.

The shape space consists of all secondary structure graphs as defined below. A variety of different algorithms [48, 71, 70, 46, 43], and different sets of thermodynamic parameters [50, 22, 63] have been used for the prediction of RNA secondary structures. Fortunately, it has been shown recently [62] that the qualitative features of the sequence-structure mappings are independent of algorithm and parameter set.

Definition 4. [64] *A secondary structure is a vertex-labeled graph on n vertices with an adjacency matrix $A = (a_{i,k})_{1\le i,k\le n}$ fulfilling*

(1) $a_{i,i+1} = 1$ *for* $1 \le i \le n-1$*;*

(2) For each i there is at most a single $k \ne i-1, i+1$ such that $a_{i,k} = 1$;

(3) If $a_{i,j} = a_{k,l} = 1$ and $i < k < j$ then $i < l < j$.

We call an edge (i,k), $|i-k| \ne 1$ 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 $\mathrm{p} := \beta/\alpha^2$, which is the probability that two arbitrarily chosen letters shall be 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 and that (3) implies that a secondary structure is a knotfree planar graph. 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 $[y, x] \in \Pi$ i.e. a symmetric relation. In the following we shall consider secondary structures over arbitrary alphabets with arbitrary pairing rules.

Definition 5. *Let s be a secondary structure (see def. 4 above) and*

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

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

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 fact we have the embedding $\mathbf{C} : \mathcal{S}_n \hookrightarrow \{U \subset \mathrm{v}[\mathcal{C}_n]\}$, $s \mapsto \mathbf{C}[s]$ (see also section 4.3).

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

Definition 6. *Let S_n be the symmetric group in n letters. We write a transposition $\tau \in S_n$ as $\tau = (i, k)$. Then*

$$\begin{array}{llll} \imath : & \mathcal{S}_n & \rightarrow & S_n \\ & s & \mapsto & \imath(s) \quad := \prod_{[i,k]\in\Pi(s)} (i, k)\,. \end{array}$$

The map $\imath$ is clearly 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 [57]. $\imath$ naturally gives rise to the mapping

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

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 Π be a nonempty pairing rule on $\mathcal{A}$ and s and s' be arbitrary (nonempty) secondary structures. Then we have w.r.t. Π*

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

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. The statement of theorem 5 does not hold true for 3 different structures.

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 $f_n : \mathcal{Q}^n_\alpha \to \mathcal{S}_n$. 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 $\mathcal{Q}^n_\alpha[\mathbf{C}[f_n(x)]]$ 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 7. *Let s be a secondary structure, then the graph of compatible sequences is*

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

Remark. Obviously $\mathcal{C}[s]$ has the vertex set $\mathbf{C}[s]$ and 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 $[i,j] \in s$.

[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. 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 8. *Suppose $f_n : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$ 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.,*

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

We shall now construct neutral networks as random graphs by means of a simple random process. More precisely we consider random 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 (as in model II of chapter 3). The fact that *a priori* there is no reason why the probability of being neutral neighbor should be the same for both single base and base pair mutations motivates:

Model III: *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 (see model II). Writing $\chi_{u,p} := \lambda_u + \lambda_p - \lambda_u \lambda_p$ we set*

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

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

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

Remark. We can intuitively construct a neutral network $\Gamma_n^{\mathrm{III}}[s]$ 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 equally–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 we proceed with the analysis of model III we introduce some terminology.

Definition 9. *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:*

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

We now deduce, by application of the theory developed in chapter 3 a sufficient criterion for the density and the connectivity property of random subgraphs $\Gamma_n^{\mathrm{III}}[s] < \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$.

The connection to the theory in chapter 3 is established as follows: For $(x,y) \in \mathrm{v}[\Gamma_n^{\mathrm{III}}[s]]$ either x or y have been chosen. Then for x has not been chosen we have $\Phi_x^{\Gamma_n^{\mathrm{III}}[s]} \cong \Gamma_{n_p}$ ($\Phi_x^{\Gamma_n^{\mathrm{III}}[s]} \cong \mathcal{Q}_\beta^{n_p}$ else) and for y has not been chosen $\Phi_y^{\Gamma_n^{\mathrm{III}[s]}} \cong \Gamma_{n_u}$ ($\Phi_y^{\Gamma_n^{\mathrm{III}[s]}} \cong \mathcal{Q}_\alpha^{n_u}$ else) where the underlying probability measures are given by

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

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

Theorem 6. *Let $\Gamma_n^{\mathrm{III}}[s] < \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ be a random graph constructed according to model III. Suppose that for all $(v_1, v_2) \in \mathrm{v}[\Gamma_n^{\mathrm{III}}[s]]$ holds:*

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

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

Proof. We write for short $\Gamma := \Gamma_n^{\mathrm{III}}[s]$. 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 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

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

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 11 of the appendix we observe that a.a.s. for all pairs $(v_1, v_2), (v_1', v_2')$ there exits an x such that $(v_1, x), (v_1', x) \in \mathrm{v}[\Gamma]$ are connected by a path of finite length of the form

$$(v_1, x), (w_1, x), ..., (w_1', x), (v_1', x)$$

since all fibers $\Phi_x^{\Gamma} \cong \Gamma_{n_u}$ are by assumption connected. Along these lines we further inspect that a.a.s. for all pairs $(v_1, v_2), (v_1', v_2')$ there exists a path of finite length of the form

$$(v_1', x), (v_1', y), ..., (v_1', z), (v_1', v_2')$$

since $\Phi_{v_1'}^{\Gamma} \cong \Gamma_{n_p}$ is also 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. path $\pi \in \Pi(\Gamma)$ in which both vertices occur. ∎

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

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

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

[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.

Proof. It remains to verify the condition

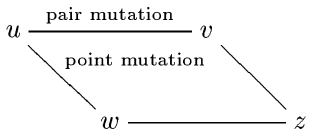
$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\,\Gamma_n^{\mathrm{III}}[s]\,|\,\forall(v_1,v_2)\in \mathrm{v}[\Gamma_n^{\mathrm{III}}[s]]:\ \Phi_{v_1}^{\Gamma_n^{\mathrm{III}}[s]},\ \Phi_{v_2}^{\Gamma_n^{\mathrm{III}}[s]} \text{ are dense and connected}\,\} = 1\,.$$

From the proof of lemma 7 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 $\Phi_{v_1}^{\Gamma_n^{\mathrm{III}}[s]}$ and accordingly β^n fibers $\Phi_{v_2}^{\Gamma_n^{\mathrm{III}}[s]}$, we can choose ℓ sufficiently large such that

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

Consequently the above equation holds if $\lambda_u > 1 - \sqrt[\alpha-1]{\alpha^{-1}}$, $\lambda_p > 1 - \sqrt[\beta-1]{\beta^{-1}}$ and the corollary follows. ∎

Before we proceed with the analysis of model III 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 III we have assumed that there is asymptotically no correlation between point mutations and pair mutations. Each fiber $\Phi_{v_2}^{\Gamma_n^{\mathrm{III}}[s]}$ is isomorphic to a random graph Γ_{n_u} and accordingly $\Phi_{v_1}^{\Gamma_n^{\mathrm{III}}[s]} \cong \Gamma_{n_p}$.

The other extreme is to consider these two types of mutations as completely correlated in the following sense:

If any three vertices in the parallelogram above are chosen, the fourth vertex has to be chosen as well.

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

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

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

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

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

We have $\Phi_{v_1}^{\Gamma_n^{\mathrm{IV}}[s]} \cong \Gamma_{n_p}$ and $\Phi_{v_2}^{\Gamma_n^{\mathrm{IV}}[s]} \cong \Gamma_{n_u}$ 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:

$$\begin{tikzcd}
 & \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p} & \\
 & \Gamma_{n_u} \times \Gamma_{n_p} \arrow[u, hook, "\iota"] & \\
\Gamma_{n_u} \arrow[ur, hook, "in"] & & \Gamma_{n_p} \arrow[ul, hook', "in"']
\end{tikzcd}$$

Theorem 6 has the following analogue for model IV:

Theorem 7. *Let $\Gamma_n^{\mathrm{IV}}[s] < \mathcal{C}[s]$ be a random subgraph such that the following holds:*

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

Then we have

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

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

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

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

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

Lemma 9. *For the orders of the random graphs $\Gamma_n^{\mathrm{III}}[s], \Gamma_n^{\mathrm{IV}}[s]$ we have for $N, N_u, N_p \in I\!N$:*

$$\boldsymbol{\mu}_n\{\Gamma_n^{\mathrm{III}}[s] = N\} = \quad B(N, \alpha^{n_u}\beta^{n_p}, [\lambda_u + \lambda_p - \lambda_u\,\lambda_p])$$

$$\boldsymbol{\mu}_n\{\Gamma_n^{\mathrm{IV}}[s] = N\} = \sum_{N_u\,N_p=N} B(N_u, \alpha^{n_u}, \lambda_u)\,B(N_p, \beta^{n_p}, \lambda_p)\,.$$

In particular the distributions of the orders become asymptotically Gaussian.

4.3. Shape Space Covering

The combination of a variety of computer simulations [21, 18, 54] 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

$$\mathrm{dist}(\Gamma_n[s_1], \Gamma_n[s_2]) := \min\{d(v_1, v_2)\,|\,v_1 \in \mathrm{v}[\Gamma_n[s_1]],\ v_2 \in \mathrm{v}[\Gamma_n[s_2]]\,\}\ .$$

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 $X = \mathrm{III}, \mathrm{IV}$. Suppose the corresponding neutral networks $\Gamma_n^{\mathrm{X}}[s_1]$ and $\Gamma_n^{\mathrm{X}}[s_2]$ are dense and connected. Then the following assertions hold*
(i) The minimum distance between the neutral networks $\Gamma_n^{\mathrm{X}}[s_1]$ and $\Gamma_n^{\mathrm{X}}[s_2]$ of any two secondary structures $s_1, s_2 \in \mathcal{S}_n$ is a.a.s. at most

$$\mathrm{dist}(\Gamma_n^{\mathrm{X}}[s_1], \Gamma_n^{\mathrm{X}}[s_2]) \le 4 \tag{24}$$

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

$$r_{s_1} = [1 - \frac{\beta}{\alpha^2}]\,n_p(s_1) + \theta_n, \quad \lim_{n\to\infty}\theta_n = 0\,.$$

Proof. (i) is an obvious consequence of theorem 5 and theorem 6.

(ii) 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 whence the theorem follows. ∎

4.4. Outlook: C*-Random Maps on Generalized Hypercubes

In this section we present a method to construct mappings $f : \mathcal{Q}_\alpha^n \longrightarrow \mathcal{S}_n$ using the random graph approach for neutral networks. Let M be a finite set. In the following we write $\mathcal{P}(M)$ for its *power set.*

Definition 10. *Let* $\mathbf{C}^* : \mathcal{S}_n \to \mathcal{P}(\mathrm{v}[\mathcal{Q}_\alpha^n])$ *and* $r : \mathcal{S}_n \to I\!N$ *be two mappings such that* $j \le i \implies r(s_j) \ge r(s_i)$.

A mapping $f : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$ *is called* $\mathbf{C}^*$-map *if and only if*

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

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

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

Remark. Clearly any RNA folding map is a $\mathbf{C}^*$-map if we set $\mathbf{C}^*[s] := \mathbf{C}[s]$ since the neutral networks are constructed *a priori* in the set of compatible sequences. In the sequel we shall assume that $\mathbf{C}^* = \mathbf{C}$.

We now restrict ourselves to the case $\Gamma_n[s] = \mathbf{C}[s]$ for $s \in \mathcal{S}_n$ and compute the distribution of the corresponding preimage sizes. In this situation the recursion formula reads

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

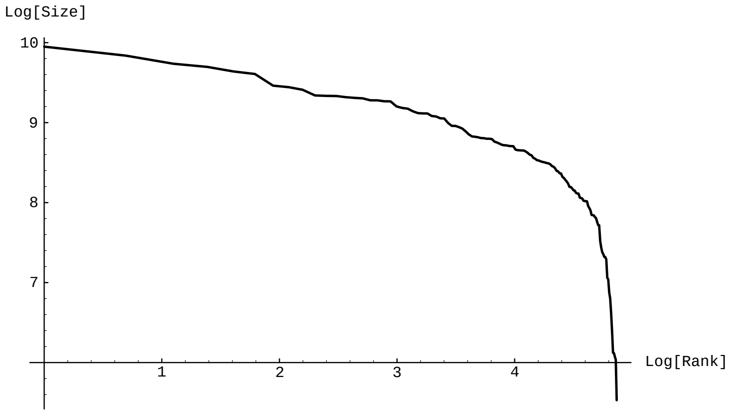


Figure 5: We report here the logarithm of the sizes of the neutral networks $f^{-1}(s)$ obtained by a $\mathbf{C}^*$-random map where the preimage are obtained from model IV (see section 2) with underlying λ parameter ($\lambda_u = \lambda_p$) equals 0.8 (see p. 48). The corresponding neutral networks are ordered on the x-axis by the logarithm of their orders. Note that the rank of the secondary structure s does not necessarily coincidence with the size of the corresponding preimage $|\, f^{-1}(s)\,|$.

We first set for fixed i and $j < i$

$$X_j := \mathbf{C}[s_{r_i}] \cap \mathbf{C}[s_{r_j}]\,.$$

According to the inclusion exclusion principle, observe immediately

$$|\bigcup_{j<i} X_j\,| = \sum_{j<i} |\,X_j\,| - 2 \sum_{j<j'<i} |\,X_j \cap X_{j'}\,| + 6 \sum_{j<j'<j''<i} |\,X_j \cap X_{j'} \cap X_{j''}\,| - \dots . \tag{25}$$

For alphabets $\mathcal{A}$ of length α having exactly complementary base pairs and $j < i$ we inspect from corollary 3 that $|\,X_j\,| = \alpha^{\Phi}$, where Φ is the number of orbits obtained from the action of $\langle \imath(s_i), \imath(s_j)\rangle$ on $\{1, ..., n\}$. By use of the pairing

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

we can view Φ as the outcome of an integer valued random variable $\hat{\Phi}$, assigning to each pair of involutions the number of orbits of the corresponding dihedral group. Therefore the first step is to determine the distribution of this random variable in order to obtain, using the inclusion-exclusion principle above, an analytical solution for the distribution of the sizes of the preimage $|\, f^{-1}(s)\,|$. Here we only report that our numerical calculations confirm that for *any rank order function and any λ parameter* the distribution of preimage sizes is given as in figure 5.

5. Error Thresholds of RNA Secondary Structures

5.1. The Mathematical Model

In this chapter we apply the mathematical modeling of neutral networks in order to verify one fundamental concept of theoretical biology, namely the existence of an error threshold and the formation of a molecular quasispecies. We shall study a finite *population* (this term shall be defined later) $\mathbf{V}$ of asexually replicating strings in a landscape induced by a neutral network $\Gamma_n[s]$ (see definition 11 below).

On the one hand we apply a birth-death process in order to create a mathematical model for the dynamics and on the other hand we simultaneously analyze the dynamics of $\mathbf{V}$ by computer simulations basing on the Gillespie algorithm [23].

In the sequel we shall restrict ourselves to the mathematical modeling and use the simulations without further discussion for comparisons. We remark that the validity of those comparisons is a standard assumption [47]. Omitting a detailed discussion of the simulations[5] we refer to appendix B. Let us begin by discussing the underlying landscape:

Suppose now that a neutral network $\Gamma_n[s]$, associated to a fixed RNA secondary structure $s \in \mathcal{S}_n$, in a graph $\mathcal{Q}_\alpha^n$ is given. We assume the latter to be obtained along the lines of Model IV of chapter 4 section 2 and illustrate the scenario by the following diagram

$$\begin{tikzcd}
 & \mathcal{Q}_\alpha^n & \\
 & \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p} \arrow[u, hook, "\iota"'] & \\
 & \Gamma_{n_u} \times \Gamma_{n_p} \arrow[u, hook, "\iota"'] & \\
\Gamma_{n_u} \arrow[uur, hook, "in"] & & \Gamma_{n_p} \arrow[uul, hook', "in"']
\end{tikzcd}$$

In the sequel we shall use the short-hand notation Γ for the neutral network $\Gamma_n^{\mathrm{IV}}[s]$. Any neutral network induces a *fitness landscape* i.e. a mapping $f_{\Gamma_n[s]} : \mathcal{Q}_\alpha^n \to \mathbb{R}_+$, as follows:

[5] In this context we give [20, 24, 47] as further references. For detailed descriptions of Gillespie algorithm see [23].

Definition 11. *Suppose a neutral network $\Gamma_n[s]$ with respect to the RNA secondary structure s is given and $\sigma \in I\!R_+$ with $\sigma > 1$. Then $\Gamma_n[s]$ induces a fitness landscape by setting:*

$$f_{\Gamma_n[s]}(v) := \begin{cases} 1 & \text{iff} \quad v \in \mathrm{v}[\mathcal{Q}_\alpha^n] \setminus \mathrm{v}[\Gamma_n[s]] \\ \sigma > 1 & \text{otherwise} \end{cases} .$$

We call $f_{\Gamma_n[s]}$ a single shape landscape.

We shall describe now a mechanism for the time evolution of a population in a single shape landscape. This landscape is on the level of secondary structures an analogue to the single peak landscape analyzed by Eigen and others. In particular we shall be interested to study the dynamics for increasing probabilities of making errorneous copies. For this purpose let us first introduce the so called “replication-deletion process”.

5.1.1. Replication Deletion Processes

Let N be a natural number such that $N \geq 2$ and let $\mathbf{V}$ be a (finite) family of vertices $(v_i \,|\, i \in \mathbb{N}_N)$ where $\{\, v_i \,|\, i \in \mathbb{N}_N \,\} \subset \mathrm{v}[\mathcal{Q}]_\alpha^n$. We shall call $\mathbf{V}$ a *population* in $\mathcal{Q}_\alpha^n$. The theory of point processes provides a powerful tool by identifying such a family $(v_i \,|\, i \in \mathbb{N}_N)$ with an integer valued measure.

$$\mathbf{V} = (v_i \,|\, i \in \mathbb{N}_N) \quad \longleftrightarrow \quad \phi := \sum_{i=1}^{N} g_{v_i}, \text{ where } g_{v_i}(v) := \begin{cases} 1 & \text{for } v \neq v_i \\ 0 & \text{otherwise} . \end{cases} \tag{26}$$

We now establish a mapping from $(v_i \,|\, i \in \mathbb{N}_N)$ to the family $(v_i' \,|\, i \in \mathbb{N}_N)$ as follows:

We select an ordered pair (v_l, v_k) where $v_l, v_k \in \{\, v_i \,|\, i \in \mathbb{N}_N \,\}$. For

$$\ell := \mathrm{res}_{\mathrm{v}[\Gamma_n[s]]}\phi(\mathrm{v}[\Gamma_n[s]])$$

the first coordinate v_l is chosen with probability $\sigma\,\ell/[(N-\ell)+\sigma\ell]$ from the elements of $\mathbf{V}$ located on the neutral network with uniform probability and from the remaining elements with uniform probability otherwise. The second coordinate of the above pair is selected with uniform probability on $(v_i \neq v_l \,|\, i \in \mathbb{N}_N)$ i.e. $1/(N-1)$. We assume the times $\hat{T}$ between these mappings to be exponentially distributed (scaled by the mean fittness)

$$\boldsymbol{\mu}\{\hat{T} \leq t\} = e^{-[(N-\ell)+\sigma\ell]\,t} .$$

Next we map $v_l = (x_1, ..., x_n)$ randomly into the vertex $v^* = (x_1', ..., x_n')$. This is performed by assigning to each coordinate x_i a $x_i' \neq x_i$ with probability p where all $x_i' \neq x_i$ are equally distributed and leave the coordinate fixed otherwise. This random mapping $v_l \mapsto v^*$ is called "replication". Finally, we delete the second coordinate of the pair (v_l, v_k), that is v_k and have a mapping $(v_l, v_k) \mapsto (v_l, v^*)$. Thereby we obtain a "new" family by substituting the v_k by the v^*. The complete mapping is called the "replication-deletion process".

Accordingly, we obtain a stochastic process $(\hat{Y}_t)_t$ in continuous time with values in

$$M_N(\mathrm{v}[\mathcal{Q}_\alpha^n]) := \{\phi \,|\, \phi \text{ is an integer valued measure on } \mathrm{v}[\mathcal{Q}_\alpha^n] \text{ and } \phi(\mathrm{v}[\mathcal{Q}_\alpha^n]) = N\,\}\,.$$

From this stochastic process we derive a further process that is defined on the natural numbers

$$(\hat{X}_t)_t := (|\,\hat{Y}_t(\mathrm{v}[\Gamma_n[s]])\,|)_t$$

that is also formulated in continuous time.

5.1.2. Some Conditional Probabilities

$f_{\Gamma_n[s]}$ induces a *bipartition* of the population $\mathbf{V}$ in $\mathcal{Q}_\alpha^n$ in the following form:
For each measure $\phi \in M_N(\mathrm{v}[\mathcal{Q}_\alpha^n])$ we consider the restrictions $\mathrm{res}_{\mathrm{v}[\Gamma_n[s]]}\phi, \mathrm{res}_{\mathrm{v}[\mathcal{Q}_\alpha^n]\setminus \mathrm{v}[\Gamma_n[s]]}\phi$. These correspond according to equation (26)

$$\mathbf{V}_\mu := \{v \in \mathbf{V} \,|\, v \in \mathrm{v}[\Gamma_n[s]]\} \quad \longleftrightarrow \quad \mathrm{res}_{\mathrm{v}[\Gamma_n[s]]}\phi$$
$$\mathbf{V}_\nu := \{v \in \mathbf{V} \,|\, v \not\in \mathrm{v}[\Gamma_n[s]]\} \quad \longleftrightarrow \quad \mathrm{res}_{\mathrm{v}[\mathcal{Q}_\alpha^n]\setminus \mathrm{v}[\Gamma_n[s]]}\phi\,.$$

whence $\mathbf{V} = \mathbf{V}_\mu \dot{\cup} \mathbf{V}_\nu$. We call call the elements of $\mathbf{V}_\mu$ *masters* (because they have a superior fitness) and those of $\mathbf{V}_\nu$ *non masters.*

Let (v_l, v_k) be a pair of vertices selected as follows:
The first coordinate of the above pair is a master vertex with probability $\sigma\,\ell/[(N-\ell)+\sigma\ell]$ and a non master vertex otherwise.
The second coordinate of the above pair is selected with uniform probability on $(v_i \neq v_l \,|\, i \in \mathbb{N}_N)$ i.e. $1/N-1$.
Let now $P_{\mu,\mu}$ and $P_{\nu,\nu}$ be the probabilities that (v_l, v_k) fulfills $v_l \in \mathbf{V}_\mu, v_k \in \mathbf{V}_\mu$ and $v_l \in \mathbf{V}_\nu, v_k \in \mathbf{V}_\nu$ respectively. We obtain:

$$P_{\mu,\mu} = \frac{\sigma\ell}{(N-\ell)+\sigma\ell}\,\frac{\ell-1}{N-1} \quad \text{and} \quad P_{\nu,\nu} = \frac{(N-\ell)}{(N-\ell)+\sigma\ell}\,\frac{(N-1-\ell)}{N-1}\,.$$

The probabilities $P_{\mu,\nu}$ and $P_{\nu,\mu}$ are defined analogously.

In order to study the stochastic process $(\hat{X}_t)_t$ we have to restrict ourselves to regular neutral networks:

Definition 12. *A regular neutral network, $\tilde{\Gamma}_{n_u,n_p}$, is the graph product of the two regular graphs $\tilde{\Gamma}_{n_u} < \mathcal{Q}_\alpha^{n_u}$, and $\tilde{\Gamma}_{n_p} < \mathcal{Q}_\beta^{n_p}$: $\tilde{\Gamma}_{n_u,n_p} := \tilde{\Gamma}_{n_u} \times \tilde{\Gamma}_{n_p}$. $\tilde{\Gamma}_{n_u}$ is a $\lceil \lambda_u \cdot n_u \rceil$-regular subgraph of $\mathcal{Q}_\alpha^{n_u}$, and $\tilde{\Gamma}_{n_p}$ a $\lceil \lambda_p \cdot n_p \rceil$-regular subgraph of $\mathcal{Q}_\beta^{n_p}$ such that*

$$\lim_{n\to\infty} |\, \tilde{\Gamma}_{n_u} \,| / \lim_{n\to\infty} \lambda_u \, \alpha^{n_u} = 1 \quad \textit{and} \quad \lim_{n\to\infty} |\, \tilde{\Gamma}_{n_p} \,| / \lim_{n\to\infty} \lambda_p \, \beta^{n_p} = 1 \, .$$

We shall write for short $\tilde{\Gamma} := \tilde{\Gamma}_{n_u,n_p}$.

Regular neutral networks shall turn out to allow to apply a birth-death model ansatz and moreover the derivation of further analytical results (see also chapter 6). The regularity assumption is in fact only a technical constraint–the neutral networks are, see lemma 3, *almost* regular graphs. Therefore it is not surprising that the results remain to be valid for the neutral networks obtained from Model IV (where the simulations are based on).

Now we are prepared to introduce the probabilities $W_{\mu,\mu}^{\tilde{\Gamma}}$ and $W_{\mu,\nu}^{\tilde{\Gamma}}$. $W_{\mu,\mu}^{\tilde{\Gamma}}$ is defined to be the probability to derive from a master vertex v_l by replication (as introduced as the mapping $(v_l, v_k) \mapsto (v_l, v^*)$) v^* as a master vertex again and $W_{\mu,\nu}^{\tilde{\Gamma}}$ the probability that a master vertex v_l is mapped into a non master vertex. The regularity assumption on the neutral network guarantees that $W_{\mu,\mu}^{\tilde{\Gamma}}$ and $W_{\mu,\nu}^{\tilde{\Gamma}}$ do *by definition* not depend on the particular vertex and are hence well defined. However both probabilities do only depend on the neutral network.

Next we want to introduce the probabilities $W_{\nu,\mu}^{\tilde{\Gamma}}$ and $W_{\nu,\nu}^{\tilde{\Gamma}}$ (the *backflow-mutations*). These shall be of particular interest when almost all elements of the population are non masters. In this case we make use of the following hypothesis (++) [10, 12]:

$$(++) \qquad \text{For} \quad 0 \le k \le N \quad \boldsymbol{\mu}\{\phi(v) = k\} \quad \text{is independent of} \quad v \in \mathrm{v}[\mathcal{Q}_\alpha^n] \setminus \mathrm{v}[\Gamma_n[s]] \, .$$

This hypothesis shall enable us to compute $W_{\nu,\mu}^{\tilde{\Gamma}}, W_{\nu,\mu}^{\tilde{\Gamma}}$ as the probabilities that a non master vertex is mapped into a master vertex. In order to compute $W_{\nu,\nu}^{\tilde{\Gamma}}, W_{\nu,\mu}^{\tilde{\Gamma}}$ under hypothesis (++) we proceed by introducing a partition of the non master vertices with respect to the neutral network. Since the neutral network is a subgraph of the graph of compatible sequences, $\mathcal{C}[s]$, (see chapter 4 section 1) all $v \in \mathbf{V}_\mu$ are in particular compatible, (i.e. sequences that *could* fold into the secondary structure s). We now arrange the vertices of $\mathrm{v}[\mathcal{Q}_\alpha^n] \setminus \mathrm{v}[\Gamma]$ in classes

$$\mathcal{E}_i := \{ v \in \mathbf{V} \,|\, v \, \text{has exactly } i \text{ incompatible base pairs} \} \, .$$

Then the "densities of non masters" in the class $\mathcal{E}_i$ is $\Delta_i := \frac{|\mathcal{E}_i \cap \mathbf{V}_\nu|}{|\mathbf{V}_\nu|}$. The $(\Delta_i)_{0 \le i \le n_p}$ are a formal analogue to the different Hamming classes studied in the case of a single peak landscapes [47]. This leads to following definition:

Definition 13. *Let $\Gamma_n[s]$ be a neutral network with respect to the secondary structure s and let $v = (v_1, \ldots, v_n)$ be a sequence. Then we define the* incompatible distance *$d(\Gamma_n[s], v)$ by*

$$d(\Gamma_n[s], v) := |\,\{[v_i, v_k]\,|\,[v_i, v_k] \notin \Pi \wedge [i, k] \in \Pi(s)\}\,|\,,$$

where Π is the pairing rule of the underlying alphabet and $\Pi[s]$ the set of contacts of s (see chapter 4 section 1).

Further we introduce the *i-th incompatible class* $\mathbf{C}_i[s]$ with respect to s:

Definition 14. *Let $\Gamma_n[s]$ be a neutral network corresponding to a secondary structure s and let v be a sequence. Then the* i-th incompatible class, *$\mathbf{C}_i[s]$, is defined by*

$$\mathbf{C}_i[s] := \{v \in \mathrm{v}[\mathcal{Q}_\alpha^n] \setminus \mathrm{v}[\Gamma_n[s]]\,|\,d(\Gamma_n[s], v) = i\} \quad \forall i = 0, \ldots, n_p.$$

In order to compute the transition probabilities $W^{\tilde{\Gamma}}_{\mu,\mu}$, $W^{\tilde{\Gamma}}_{\nu,\mu}$, $W^{\tilde{\Gamma}}_{\mu,\nu}$ and $W^{\tilde{\Gamma}}_{\nu,\nu}$ we introduce some terminology. An alphabet $\mathcal{A}$ is a *$\star$-alphabet* iff

- $\mathcal{A}$ consists of complementary bases i.e $\mathcal{A}$ can be written as $\mathcal{A} = \{A_1, A_1^c, A_2, A_2^c, \ldots, A_m, A_m^c\}$ (whence in particular $|\,\mathcal{A}\,| = \alpha = 2m$)
- The induced pair-alphabet $\mathcal{B}$ (of length β) is of the form $\mathcal{B} = \{(A_1, A_1^c); (A_1^c, A_1); \ldots; (A_m, A_m^c); (A_m^c, A_m)\}$, whence $\beta = \alpha$.

Some examples for $\star$-alphabets are $\{\mathbf{G}, \mathbf{C}\}$, $\{\mathbf{A}, \mathbf{U}\}$ and $\{\mathbf{G}, \mathbf{C}, \mathbf{X}, \mathbf{K}\}$.

For binary $\star$-alphabets of length 2 we obtain

$$|\,\mathbf{C}_i[s]\,| = \begin{cases} \binom{n_p}{i}\, 2^{n_p+n_u} & \text{for } 1 \le i \le n_p \\ 2^{n_u+n_p} - |\,\Gamma_n[s]\,| & \text{for}\, i = 0 \end{cases}$$

and consequently we obtain assuming (++) $\Delta_i = \begin{cases} \frac{\binom{n_p}{i}\, 2^{n_p+n_u}}{2^n - |\,\Gamma_n[s]\,|} & \text{for } 1 \le i \le n_p \\ \frac{2^{n_u+n_p} - |\,\Gamma_n[s]\,|}{2^n - |\,\Gamma_n[s]\,|} & \text{otherwise .} \end{cases}$

Finally we conclude this section with the following lemma

Lemma 10. *Suppose $\tilde{\Gamma} < \mathcal{Q}_\alpha^n$ is a fixed regular neutral network and $\mathcal{A}$ is a $\star$-alphabet. Suppose that we have a random mapping $v = (x_1, ..., x_n) \mapsto v' = (x'_1, ..., x'_n)$, $v, v' \in \mathrm{v}[\mathcal{Q}_\alpha^n]$ that is defined as follows: We set $x_i = x'_i$ with probability $1 - p$ and $x_i \neq x'_i$ with uniform probability p. If hypothesis $(++)$ is fulfilled, then*

$$W_{\mu,\mu}^{\tilde{\Gamma}} = [1 - (1-p)^{n_u}]\lambda_u(1-p)^{2n_p} + (1-p)^{n_u}\lambda_p\,\Phi(p) + [1-(1-p)^{n_u}]\lambda_u\lambda_p\,\Phi(p) + (1-p)^n$$

with $\Phi(p) := [(\frac{p^2}{\alpha-1} + (1-p)^2)^{n_p} - (1-p)^{2n_p}]$ and furthermore

$$W_{\nu,\mu}^{\tilde{\Gamma}} = \sum_{h=1}^{n}\sum_{\ell=0}^{h}\sum_{i=0}^{\ell} \Delta_i \binom{n_u}{h-\ell}\,\lambda_u^x(\frac{2}{\alpha-1})^i \binom{n_p - i}{\ell - i}\,(\alpha-1)^{\ell-i}\,\lambda_p\,p^h\,(1-p)^{n-h}\,,$$

with $x = \begin{cases} 0 & \textit{if } h - \ell = 0 \\ 1 & \textit{otherwise.} \end{cases}$

Proof. (i) Denoting an error at the unpaired positions with $(-,\quad)$ and at the paired positions with $(\quad,-)$, we can distinguish the following four types: $(+,+), (-,+), (+,-)$ and $(-,-)$. The probability for $(+,+)$ is obvious. For $(-,+)$, we have

$$(1-p)^{n-n_u}\sum_{k=1}^{n_u}\binom{n_u}{k}p^k(1-p)^{n_u-k}\lambda_u = [1-(1-p)^{n_u}]\lambda_u(1-p)^{2n_p}\,.$$

An error at the paired positions implies for alphabets with unequivocal complementary base pairs that *both* positions have to be changed in order to obtain a compatible pair again. With this information the cases $(+,-)$ and $(-,-)$ are straightforwardly to compute.

For other alphabets it suffices to observe that the probability to obtain a vertex $v \in \mathrm{v}[\tilde{\Gamma}]$ by a mutation event at the paired positions is given by

$$\sum_{j=1}^{n_p}\binom{n_p}{j}\left[\frac{p^2}{\alpha-1}\right]^j(1-p)^{2(n_p-j)} = \left[\frac{p^2}{\alpha-1} + (1-p)^2\right]^{n_p} - (1-p)^{n_p}\,.$$

(ii) For an incompatible configuration with exactly i incompatible positions we now assume an mutation at exactly h positions. There can be $0 \leq \ell \leq n_p$ errors on the paired positions. In order to obtain a compatible configuration, it is necessary to make an appropriate mutation in each of the i incompatible pairs. There are 2^i different choices to do so with corresponding probability $\frac{1}{\alpha-1}$. The remaining $\ell - i$ errors have to occur pairwise in the other pairs which can be done in $\binom{n_p-i}{\ell-i}$ $(\alpha-1)^{\ell-i}$ different ways. This completes the proof of the lemma. ∎

5.2. Birth-Death Models

In this section we intend to study the random variable $\hat{X}_t$ that counts the number of strings of $\hat{X}_t = |\,\hat{Y}_t(\mathrm{v}[\Gamma_n[s]])\,|$. We shall approximate the above stochastic process by a birth-death process in continuous time [31, 32, 6]. Our ansatz for the birth and death rates in the next two subsection is completely analogous to that of Nowak and Schuster [47]. Let $P_{\ell,\ell'}(t) = \mu\{X_{t+s} = \ell' \,|\, X_s = \ell\}$ be independent of s i.e. the process X is *homogeneous.* We first state an ergodic theorem that implies the existence of a stationary distribution for our birth-death process.

Theorem 9. *Let X_t be a homogeneous Markov process with finitely many states $0, ..., N$. If there exists a $0 < t^* < \infty$ such that*

$$P_{i,k}(t^*) > 0 \quad \textit{for} \quad 0 \le i, k \le N\,.$$

Then there exists the limits

$$\lim_{t\to\infty} P_{i,k}(t) = P_k \quad \textit{for} \quad 0 \le i, k \le N\,.$$

Now for $\lim_{h\to 0}\psi(h) = 0$ by definition of a birth-death process the following situation is given:

$$\begin{aligned}
P_{\ell,\ell+1}(h) &= \mathbf{P}_{\ell,\ell+1}\, h + \psi(h) && \text{for}\, h \searrow 0,\ \ell \ge 0\\
P_{\ell,\ell-1}(h) &= \mathbf{P}_{\ell,\ell-1}\, h + \psi(h) && \text{for}\, h \searrow 0,\ \ell \ge 1\\
P_{\ell,\ell}(h) &= 1 - (\mathbf{P}_{\ell,\ell+1} + \mathbf{P}_{\ell,\ell-1})\, h + \psi(h) && \text{for}\, h \searrow 0,\ \ell \ge 0\\
P_{\ell,\ell'}(0) &= \delta_{\ell,\ell'},\ \mathbf{P}_{0,-1} = 0,\ \mathbf{P}_{\ell,\ell+1},\ \mathbf{P}_{\ell,\ell-1} > 0 && \text{for}\ \ell > 0.
\end{aligned}$$

For $i \ge k$ we have $P_{i,k}(h\,(i-k)) = \sum_\ell P_{i,\ell}(h)\, P_{\ell,k}(h\,(i-(k+1))) \ge P_{i,i-1}(h)\, P_{i-1,k}(h\,(i-(k+1)))$. Further for any birth-death process we immediately verify by induction on $i-k$ that

$$P_{i,k}(h\,(i-k) \ge \prod_{\ell=0}^{i-k-1} P_{i-\ell,i-1-\ell}(h) > 0\,.$$

Along these lines we further obtain for $k \ge i$ $P_{i,k}(h\,(k-i) \ge \prod_{\ell=0}^{k-i-1} P_{i+\ell,i+1+\ell}(h) > 0$ whence the above theorem applies. The corresponding P_k can be computed by use of the Chapman-Kolmogorov backward equation and we determine the distribution in the next section.

Next we introduce the *transition rates* $\mathbf{P}_{\ell,\ell+1}, \mathbf{P}_{\ell,\ell-1}$.

$$\begin{aligned} \text{For} \quad 0 \le \ell \le N-1 \quad \mathbf{P}_{\ell,\ell+1}\, dt &:= \frac{(N-\ell)}{[\sigma\ell + (N-\ell)](N-1)} \left[\sigma\ell\, W^{\tilde{\Gamma}}_{\mu,\mu} + (N-1-\ell)\, W^{\tilde{\Gamma}}_{\nu,\mu}\right] dt \\ \text{For} \quad 1 \le \ell \le N \quad \mathbf{P}_{\ell,\ell-1}\, dt &:= \frac{\ell}{[\sigma\ell + (N-\ell)](N-1)} \left[\sigma(\ell-1)\, W^{\tilde{\Gamma}}_{\mu,\nu} + (N-\ell)\, W^{\tilde{\Gamma}}_{\nu,\nu}\right] dt \end{aligned}$$

It can be appropriate to consider a time scale depending on the *mean fitness* of the population [23]. This ansatz takes into account that a population with higher mean fitness is expected to replicate faster in time than populations with lower average fitness values. This is expressed by reaction rates in the *reactor time* $\tilde{t}$, where t, and $\tilde{t}$ can be transformed into each other by $dt = [1/N]\,[(\sigma-1)\,\ell + N]\, d\tilde{t}$. In other words $dt/d\tilde{t} = [1/N][(\sigma-1)\,\ell + N]$. The above birth and death rates then imply corresponding birth and death rates with respect to the reactor time $\tilde{t}$ as follows

$$\begin{aligned} \text{For} \quad 0 \le \ell \le N-1 \quad \mathbf{P}'_{\ell,\ell+1} &= \frac{(N-\ell)}{N(N-1)} \left[\sigma\ell\, W^{\tilde{\Gamma}}_{\mu,\mu} + (N-1-\ell) W^{\tilde{\Gamma}}_{\nu,\mu}\right] \\ \text{For} \quad 1 \le \ell \le N \quad \mathbf{P}'_{\ell,\ell-1} &= \frac{\ell}{N(N-1)} \left[\sigma(\ell-1)\, W^{\tilde{\Gamma}}_{\mu,\nu} + (N-\ell)\, W^{\tilde{\Gamma}}_{\nu,\nu}\right] \end{aligned}$$

5.2.1. Stationary Distribution

Now we compute the stationary distribution of a birth-death process whose birth and death rates can be written as

$$\begin{aligned} \text{For} \quad 0 \le \ell \le N-1 \quad \mathbf{P}_{\ell,\ell+1} &= \left[\frac{f_1(\sigma,\ell,N)}{f_2(\sigma,\ell,N)}\right] \; \Lambda_1\; [1 + \frac{C_1}{\ell}] \\ \text{For} \quad 1 \le \ell \le N \quad \mathbf{P}_{\ell,\ell-1} &= \left[\frac{f_1(\sigma,\ell,N)}{f_2(\sigma,\ell,N)}\right] \; \Lambda_2\, [1 + \frac{C_2}{(N-\ell)}]\,. \end{aligned}$$

We call birth-death processes with this property P*-processes.*

According to [15, 32] the stationary distribution $\boldsymbol{\mu}_p$ is for $1 \le k \le N$: $\boldsymbol{\mu}_p(k) = \pi_p(k)/\sum_k \pi_p(k)$, where

$$\pi_p(k) = \prod_{\ell=1}^{k} \frac{\mathbf{P}_{\ell-1,\ell}}{\mathbf{P}_{\ell,\ell-1}}\,.$$

In the following proposition we shall make use of the relation $B(z-y,y) = \frac{\Gamma(z-y)\,\Gamma(y)}{\Gamma(z)}$ where $B(x,y)$ is the Beta and $\Gamma(x)$ the Gamma function. The above relation is a classical result and was proved by Dirichlet.

Proposition 2. *Suppose a birth-death process is a* P*-process. Then for* $1 \le k \le N$ *its stationary distribution is determined by*

$$\pi_p(k) = \frac{\mathbf{P}_{0,1}}{\mathbf{P}_{k,k-1}} \, \frac{B(N,C_2)}{(k+C_1)\, B(1+C_1,k)\; B(N-(k-1),C_2)} \, [\frac{\Lambda_1}{\Lambda_2}]^{k-1} \, .$$

Proof. By assumption we can write

$$\begin{aligned} \mathbf{P}_{\ell,\ell+1} &= \left[\frac{f_1(\sigma,\ell,N)}{f_2(\sigma,\ell,N)}\right] \; \Lambda_1 \; [1+\frac{C_1}{\ell}] \\ \mathbf{P}_{\ell,\ell-1} &= \left[\frac{f_1(\sigma,\ell,N)}{f_2(\sigma,\ell,N)}\right] \; \Lambda_2 \, [1+\frac{C_2}{(N-\ell)}] \, , \end{aligned}$$

whence

$$\pi_p(k) = \prod_{\ell=1}^{k-1} \left[\frac{1+\frac{C_1}{\ell}}{1+\frac{C_2}{(N-\ell)}}\right] [\frac{\Lambda_1}{\Lambda_2}]^{k-1} \, \frac{\mathbf{P}_{0,1}}{\mathbf{P}_{k,k-1}} \, .$$

This can be rewritten as

$$\pi_p(k) = \frac{B(N+C_2-(k-1),k-1)}{(k+C_1)\; B(1+C_1,k)\; B(N-(k-1),k-1)} \, [\frac{\Lambda_1}{\Lambda_2}]^{k-1} \, \frac{\mathbf{P}_{0,1}}{\mathbf{P}_{k,k-1}} \, .$$

With $B(x,y)B(x+y,z) = B(y,z)B(y+z,x)$ we conclude

$$B(N+C_2-(k-1),k-1)/B(N-(k-1),k-1) = B(N,C_2)/B(N-(k-1),C_2)$$

and the proposition is proved. ∎

Remark. In order to apply the above proposition we introduce

$$\begin{aligned} \Lambda_1 \quad &:= \left[\sigma\, W^{\tilde{\Gamma}}_{\mu,\mu} - W^{\tilde{\Gamma}}_{\nu,\mu}\right] \quad & \Lambda_2 \quad &:= \left[W^{\tilde{\Gamma}}_{\nu,\nu} - \sigma\, W^{\tilde{\Gamma}}_{\mu,\nu}\right] \\ C_1 \quad &:= \tfrac{(N-1)\, W^{\tilde{\Gamma}}_{\nu,\mu}}{\Lambda_1} \, , \text{ and } & C_2 \quad &:= \tfrac{(N-1)\, \sigma\, W^{\tilde{\Gamma}}_{\mu,\nu}}{\Lambda_2} \, . \end{aligned} \tag{27}$$

With this notation we are prepared to compute the stationary distribution of $\hat{Z}_{t,p}$. We shall write $\hat{Z}_p$ for the stationary distribution and we further rewrite the transition probabilities as

$$\begin{aligned} \mathbf{P}_{\ell,\ell+1} &= \left[\frac{(N-\ell)\,\ell}{[\sigma\ell+(N-\ell)]\,(N-1)}\right] \; \Lambda_1 \; [1+\frac{C_1}{\ell}] \qquad \text{and} \\ \mathbf{P}_{\ell,\ell-1} &= \left[\frac{(N-\ell)\,\ell}{[\sigma\ell+(N-\ell)]\,(N-1)}\right] \; \Lambda_2 \, [1+\frac{C_2}{(N-\ell)}] \, . \end{aligned}$$

Corollary 8. *Suppose C_1, C_2, Λ_1 and Λ_2 are defined as in the above remark and $1 \leq k \leq N$. Then the stationary distribution of the above birth-death process $\hat{Z}_{t,p}$ is determined by*

$$\pi_p(k) = \frac{W^{\tilde{\Gamma}}_{\nu,\mu}}{\mathbf{P}_{k,k-1}} \, \frac{B(N, C_2)}{(k + C_1)\, B(1 + C_1, k)\ B(N - (k-1), C_2)} \left[\frac{\sigma\, W^{\tilde{\Gamma}}_{\mu,\mu} - W^{\tilde{\Gamma}}_{\nu,\mu}}{W^{\tilde{\Gamma}}_{\nu,\nu} - \sigma\, W^{\tilde{\Gamma}}_{\mu,\nu}}\right]^{k-1} ,$$

where for reactor time $\tilde{t}$ we have

$$\frac{W^{\tilde{\Gamma}}_{\nu,\mu}}{\mathbf{P}_{k,k-1}} = \frac{W^{\tilde{\Gamma}}_{\nu,\nu}(N-1)}{k\,[\sigma(k-1)W^{\tilde{\Gamma}}_{\mu,\nu} + (N-k)W^{\tilde{\Gamma}}_{\nu,\nu}]} .$$

5.3. Numerical Localization of the Error Threshold

In this section we shall apply proposition 2 to study the stationary distribution of the random variable $\hat{Z}_p$ as a function in p.

Let us recall (see corollary 8) that the stationary distribution of $\hat{Z}_p$ can (up to the factor $(\sum_k \pi[k])^{-1}$) be written as

$$\pi_p(k) = \frac{W^{\tilde{\Gamma}}_{\nu,\mu}}{\mathbf{P}_{k,k-1}} \, \frac{B(N, C_2)}{(k + C_1)\, B(1 + C_1, k)\ B(N - (k-1), C_2)} \left[\frac{\sigma\, W^{\tilde{\Gamma}}_{\mu,\mu} - W^{\tilde{\Gamma}}_{\nu,\mu}}{W^{\tilde{\Gamma}}_{\nu,\nu} - \sigma\, W^{\tilde{\Gamma}}_{\mu,\nu}}\right]^{k-1} ,$$

where $C_1, C_2, \Lambda_1, \Lambda_2$ have been defined in equation (27). We shall discuss the following two extreme cases. On the one hand we can assume that the population size N is infinite and on the other hand that $N \ll |\, \mathcal{Q}^n_\alpha \,|$. In the first case, since n is assumed to be fixed, the concentrations of masters c_μ is *nonzero* for *all* error probabilities p. In particular for any finite k holds $\lim_{N \to \infty} \boldsymbol{\mu}_p(k) = 0$. This fact can easily be obtained from the following discussion. We proceed by analyzing the $\pi(k)$. First we observe

$$\forall p > 0, N \in \mathbb{N} : \left[\frac{\sigma\, W^{\tilde{\Gamma}}_{\mu,\mu} - W^{\tilde{\Gamma}}_{\nu,\mu}}{W^{\tilde{\Gamma}}_{\nu,\nu} - \sigma\, W^{\tilde{\Gamma}}_{\mu,\nu}}\right] > 1 .$$

Further, for k, N large enough, we can approximate the Beta functions $B(N, C_2)$, $B(N-(k-1), C_2)$ and $B(k, 1 + C_1)$ by use of $\Gamma(z) = z^{z-1/2}\, e^{-z}\, (2\pi)^{1/2} + O(1/z)$. Introducing

$$\xi_1 := \frac{W^{\tilde{\Gamma}}_{\mu,\nu}}{W^{\tilde{\Gamma}}_{\nu,\nu} - \sigma\, W^{\tilde{\Gamma}}_{\mu,\nu}} \quad \text{and} \quad \xi_2 := \frac{W^{\tilde{\Gamma}}_{\nu,\mu}}{\sigma\, W^{\tilde{\Gamma}}_{\mu,\mu} - W^{\tilde{\Gamma}}_{\nu,\mu}} ,$$

we compute immediately

$$
\begin{aligned}
B(N,C_2) &= \left[\frac{(2\pi)\,(1+\xi_1)}{N\,\xi_1}\right]^{1/2} \left[1+\xi_1\right]^{-N} \left[1+\xi_1^{-1}\right]^{-N\,\xi_1} + O(1/N)\\
B(1+C_1,k) &= \left[\frac{(2\pi)\,(k/N+\xi_1)}{k\,\xi_1}\right]^{1/2} \left[1+\frac{\xi_1\,N}{k}\right]^{k} \left[1+\frac{k}{N\xi_1}\right]^{-N\,\xi_1} + O(1/N)\\
B(N+1-k,C_2) &= \left[\frac{2\pi((N-k)/N+\xi_2)}{(N-k)\xi_2}\right]^{1/2} \left[1+\frac{\xi_2(N-k)}{N}\right]^{k-N} \left[1+\frac{N-k}{N\xi_2}\right]^{-N\xi_2} + O(1/N).
\end{aligned}
$$

The above approximations allow to compute the stationary distribution of $\hat{Z}_p$ in the limes of infinite population size (since $\lim_{N\to\infty}\boldsymbol{\mu}(\hat{Z}_p = k) > 0$ implies $k \nearrow \infty$). Clearly, if p increases to $p = 1/2$, the concentration of masters decreases up to $\frac{|\tilde{\Gamma}|}{\alpha^n}$. We monitor the stationary distribution in figure 6.

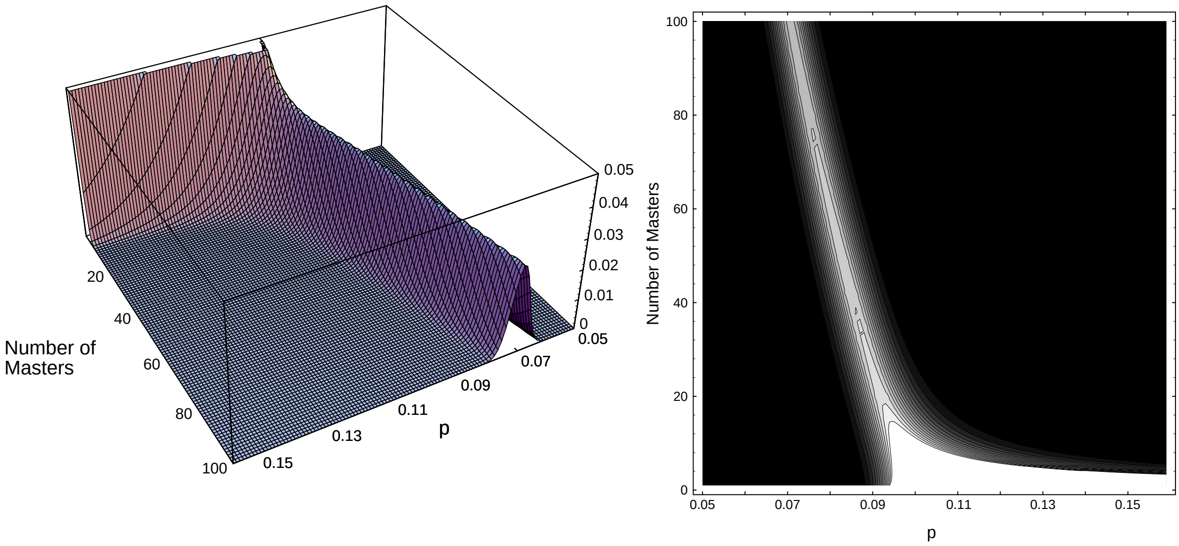


Figure 6: **a)** For a regular neutral network $\tilde{\Gamma}$ with parameters $\lambda_u = 0.5$, $\lambda_p = 0.5$ and $\sigma = 10$ we plot the stationary distribution of $\hat{Z}_p$. This means we show the density of the number of masters $\mathbf{V}_\mu$ in the population on the z-axis.
b) A contour plot of the stationary distribution of $\hat{Z}_p$ as in case **a)**.

Let us consider next the case $N \ll |\mathcal{Q}_\alpha^n| = \alpha^n$ i.e the population size is small compared to the number of all sequences. Since for any RNA secondary structures holds $n_p = O(n)$ and $n_u = O(n)$ we observe (for sufficiently large n) $|\tilde{\Gamma}| = \alpha^{n_u}\,\beta^{n_p} << \alpha^n$ or equivalently $\frac{|\tilde{\Gamma}|}{\alpha^n} << 1$. Consequently for $p \nearrow 1/2$ $\boldsymbol{\mu}(\hat{Z}_p = 0)$ is expected to become the maximum of the distribution function (although 0 is not an absorbing state). We now propose

Definition 15. *Suppose a population* $\mathbf{V} \subset \mathrm{v}[\mathcal{Q}_\alpha^n]$ *of* N *strings replicates on a regular neutral network* $\tilde{\Gamma}_n[s]$. *Then*

$$p_N^* := \max\left\{ p \mid \mathbf{V}[\hat{Z}_p] = \left[\mathbf{E}[\hat{Z}_p] - \frac{|\tilde{\Gamma}[s]|}{\alpha^n}\right]^2 \right\} \tag{28}$$

is the error threshold *of* $\mathbf{V}$ *with respect to the secondary structure* $s \in \mathcal{S}_n$. *We call* p_∞^* *the* error threshold *of the secondary structure* s.

Remark. We immediately inspect that the above mentioned criterion generalizes the one used in the case of infinite population size in the ansatz of Eigen [10] (see also [60]) that is a mean field approximation for all sequences except the master sequence. In this situation p_∞^* is the solution of $c_\mu(p^*) = 1/\alpha^n$.

Let us discuss now the case of infinite population size. In this situation we can apply a completely deterministic ansatz solving a (well-known) rate equation for the corresponding *concentrations* of master c_μ and non master vertices c_ν, respectively.

Lemma 11. *Let* $W_{\mu,\mu}^{\tilde{\Gamma}}$ *and* $W_{\nu,\mu}^{\tilde{\Gamma}}$ *be the probabilities as stated in lemma 10. Then*

$$c_\mu = \frac{\sigma W_{\mu,\mu}^{\tilde{\Gamma}} - (1 + W_{\nu,\mu}^{\tilde{\Gamma}})}{2(\sigma-1)} + \left[\left(\frac{\sigma W_{\mu,\mu}^{\tilde{\Gamma}} - (1 + W_{\nu,\mu}^{\tilde{\Gamma}})}{2(\sigma-1)}\right)^2 + \frac{W_{\nu,\mu}^{\tilde{\Gamma}}}{\sigma - 1}\right]^{1/2} .$$

Proof. In the long time limes $t \nearrow \infty$ holds

$$c_\mu\, \sigma W_{\mu,\mu}^{\tilde{\Gamma}} + [1 - c_\mu]\, W_{\nu,\mu}^{\tilde{\Gamma}} = c_\mu\, [(\sigma - 1)c_\mu + 1],$$

and the lemma follows. ∎

Remark. Assuming $W_{\nu,\mu}^{\tilde{\Gamma}} \approx 0$, i.e. neglecting back-flow mutations [10] and $\frac{|\tilde{\Gamma}|}{\alpha^n} \approx 0$, we derive

$$c_\mu \approx \left[\frac{\sigma W_{\mu,\mu}^{\tilde{\Gamma}} - (1 + W_{\nu,\mu}^{\tilde{\Gamma}})}{\sigma - 1}\right] \quad , \quad W_{\mu,\mu}^{\tilde{\Gamma}} \approx 1/\sigma \iff c_\mu = 0\,.$$

Using the above remark we can approximate (without taking into account back-flow mutations) the error thresholds in the case of infinite population size, see figure 7.

Table 1.1

Theoretical and numerical Error Thresholds (for $\sigma = 10$)

λ_u	λ_p	Theory		Gillespie Simulation
		$N = \infty$	$N = 1000$	$N = 1000$
0.1	0.1	0.079	0.071	0.065
0.27	0.5	0.081	0.08	0.0854
0.5	0.5	0.105	0.095	0.095
0.8	0.8	0.118	0.116	0.11

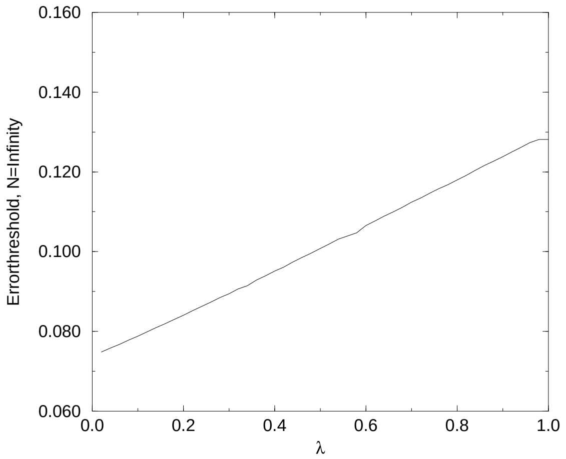


Figure 7: **a)** The error thresholds p^* of a secondary structure s with $n_u = 12$ and $n_p = 9$ for chain length 30. p^* is written as surface in the parameters λ_u, λ_p. The curve is computed with *Mathematica* [66] by numerical solution of $W^{\Gamma}_{\mu,\mu} = 1/\sigma$, where $\sigma = 10$.
b) The error thresholds p^* of a secondary structure s as described in figure 7 **a)**. Here p^* is plotted as function of $\lambda := \lambda_u = \lambda_p$.

Using the threshold criterion of definition 15 we can localize the error thresholds numerically for some population sizes and different single shape landscapes[6] with $\sigma = 10$ as superiority. The deterministic threshold values are obtained by solving $W^{\tilde{\Gamma}}_{\mu,\mu} \approx 1/\sigma$ for p (table 1). However, we have so far no analytical expression for p^*_N and the expressions derived so far don't raise hopes in finding one[7].

[6] The calculations were done with *Mathematica* [66].

[7] The known formula published in [47] in the case of Eigens single peak landscape has been derived from a different criterion, namely the vanishing of a certain local maximum in the stationary distribution of $\hat{Z}_p$. Unfortunately, this criterion is only valid for a very restricted parameter region.

Finally we end this section by plotting the densities of the i-th incompatible classes $\mathbf{C}_i[s]$ (see section 2 definition 14) of the population obtained from our simulations[8]. We observe that at the error threshold there is a *sharp transition* from a population that is localized on the neutral network to a population that is uniformly distributed in sequence space.

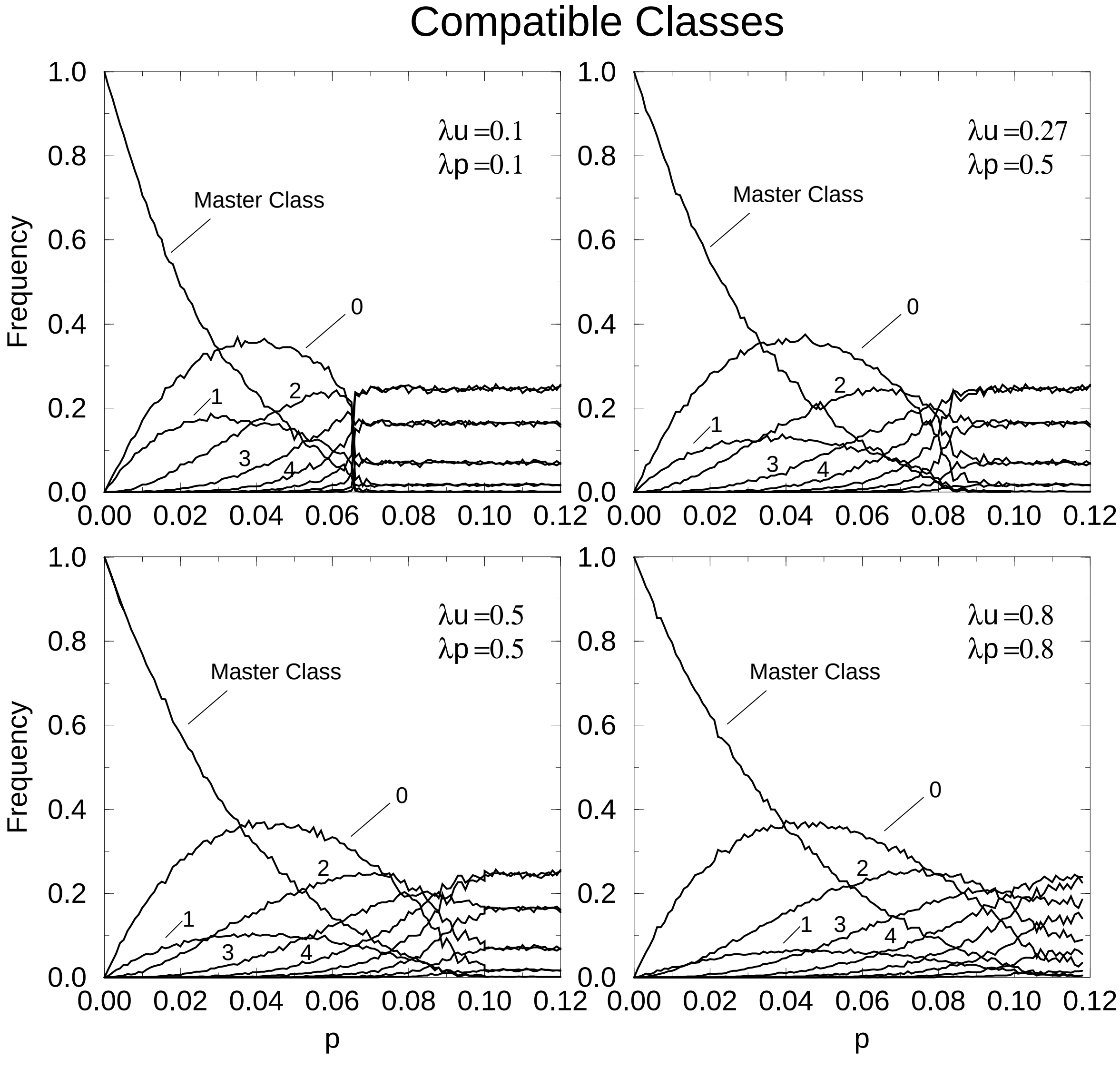


Figure 8: In this figure we plot the frequencies of masters and non masters with respect to the error-classes in incompatible distances $\mathbf{C}_i[s]$ for different single shape landscapes. $0, 1, \ldots$ represent the classes of incompatible sequences $\mathbf{C}_0[s], \mathbf{C}_1[s], \ldots$. The masters coincide with the strings that are localized on the neutral network. The underlying population size for the Gillespie simulation is $N = 1000$ and the chain length is $n = 30$.

[8] In difference to the ansatz of constant population size, (the basic assumption for the birth-death model), the simulations are obtained by use of the Gillespie algorithm [23].

6. Distribution on the Neutral Network

This chapter shall be a further application of the concept of neutral networks as random graphs. We study the population structure on the neutral network and put the results reported by Huynen and coworkers [30] on a solid mathematical basis.

Let us assume again that a secondary structure $s \in \mathcal{S}_n$ and its corresponding neutral network $\Gamma_n[s]$ are fixed. We assume $\Gamma_n[s]$ to be obtained from Model IV. In this chapter we shall study the *distribution* of the strings on the neutral network i.e. the distribution of $\mathbf{V}_\mu$. Here we understand distribution as distribution in *Hamming distances*. For this purpose we introduce the random variable

$$\hat{d}_\Gamma^\mu : \mathbf{V}_\mu \times \mathbf{V}_\mu \longrightarrow \mathbb{R} \text{ , where } \hat{d}_\Gamma(v, v') := h(v, v') \, , \tag{29}$$

and $h(\,.\,,\,.\,)$ denotes the Hamming distance. The shape of the distribution is basically determined by the following factors:

- the distribution of the random variable $\hat{Z}_\mu$ whose states are the number of offspring (that we shall assume to be independent of the particular master-vertex).
- the *structure* of the neutral network $\Gamma_n[s]$, given by the basic parameters for the construction of the random graph, $\{\lambda_u, \lambda_p, n_u, n_p\}$.
- the single digit error rate p for the replication-deletion process.

We shall assume in the sequel that $|\,\mathbf{V}_\mu\,|$ i. e. the number of strings located on the neutral network is constant. Using the results of the previous chapter we set $|\,\mathbf{V}_\mu\,| := \mathbf{E}[\hat{X}_p]$.

Our analysis can be decomposed as follows:

(i) we study abstract *genealogies* following [7, 9] in order to compute the probability for two individuals of having a common ancestor in the i-th generation.

(ii) we correlate the genealogies with random walks on neutral networks. Thereby we transform the information from the genealogies in Hamming distances.

(iii) combining (i) and (ii) we derive an analytical density function for the probability of pairs of elements of the population having a given Hamming distance.

(iv) we apply our results to the case of binary alphabets having complementary base pairs.

6.1. Step I: Reproduction Schemes and Genealogies

6.1.1. Reproduction Schemes

We shall assume that replication-deletion events to happen at discrete times $t_0, \ldots, t_1$ as follows: For subsequent times t_0, t_1 we remove the complete population $\mathbf{V}(t_0)$ at t_1 and generate the offspring of each element with the single digit error probability p. For reproduction we first use the integer valued random variable $\hat{Z}_\mu$, which counts the offspring of each $v \in \mathbf{V}_\mu(t_0)$. $\hat{Z}_\mu$ has the following two characteristics:

- $\mathbf{E}[\hat{Z}_\mu] = 1$ resulting from the constraint of constant population size.
- $\hat{Z}_\mu$ does not depend on the particular vertex $v \in \mathbf{V}_\mu$.

Accordingly, we introduce the random variable $\hat{Z}_\nu$ defined on $\mathbf{V}_\nu$. Resulting from the inferior non master-fitness we observe $\mathbf{E}[\hat{Z}_\nu] \leq \mathbf{E}[\hat{Z}_\mu]$. The complete random process is called the *reproduction scheme* $\mathcal{R}$ of $\mathbf{V}$. For a reproduction scheme $\mathcal{R}$ in general $\mathbf{V}_\mu(t_1)$ has been produced by $\mathbf{V}_\mu(t_0)$ *and* $\mathbf{V}_\nu(t_0)$. Nevertheless it is possible to introduce a particular type of reproduction scheme $\mathcal{R}^*$, that decouples master and non master vertices by increase of the offspring-production of $\mathbf{V}_\mu(t_0)$ while defining $\hat{Z}_\nu \equiv 0$. The use of the reproduction scheme $\mathcal{R}^*$ does not lead to an essentially different distribution of $\mathbf{V}_\mu$ on Γ, as long as the individuals of $\mathbf{V}_\nu \subset \mathbf{V}$ are located mainly in small Hamming classes relative to the network. Moreover lemma 10 shows that the $\mathbf{V}_\mu$ offspring is completely produced by $\mathbf{V}_\mu$ itself, as long as the superiority of $\mathbf{V}_\mu$ is high enough.[9]

6.1.2. Genealogies

We shall now study *genealogies* [8] resulting from a reproduction scheme $\mathcal{R}^*$in complete analogy to [9]. As already mentioned the number of masters N_μ is assumed to be constant i.e. $N_\mu := |\,\mathbf{V}_\mu\,|$. Let us compute the probability $\wp_1$ that two elements $v, v' \in \mathbf{V}_\mu$ have a common ancestor in the previous generation. Suppose $v \in \mathbf{V}_\mu$ has an offspring of $k \geq 2$ elements. Then we can choose $\binom{k}{2}$ different sets $\{v', v''\}$ having v as common ancestor. The expected number of those sets is consequently $N_\mu \binom{k}{2} \boldsymbol{\mu}\{\, \hat{Z}_\mu = k \,\}$ and summing over k we obtain $\frac{1}{2} N_\mu \, \mathbf{E}[\hat{Z}_\mu]_2$, whence $\wp_1 = \mathbf{E}[\hat{Z}_\mu]_2/(N_\mu - 1) = \mathbf{V}[\hat{Z}_\mu]\,/(N_\mu - 1)$. Accordingly $w_1 = 1 - \wp_1$ is the corresponding probability to have *different* ancestors one generation ago. (The probability that r vertices have a common ancestor in the previous generation is $\wp_1^{(r)} = \mathbf{E}[\hat{Z}_\mu]_r/(N_\mu - (r-1))$.) Using the result on $\wp_1$, we can express the probability of having different ancestors in *all* i previous generations:

$$w_i = \left[1 - \frac{\mathbf{V}[\hat{Z}_\mu]}{(N_\mu - 1)}\right]^i \approx e^{-\mathbf{V}[\hat{Z}_\mu]\; i/(N_\mu - 1)}\,, \tag{30}$$

[9] This means that in most scenarios there is in fact no need for rescaling the $\mathbf{V}_\mu$-offspring.

since the production of offspring are independent random events. Introducing the variable

$$\tau := t/[N_\mu - 1]\,, \tag{31}$$

we can write $w_t \approx e^{-\mathbf{V}[\hat{Z}_\mu]\,\tau}$ and

$$\frac{d}{d\tau}\{1 - w_t\} = \mathbf{V}[\hat{Z}_\mu]\; e^{-\mathbf{V}[\hat{Z}_\mu]\,\tau}\,,$$

which corresponds to the probability of finding a common ancestor in the interval $(\tau, \tau + d\tau)$. We finally summarize before proceeding the different time scales that are involved:

- τ, the discrete time in the scaling of an *elementary* reaction-step
- t, the discrete time in the scaling of generations
- $\tilde{t}$, the reactor time in the scaling of generations
- $\tilde{\tau}$, the reactor time in the scaling of an elementary reaction-step.

Remark. Introducing an extended formalism for a reproduction scheme that involves offspring of masters *and* non masters, we can compute the probabilities of emerging $\{v_\mu, v'_\mu\}, \{v_\mu, v'_\nu\}$ and $\{v_\nu, v'_\nu\}$ as offspring. In order to express the probability for two elements having different ancestors in all i previous generations (c.f. equation (30)), this probability depends on the pairs of ancestors and not only on the number of generations we trace backwards. In other words, we need to know the complete genealogy of the elements. Although all probabilities can be expressed explicitly, the formalism becomes too difficult.

6.2. Step II: Random Walks on Neutral Networks

This section is devoted to *random walks* on the neutral network $\Gamma_n[s]$. For this purpose we introduce the probability $\varphi_{\Gamma_n[s]}(t,h)$ of traveling a Hamming distance h on $\Gamma_n[s]$ by a random walk lasting t generations. We recall that $\Gamma_n[s] < \mathbf{C}[s]$ and $\mathbf{C}[s] \cong \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ and introduce the projection mappings

$$\pi_u : \Gamma_n[s] \longrightarrow \Gamma_{n_u}$$

$$\pi_p : \Gamma_n[s] \longrightarrow \Gamma_{n_p}\,.$$

Since the errors occur independently in each digit we can decompose the random walk in $\Gamma_n[s]$ in *two independent walks*: one in $\pi_u(\Gamma_n[s])$ and the other one in $\pi_p(\Gamma_n[s])$.

Clearly, $\pi_p(\Gamma_n[s]), \pi_u(\Gamma_n[s])$ are random graphs and the vertex degrees $\hat{\delta}_{v_u}, \hat{\delta}_{v_p}$ are random variables that are Gaussian (see chapter 3 section 2). Accordingly, in order to study *random walks* on neutral networks we have to restrict ourselves again to *regular* neutral networks, $\tilde{\Gamma}_{n_u,n_p}$ c.f. definition 12 in section 1. Then, writing again $\tilde{\Gamma} = \tilde{\Gamma}_{n_u,n_p}$, for each vertex $v \in \mathrm{v}[\tilde{\Gamma}]$ each adjacent vertex v' is contained in $\mathrm{v}[\tilde{\Gamma}]$ with probability λ_u (for the unpaired projection) and λ_p (for the paired projection) respectively. Further we restrict ourselves in this section to $\star$-alphabets (i.e alphabets consisting of complementary bases that admit only complementary base pairs; for example $\{\mathbf{G},\mathbf{C}\}$ or $\{\mathbf{G},\mathbf{C},\mathbf{X},\mathbf{K}\}$). We shall write for short $\varphi_{\lambda_u} = \varphi_{\pi_u(\tilde{\Gamma})}, \varphi_{\lambda_p} = \varphi_{\pi_p(\tilde{\Gamma})}$.

Lemma 12. *Suppose $\mathcal{Q}_\alpha^{n_u}, \mathcal{Q}_\beta^{n_p}$ are generalized hypercubes where $\mathcal{A}$ (the alphabet of the unpaired digits) is a $\star$-alphabet (see definition 11) and $\mathcal{B}$ is the corresponding pair alphabet. Let $\varphi_{\tilde{\Gamma}}(t,h)$ be the probability to travel a Hamming distance h by a random walk in $\tilde{\Gamma}$ lasting t generations. Further suppose that φ_{λ_u}, φ_{λ_p} are the corresponding probabilities for random walks in $\pi_u(\tilde{\Gamma}), \pi_p(\tilde{\Gamma})$. Then*

$$\varphi_{\tilde{\Gamma}}(t,h) = \sum_{h_u+2h_p=h} \varphi_{\lambda_u}(t,h_u)\,\varphi_{\lambda_p}(t,h_p)\,.$$

We consider the reproduction-deletions as point-events, i.e. we consider the random walks in continuous time. Making use of the regularity assumption on the neutral network, we obtain *infinitesimal error rates* (for unpaired and paired digits), $\lambda_u\, p\, dt$ and $\lambda_p\, p^2\, dt$.

Next we derive an ODE for the measures φ_{λ_u} and φ_{λ_p}.

Lemma 13. *Suppose $\mathcal{A}$ is a $\star$-alphabet with pair alphabet $\mathcal{B}$. For a random walk in $\pi_u(\tilde{\Gamma}) < \mathcal{Q}_\alpha^{n_u}$ and $\pi_p(\tilde{\Gamma}) < \mathcal{Q}_\beta^{n_p}$ the corresponding probabilities $\varphi_{\lambda_u}, \varphi_{\lambda_p}$ fulfill the equations*

$$\begin{aligned}\tfrac{d}{dt}\{\varphi_{\lambda_u}(t,h_u)\} &= \lambda_u\, p\Big[\tfrac{h_u+1}{\alpha-1}\,\varphi_{\lambda_u}(t,h_u+1) + (n_u-h_u+1)\,\varphi_{\lambda_u}(t,h_u-1)\\ &\qquad -(n_u-(1-\tfrac{1}{\alpha-1})h_u)\,\varphi_{\lambda_u}(t,h_u)\Big]\\ \tfrac{d}{dt}\{\varphi_{\lambda_p}(t,h_p)\} &= \lambda_p\, p^2\Big[\tfrac{h_p+1}{\beta-1}\,\varphi_{\lambda_p}(t,h_p+1) + (n_p-h_p+1)\,\varphi_{\lambda_p}(t,h_p-1)\\ &\qquad -(n_p-(1-\tfrac{1}{\beta-1})h_p)\,\varphi_{\lambda_p}(t,h_p)\Big]\,.\end{aligned}$$

Proof. Obviously,

$$\begin{aligned}\varphi_{\lambda_u}(t+dt,h_u) =&\Big[\varphi_{\lambda_u}(t,h_u-1)\,\frac{n_u-(h_u-1)}{n_u} + \varphi_{\lambda_u}(t,h_u+1)\,\frac{h_u+1}{(\alpha-1)\,n_u}\\ &+\varphi_{\lambda_u}(t,h_u)\,\frac{h_u}{n_u}\frac{\alpha-2}{\alpha-1}\Big]\times\left[n_u\,\lambda_u p dt\,(1-\lambda_u p dt)^{n_u-1}\right]\\ &+\varphi_{\lambda_u}(t,h_u)\,(1-\lambda_u p dt)^{n_u}\,.\end{aligned}$$

This is equivalent to

$$\begin{aligned}\varphi_{\lambda_u}(t+dt,h_u) =& \Big[\varphi_{\lambda_u}(t,h_u-1)\,\frac{n_u-(h_u-1)}{n_u} + \varphi_{\lambda_u}(t,h_u+1)\,\frac{h_u+1}{(\alpha-1)\,n_u}\\ &+\varphi_{\lambda_u}(t,h_u)\,\frac{h_u}{n_u}\frac{\alpha-2}{\alpha-1}\Big]\times n_u\,\lambda_u\, p dt\\ &+\varphi_{\lambda_u}(t,h_u)(1-n_u p dt) + O(dt^2)\,.\end{aligned}$$

Taking the limes $dt \to 0$ we obtain the first claim. The corresponding equation for a random walk in $\pi_p(\tilde{\Gamma})$ is derived analogously and the proof of the lemma is complete. ▮

Lemma 14. *Suppose $\mathcal{A}$ is a $\star$-alphabet with pair alphabet $\mathcal{B}$. We set*

$$\wp_u(t) := \frac{\alpha-1}{\alpha}\,(1-e^{-\frac{\alpha}{\alpha-1}\lambda_u p\,t}) \text{ and } \wp_p(t) := \frac{\beta-1}{\beta}\,(1-e^{-\frac{\beta}{\beta-1}\lambda_p p^2\,t})\,.$$

Then $\varphi_{\lambda_u}(t,h_u)$ and $\varphi_{\lambda_p}(t,h_p)$ are given by

$$\varphi_{\lambda_u}(t,h_u) = \binom{n_u}{h_u}\,\wp_u(t)^{h_u}(1-\wp_u(t))^{n_u-h_u}\ ,\ \varphi_{\lambda_p}(t,h_p) = \binom{n_p}{h_p}\,\wp_p(t)^{h_p}(1-\wp_p(t))^{n_p-h_p}\,.$$

We can view $\hat{h}_u, \hat{h}_p$ as random variables with expectation values

$$\mathbf{E}[\hat{h}_u]\,(t) = \frac{\alpha-1}{\alpha}\,n_u\left[1-e^{-\frac{\alpha}{\alpha-1}\lambda_u\,pt}\right],\ \mathbf{E}[\hat{h}_p]\,(t) = \frac{\beta-1}{\beta}\,n_p\left[1-e^{-\frac{\beta}{\beta-1}\lambda_p\,p^2 t}\right]\,.$$

Proof. We first separate the variables h, t by the ansatz:

$$\varphi(t,h) := \binom{n}{h}G(t)^h\,(1-G(t))^{n-h}\,.$$

Thereby we obtain equivalent ordinary differential equations (parameterized by h) in t.

$$\forall\,1\le h\le n:\ \frac{d}{dt}G\cdot[h-n\,G] = A\,\left[h+G\cdot[-n-\frac{\alpha}{\alpha-1}h]+G^2\cdot\frac{\alpha}{\alpha-1}n\right]\ , \tag{32}$$

where $A=\lambda_u\,p$ (in the projection to the unpaired cube) or $A=\lambda_p\,p^2$ (in the projection to the paired cube), respectively. Equation (32) is obviously equivalent to

$$\frac{dG}{dt} = A\,\left[1-\frac{\alpha}{\alpha-1}\,G\right]\,.$$

We write

$$\begin{aligned} &\int\frac{1}{1-\frac{\alpha}{\alpha-1}\,G}\,\frac{dG}{dt}\,dt = \int A\,dt + C \\ \Longrightarrow\qquad &-\frac{\alpha-1}{\alpha}\,\ln\left[1+\frac{\alpha}{\alpha-1}\,G\right] = A\,t + C\,, \end{aligned}$$

with $C\in\mathbb{R}$. Using the above equation we finally end up with

$$G(t) = \frac{\alpha-1}{\alpha}(1-e^{-\frac{\alpha}{\alpha-1}\,A\,t})$$

proving the lemma. ∎

6.3. Synthesis: Distribution of Pair Distances

The results of the two previous sections allow to compute the distribution of $\hat{d}^{\mu}_{\tilde{\Gamma}}$. For every randomly chosen pair of elements (v, v') we define

$$\boldsymbol{\mu}\{\, \tau \,\}^{(a)} := \mathbf{V}[\hat{Z}_\mu]\; e^{-\mathbf{V}[\hat{Z}_\mu]\,\tau}\, d\tau\,,$$

i.e. the probability for a pair of vertices to have a common ancestor in the interval $(\tau, \tau + d\tau)$ (see section 1). We observe that v and v' are connected by a random walk lasting the time $t = 2\,[N_\mu - 1]\,\tau$. We study the random variable $\hat{\varphi}_{\tilde{\Gamma}}(2\,[N_\mu - 1]\,\tau, h)$, defined on the probability space $\big(\{\tau \,|\, \tau \in [0, \infty)\,\}, \boldsymbol{\mu}\{\,\tau\,\}^{(a)}\big)$ in order to compute the distribution of possible Hamming distances for a given time $2\,[N_\mu - 1]\,\tau$. In the limes of infinite chain length i.e. $n \nearrow \infty$ lemma 14 shows that there is a *mapping* between times and corresponding Hamming distances. This can easily be deduced from the fact that $\hat{\varphi}_{\tilde{\Gamma}}(t, h)$ becomes localized at $\mathbf{E}[\hat{h}_u] + 2\mathbf{E}[\hat{h}_p]$, explicitly

$$\lim_{n\to\infty} \hat{\varphi}_{\Gamma}(t, h) = \begin{cases} 1 & \text{if } h = \frac{\alpha-1}{\alpha}\, n_u\,[1 - e^{\frac{\alpha}{\alpha-1}\,2\,t}] + 2\,\frac{\beta-1}{\beta}\, n_p\,[1 - e^{\frac{\beta}{\beta-1}\,2\,t}] \\ 0 & \qquad\qquad\qquad \text{else}\,. \end{cases}$$

Accordingly, we obtain $\hat{d}^{\mu}_{\tilde{\Gamma}}$ for $\star$-alphabets.[10]. The main result of this section is

Theorem 10. *Suppose $\mathcal{A}$ is an arbitrary alphabet. Let $\tilde{\Gamma}$ (see definition 11) be the regular neutral network with respect to the fixed secondary structure s. We further assume that $|\,\mathbf{V}_\mu\,| = N_\mu$. Then*

$$\boldsymbol{\mu}\{\, \hat{d}^{\mu}_{\Gamma} = h \,\} = \mathbf{V}[\hat{Z}_\mu]\; \int_0^\infty \varphi_{\tilde{\Gamma}}(2\,[N_\mu - 1]\,\tau, h)\; e^{-\mathbf{V}[\hat{Z}_\mu]\,\tau}\, d\tau\,. \tag{33}$$

The theorem implies expressing $\varphi_{\tilde{\Gamma}}(2[N_\mu - 1]\,\tau, h)$ and using lemma 14:

Corollary 9. *Suppose $\mathcal{A}$ is a $\star$-alphabet with pair alphabet $\mathcal{B}$. Then we have*

$$\begin{aligned} &\boldsymbol{\mu}\{\, \hat{d}^{\mu}_{\Gamma} = h \,\} = \\ &\mathbf{V}[\hat{Z}_\mu] \sum_{h_u + 2h_p = h} \int_0^\infty B(n_u, \wp_u(2[N_\mu - 1]\,\tau), h_u)\, B(n_p, \wp_p(2[N_\mu - 1]\,\tau), h_p)\; e^{-\mathbf{V}[\hat{Z}_\mu]\,\tau}\, d\tau\,, \end{aligned}$$

where $\wp_u((2[N_\mu - 1]\,\tau), h_u), \wp_p((2[N_\mu - 1]\,\tau), h_p)$ are defined in the previous lemma.

Using corollary 9 we can now compare the analytical distributions of $\hat{d}^{\mu}_{\tilde{\Gamma}}$ with our simulations done in the case of *binary* alphabets (see figure 9).

[10] Note that this restriction results from lemma 10 and the assumption that any two pairs have corresponding Hamming distance two.

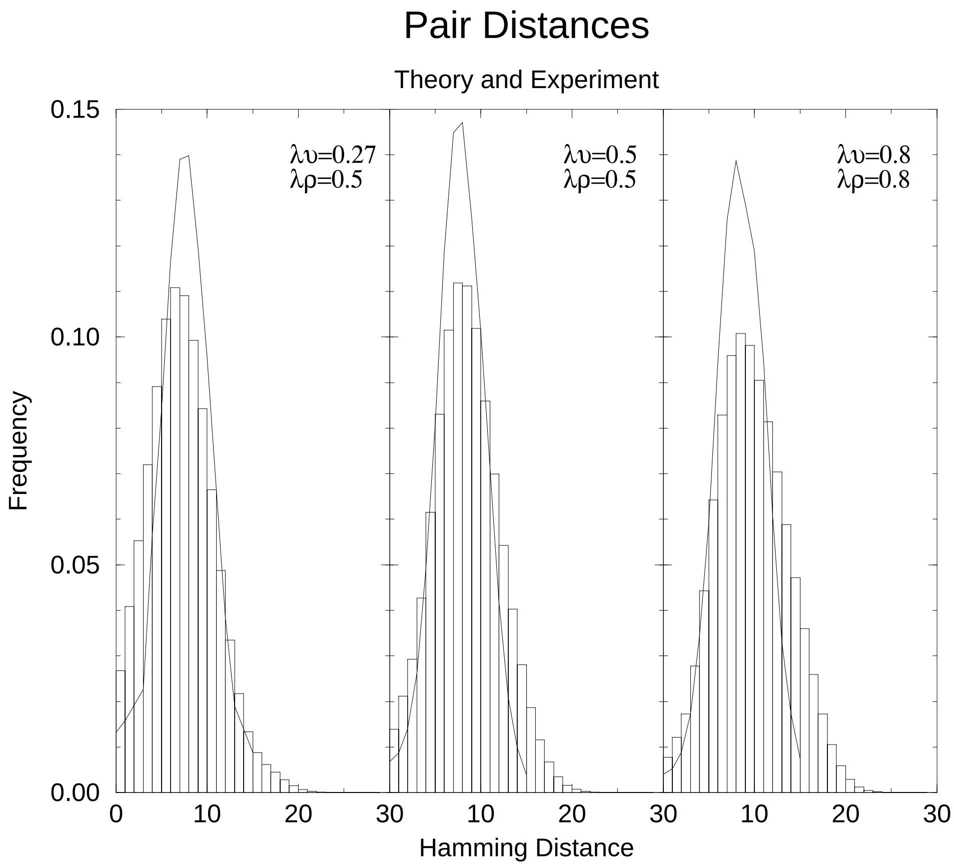


Figure 9: The distribution of $\hat{d}^{\mu}_{\tilde{\Gamma}}$ in comparison to computer simulations that base on the Gillespie algorithm [23]. The simulation data are an time average for 200 generations with an underlying chain length of $n = 30$. The solid lines denote the analytical values, the histograms show the numerical data.

Remark. The difference between the experimental and theoretical density curves is due to an effect known as *buffering* [30]. We recall the following situation to be given: $\Gamma_n[s] \hookrightarrow \mathcal{Q}^n_\alpha$ and $\Gamma_n[s]$ is a random graph whose vertex degrees are random variables (see p. 53). One observes in corresponding Gillespie simulations that on neutral networks a population is located preferably at vertices with higher degrees i.e.

$$v \in \mathrm{v}[\Gamma_n] : \, \delta_v \gg \lambda_u \, n_u + \lambda_p \, n_p \, .$$

For binary alphabets in particular the expected distance of pairs (v, v') with $\delta_v, \delta_{v'} \gg \lambda_u \, n_u + \lambda_p \, n_p$ is $n/2$, since the distance sequence of the Boolean hypercube is given by $\binom{n}{k}$. Therefore we observe a shift to higher pair distances in the population than the theory predicts for regular neutral networks.

6.4. Application: Binary Alphabets with Complementary Base Pairs

We now apply corollary 9 to describe $\hat{d}^{\mu}_{\tilde{\Gamma}}$ for the replication-deletion process on $\tilde{\Gamma}$. All results derived in this section can be formulated analogously for arbitrary alphabets lengths as long as alphabets with strict complementary base pairs (i.e. $\alpha = \beta$) are used. The first result is an immediate consequence of corollary 9.

Proposition 3. *Suppose the conditions of corollary 9 are given and furthermore $n \nearrow \infty$, $n_u = O(n)$ and $n_p = O(n)$. Then $\mathbf{E}[\hat{d}^{\mu}_{\tilde{\Gamma}}]$ and $\mathbf{V}[\hat{d}^{\mu}_{\tilde{\Gamma}}]$ are given by*

$$\begin{aligned}
\mathbf{E}[\hat{d}^{\mu}_{\tilde{\Gamma}}] &\sim \frac{n_u}{2}\left[\frac{\chi_u}{\chi_u + \mathbf{V}[\hat{Z}_\mu]}\right] + n_p\left[\frac{\chi_p}{\chi_p + \mathbf{V}[\hat{Z}_\mu]}\right] \\
\mathbf{V}[\hat{d}^{\mu}_{\tilde{\Gamma}}] &\sim \frac{(n_u/2)^2\,\chi_u^2\,\mathbf{V}[\hat{Z}_\mu]}{(\chi_u + \mathbf{V}[\hat{Z}_\mu]\,)^2(2\chi_u + \mathbf{V}[\hat{Z}_\mu]\,)} + \frac{n_p^2\,\chi_p^2\,\mathbf{V}[\hat{Z}_\mu]}{(\chi_p + \mathbf{V}[\hat{Z}_\mu]\,)^2(2\chi_p + \mathbf{V}[\hat{Z}_\mu]\,)} \\
&+ \frac{n_u\,n_p\,\mathbf{V}[\hat{Z}_\mu]\;\chi_u\,\chi_p}{(\chi_u + \mathbf{V}[\hat{Z}_\mu]\,)(\chi_p + \mathbf{V}[\hat{Z}_\mu]\,)(\chi_p + \chi_u + \mathbf{V}[\hat{Z}_\mu]\,)}
\end{aligned}$$

where $\chi_u := 4[N_\mu - 1]\,\lambda_u\,p$ and $\chi_p := 4[N_\mu - 1]\,\lambda_p\,p^2$.

Proof. The assumption implies (c.f. [15]) that the binomial measures are localized at their expectation value. Therefore we obtain the Hamming distance

$$H(2[N_\mu - 1]\tau) = \mathbf{E}[\hat{h}_u]\,(2[N_\mu - 1]\tau) + 2\mathbf{E}[\hat{h}_p]\,(2[N_\mu - 1]\tau)$$

iff the chosen pair had a common ancestor in the interval $(\tau, \tau + d\tau)$. Consequently,

$$\mathbf{E}[\hat{d}^{\mu}_{\tilde{\Gamma}}] \sim \mathbf{V}[\hat{Z}_\mu]\int_0^\infty H(2[N_\mu - 1]\tau)\;\boldsymbol{\mu}\{\,\tau\,\}^{(a)}d\tau$$

and the first claim follows. Next we observe

$$\mathbf{V}[\hat{Z}_\mu]\int_0^\infty H^2\,\boldsymbol{\mu}^{(a)}\,d\tau = 2\int_0^\infty H\,\frac{d}{d\tau}H\,\boldsymbol{\mu}^{(a)}d\tau$$

and putting $\chi_u := 4[N_\mu - 1]\,\lambda_u p$, $\chi_p := 4[N_\mu - 1]\,\lambda_p p^2$ we derive

$$\mathbf{V}[\hat{d}^{\mu}_{\tilde{\Gamma}}] = 2\int_0^\infty H\,\frac{d}{d\tau}H\,\boldsymbol{\mu}^{(a)}d\tau - \left[\mathbf{V}[\hat{Z}_\mu]\int_0^\infty H\,\boldsymbol{\mu}^{(a)}d\tau\right]^2 .$$

After a lengthy computation we derive:

$$\begin{aligned}
\mathbf{V}[\hat{d}^{\mu}_{\tilde{\Gamma}}] &= \frac{(n_u/2)^2\,\chi_u^2\,\mathbf{V}[\hat{Z}_\mu]}{(\chi_u + \mathbf{V}[\hat{Z}_\mu]\,)^2(2\chi_u + \mathbf{V}[\hat{Z}_\mu]\,)} + \frac{n_p^2\,\chi_p^2\,\mathbf{V}[\hat{Z}_\mu]}{(\chi_p + \mathbf{V}[\hat{Z}_\mu]\,)^2(2\chi_p + \mathbf{V}[\hat{Z}_\mu]\,)} \\
&+ \frac{n_u\,n_p\,\mathbf{V}[\hat{Z}_\mu]\;\chi_u\,\chi_p}{(\chi_u + \mathbf{V}[\hat{Z}_\mu]\,)(\chi_p + \mathbf{V}[\hat{Z}_\mu]\,)(\chi_p + \chi_u + \mathbf{V}[\hat{Z}_\mu]\,)}
\end{aligned}$$

and the proof of the proposition is complete. ∎

Suppose $\lambda_u = \lambda_p$, i.e. the fraction of neutral unpaired and paired bases coincides, then proposition 3 allows to compute the surface $\mathbf{F}(p, \lambda) := \mathbf{E}[\hat{d}^{\mu}_{\tilde{\Gamma}}]$ (see figure 10).

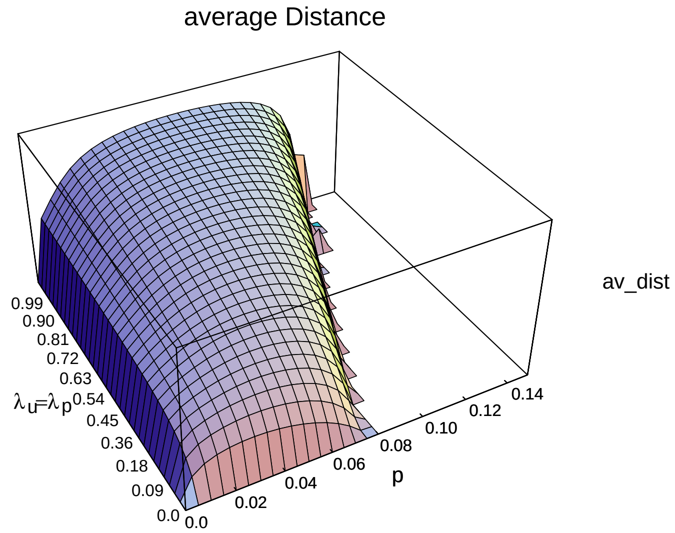


Figure 10: The average pair distance $\mathbf{E}[\hat{d}^{\mu}_{\Gamma}]$ of master fraction of the population $\mathbf{V}$ on the neutral network $\tilde{\Gamma}$ in the long time limes. We assume that $\lambda = \lambda_u = \lambda_p$. The distance is plotted as function of the single digit error rate p and the fraction of neutral neighbors for the paired and unpaired digits, λ. We observe that for wide parameter ranges the average pair distance of $\mathbf{V}_\mu$ is plateau-like. In particular the average pair distance becomes 0 at the shape-error threshold.

Definition 16. *Let $t \geq t_0$ be two times. Then*

$$\mathrm{dist}(\mathbf{V}(t), \mathbf{V}(t_0)) := \frac{1}{N_\mu^2} \sum_{\substack{v \in \mathbf{V}_\mu(t) \\ v' \in \mathbf{V}_\mu(t_0)}} d(v, v') \quad \text{and}$$

$$\mathrm{avdist}(\mathbf{V}(t), \Delta t) := \langle \mathrm{dist}(\mathbf{V}(t'), \mathbf{V}(t' + \Delta t)) \rangle_{t'} \,.$$

where $\langle\ \rangle_{t'}$ denotes the time average.

Proposition 4. *Under the conditions of corollary 9 and for binary $\star$-alphabets with complementary base pairs we obtain for a fixed time t in the limes of infinite chain length:*

$$\mathrm{avdist}[\mathbf{V}_\mu(t), \Delta t] \sim$$
$$n_u/2 \left[\frac{\chi_u}{\chi_u + \mathbf{V}[\hat{Z}_\mu]}\right] [1 - e^{-2\lambda_u\, p\, \Delta t}] + n_p \left[\frac{\chi_p}{\chi_p + \mathbf{V}[\hat{Z}_\mu]}\right] [1 - e^{-2\lambda_p\, p^2\, \Delta t}] \,.$$

Proof. We first observe that in the limes $n \nearrow \infty$ the Gaussian distributions φ_{λ_u} and φ_{λ_p} become delta distributions [15]. Hence

$$\int_0^\infty \left\{ \mathbf{E}[\hat{h}_u(2[N_\mu - 1]\,\tau) + (t_0 - t_0')] \; + 2\mathbf{E}[\hat{h}_p(2[N_\mu - 1]\,\tau) + (t_0 - t_0')] \right\} \boldsymbol{\mu}\{\,\tau\,\}^{(a)} d\tau$$
$$= \frac{n_u}{2} \left[\frac{\chi_u}{\chi_u + \mathbf{V}[\hat{Z}_\mu]}\right] [1 - e^{-2\,\lambda_u\, p\, (t_0 - t_0')}] + n_p \left[\frac{\chi_p}{\chi_p + \mathbf{V}[\hat{Z}_\mu]}\right] [1 - e^{-2\,\lambda_p\, p^2\, (t_0 - t_0')}] \,,$$

where $\chi_u = 4\,[N_\mu - 1]\,\lambda_u\,p$ and $\chi_p = 4\,[N_\mu - 1]\,\lambda_p\,p^2$ (as introduced in proposition 3) and the proof of the proposition is complete. ∎

Next we turn to the displacement of the *barycenter* of the population $\mathbf{V}_\mu$. For this purpose it is convenient to write the complementary digits v_i of the sequence $v = (v_1, ..., v_n)$ as -1 and 1 respectively. We write shall $v \cdot v' := \sum_{i=1}^{n} v_i\, v_i$.

Definition 17. *The* barycenter *of the fraction of masters $\mathbf{V}_\mu \subset \mathbf{V}$ where $|\,\mathbf{V}_\mu\,| = N_\mu$, denoted by $M^\mu(t)$, is*

$$M^\mu(t) := \frac{1}{N_\mu} \sum_{v \in \mathbf{V}_\mu} v\,.$$

Remark. $M^\mu(t)$ is an element of the n-dimensional simplex.

In the next theorem we compute the so called *diffusion-coefficient* of the barycenter $M^\mu(t)$ in the long time limes.

Theorem 11. *Suppose a population $\mathbf{V}$ replicates on the regular neutral network $\tilde{\Gamma}$ with superior fitness $\sigma > 1$. We assume that $\mathbf{V}_\mu \subset \mathbf{V}$ fulfills $|\,\mathbf{V}_\mu\,| = N_\mu$ implying a constant mean fitness $\overline{\sigma} = \frac{(\sigma-1)N_\mu + N}{N}$. Then we have*

$$\begin{aligned} &\frac{1}{\Delta t}\,\langle [M^\mu(t+\Delta t) - M^\mu(t)]^2 \rangle_t \approx \\ &2\,\lambda_u\,n_u\,p\left[\frac{\chi_u}{\chi_u + \mathbf{V}[\hat{Z}_\mu]}\right] + 4\,\lambda_p n_p p^2 \left[\frac{\chi_p}{\chi_p + \mathbf{V}[\hat{Z}_\mu]}\right]. \end{aligned}$$

and

$$\frac{1}{\Delta\tilde{t}}\langle [M^\mu(\tilde{t}+\Delta\tilde{t}) - M^\mu(\tilde{t})]^2 \rangle_{\tilde{t}} = \overline{\sigma}\,\left[\frac{1}{\Delta t}\langle [M^\mu(t+\Delta t) - M^\mu(t)]^2 \rangle_t\right],$$

where $\chi_u = 4\,\lambda_u\,p\,(N_\mu - 1)$ and $\chi_p = 4\,\lambda_p\,p^2\,(N_\mu - 1)$.

Proof. We first write $t_0 = t - \Delta t$ then

$$\langle [M^\mu(t) - M^\mu(t_0)]^2 \rangle_t = \langle M^\mu(t)^2 \rangle_t + \langle M^\mu(t_0)^2 \rangle_t - 2\langle M^\mu(t) M^\mu(t_0) \rangle_t$$

noting that in the limes of long times $\langle M^\mu(t)^2 \rangle_t = \langle M^\mu(t_0)^2 \rangle_t$. Then we inspect

$$\begin{aligned} \mathrm{dist}(\mathbf{V}(t), \mathbf{V}(t_0)) &= \frac{1}{4\,N_\mu^2} \sum_{\substack{v \in \mathbf{V}_\mu(t) \\ v' \in \mathbf{V}_\mu(t_0)}} \sum_i (v_i - v_i')^2 \\ &= \frac{1}{4}[M^\mu(t)^2 + M^\mu(t_0)^2] + \frac{1}{4\,N_\mu^2} \sum_{\substack{v \in \mathbf{V}_\mu(t) \\ v' \in \mathbf{V}_\mu(t_0)}} [-2\,v \cdot v']\,. \end{aligned}$$

Taking the time average we obtain

$$\langle M^\mu(t)^2 + M^\mu(t_0)^2 - 2M^\mu(t)\,M^\mu(t_0) \rangle_t = 4\,\mathrm{avdist}(\mathbf{V}_\mu(t), t - t_0)\,.$$

Consequently it remains to compute $\frac{1}{\Delta t}\,\mathrm{avdist}(\mathbf{V}_\mu(t), \Delta t)$. For this purpose we can approximate this expression assuming that Δt is small by differentiating with respect to t and $\tilde{t}$ (note $\Delta t / \Delta \tilde{t} = \overline{\sigma}$).

$$\frac{d}{dt}\mathrm{avdist}(\mathbf{V}(t_0), dt) = 2\ \lambda_u\, n_u\, p\ \left[\frac{\chi_u}{\chi_u + \mathbf{V}[\hat{Z}_\mu]}\right] + 4\,\lambda_p n_p p^2\ \left[\frac{\chi_p}{\chi_p + \mathbf{V}[\hat{Z}_\mu]}\right]\ .$$

This completes the proof of the theorem. ∎

7. Algebraic Representation of RNA Secondary Structures

For the investigation of the structure of the set of compatible sequences we used in chapter 4 section 1 the following algebraic framework:

Let S_n be the *symmetric group in n letters* and write a transposition $\tau \in S_n$ as $\tau = (x_i, x_k)$. Then

$$\begin{array}{rccll} \imath: & \mathcal{S} & \rightarrow & S_n & \\ & s & \mapsto & \imath(s) & := \prod_{[i,k]\in\Pi(s)}(i,k)\,. \end{array}$$

We further recall that $\imath$ naturally induced the mapping

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

7.1. Representation of RNA Secondary Structures as Involutions

The dihedral group representation will be used to obtain a new metric on the set of secondary structures that is related to the transition probability between two neutral networks [65]. For this purpose we introduce

Definition 18. *Let G be a group. A function $|\,.\,| : G \rightarrow \mathbb{R}_{0+}$ is called a length function on G if*

(i) $|x| = 0 \iff x = e$.

(ii) $|x| = |x^{-1}|$ *for all* $x \in G$.

(iii) $|xy| \le |x| + |y|$ *for all* $x, y \in G$.

Remark. Obviously, $|\,.\,|$ is a length function on G if and only if $d(x,y) = |xy^{-1}|$ is a metric. Equivalently for a given metric on G, $|x| = d(x,e)$ is a length function. The following lemma is wellknown and stated for convenience of the reader.

Lemma 15. *Suppose T is the set of all transpositions of the symmetric group in n letters, S_n. Then*

$$\ell(\pi) = n - \Theta(\pi), \qquad \pi \in S_n,$$

is a length function on S_n, where $\Theta(\pi)$ is the number of cycles into which π decomposes.

Proof. This result is well known. We give a proof here for illustrative purposes.

(i) We show first that the minimum number of transpositions is in fact a length function on S_n. We proceed by induction on $\ell(y)$. Assume y is a transposition then we have $\ell(x)-1 \le \ell(x\,y) \le \ell(x)+1$. $\ell(x\,y) \le \ell(x)+1$ holds by definition and the assumption $\ell(x\,y) < \ell(x)-1$ leads together with the first inequality to the contradiction $\ell(x) \le \ell(x\,y)+1 < \ell(x)$. Finally we assume the inequality holds for $\ell(y)=k-1$. Then for any element y' with $\ell(y')=k$ there exists a transposition τ' such that $\ell(y'\tau')=k-1$. Applying the induction hypothesis we obtain $\ell(x\,y'\tau') \le \ell(x)+\ell(y')-1$ whence $\ell(x\,y'\tau')+1 \le \ell(x)+\ell(y')$. It remains to observe $\ell(x\,y') \le \ell(x\,y'\tau')+1$ and the claim follows by the induction principle.

(ii) Each permutation can be written uniquely in a product of pairwise disjoint cycles i.e. $\pi = \prod_{j=1}^{k} Z_j^{k_j}$ where k_j is the number of cycles of length j. This representation results from the action of the cyclic group $\langle\pi\rangle$ on the set of all positions of the string. Each cycle Z_j of length j can be written uniquely as a product of $j-1$ transpositions. Therefore we obtain

$$\ell(\pi) \le \sum_j k_j\, j - \sum_j k_j$$

and it remains to show equality since $\sum_j k_j\, j - \sum_j k_j = n - \Theta(\pi)$. It is straightforward to prove by induction on j that every cycle Z_j requires at least $j-1$ different transpositions for its representation and the lemma is proved. ∎

Proposition 5. *Let s and s' be two secondary structures of length n, and let $\imath(s)$ and $\imath(s')$ be their representations as involutions. Then*

$$d^{(i)}(s,s') := \ell(\imath(s)\circ\imath(s')^{-1}) := n - \Theta(\imath(s)\imath(s')^{-1})$$

is a metric distance.

Proof. We know that the mapping $\imath : \mathcal{S} \to S_n$ is injective and that, according to the previous lemma, ℓ is a length function on S_n. ∎

Further $d^{(i)}$ induces in a natural way a graph structure on the set of secondary structures by defining the edge set e$[\mathcal{S}]$ as

$$\mathrm{e}[\mathcal{S}] := \{\{s,s'\} \,|\, d^{(i)}(s,s') = 1\}\,.$$

Obviously $(\mathcal{S}_n, \mathrm{e}[\mathcal{S}_n])$ is connected and $d^{(i)}$ coincides with the canonical metric of $(\mathcal{S}_n, \mathrm{e}[\mathcal{S}_n])$. This follows from the fact that T is a *proper* set of generators of S_n i.e. id $\notin T$ and $t \in T \iff t^{-1} \in T$ and that for all $t \in T$ we have $\ell(t) = 1$. This implies immediately $\ell(\imath(s)\circ\imath(s')) = d_{(\mathcal{S}_n,\mathrm{e}[\mathcal{S}_n])}(\imath(s),\imath(s'))$.

7.2. Subgroup Representation of RNA Secondary Structures

Next we introduce another possibility of representing the base pairing information of a RNA secondary structure.

Definition 19. *Let s be a secondary structure with associated set of contacts $\Pi(s) = \{[i,k] \mid a_{i,k} = 1,\, k \neq i-1, i+1\}$ and $T(s) := \{(i,k) \in S_n \mid a_{i,k} = 1,\, k \neq i-1, i+1\}$ the set of transpositions corresponding to the contacts of s. Then $S(s) := \langle T(s) \rangle$ i.e. the subgroup generated by $T(s)$ in S_n is the permutation group of the the secondary structure s.*

For a finite group G we denote by $\Sigma(G)$ the set of all subgroups.

Lemma 16. *The mapping $T : \mathcal{S} \to \Sigma(S_n)$ an embedding.*

Proof. Since each base is contained in at most one base pair, the transpositions belonging to one structure are disjoint, and hence commute. Obviously now, two different structures have different base pairs and hence induce different permutation groups. ∎

Definition 20. *Let G be a finite group. For any two subgroups S and S' of G we define*

$$\psi(S, S') := \ln\,[\,SS' \,:\, S \cap S'\,]\;.$$

The following proposition shows that $\psi(\,,\,)$ serves as a metric on the set of subgroups in general. In particular we have then a new matrix on the set of secondary structures.

Theorem 12. *Let G be a finite group. Then $\psi : \Sigma(G) \times \Sigma(G) \to \mathbb{R}$ is a metric on $\Sigma(G)$.*

Proof. (i) Symmetry is trivial. (ii) Clearly $[\,SS' \,:\, S \cap S'\,] \geq 1$, and this expression can be 1 only if $S = S'$. (iii) We will show that

$$[\,SS'' \,:\, S \cap S''\,]\;[\,S''S' \,:\, S'' \cap S'\,] \geq [\,SS' \,:\, S \cap S'\,]\;.$$

This is equivalent to

$$\begin{aligned} & \frac{|S|\,|S''|}{|S \cap S''|^2}\,\frac{|S'|\,|S''|}{|S' \cap S''|^2} \geq \frac{|S'|\,|S|}{|S' \cap S|^2} \\ \Longleftrightarrow \qquad & |S''|\,|S \cap S'| \geq |S'' \cap S|\,|S'' \cap S'|\,. \end{aligned}$$

Since $S \cap S' \cap S''$ is a subgroup of S, S', and S'', we may rewrite this as

$$\begin{aligned} & |S \cap S' \cap S''|\,|S''(S' \cap S)| \geq |(S \cap S'')(S' \cap S'')||(S \cap S'') \cap (S' \cap S'')| \\ \Longleftrightarrow \qquad & |S''(S' \cap S)| \geq |(S \cap S'')(S' \cap S'')|\,. \end{aligned}$$

The latter inequality is always true since both $S \cap S''$ and $S' \cap S''$ are subgroups of S'' and hence their product is still contained in S''. ∎

8. Discussion

8.1. Neutral Neighbors of RNA Secondary Structures

In order to connect the above random graph theory with the combinatory map arising from the folding of RNA sequences into their secondary structures we need to estimate the fraction of neutral neighbors in this map. Random samples of sequences for different chain lengths have been chosen, and the distribution of neutral neighbors has been computed separately for unpaired and the paired bases. We denote the average values $\overline{\lambda}_u$ and $\overline{\lambda}_p$ respectively. For the above theory to be applicable we require that $\overline{\lambda}_u$ and $\overline{\lambda}_p$ have the following two properties:

(1) Sequences folding into a given structure s are distributed approximately uniformly in the graph of compatible sequences $\mathcal{C}[s]$. This is the assumption on which our models are based in the first place.

(2) The fractions of neutral neighbors, λ_u and λ_p, become independent of the chain length n for long chains. This condition is necessary in order to ensure a finite $\bar{\lambda}$ for long sequences.

(3) The variance of the fraction of both the unpaired and the paired neutral neighbors vanishes for long chains, i.e., the *relative* vertex degree δ_v/γ becomes constant for large chains. This is a prediction of lemma 2 (p. 22).

An inverse folding algorithm [28] has been used to produce large independent samples of sequences folding, e.g., into the secondary structure of a tRNA. The distribution of these structures is indistinguishable from a sample of random sequences as far as the statistics of the pair distances is concerned [54].

Table 2. Asymptotic values for λ_u and λ_p

α	*Model II*	Alphabet	*Experimental*	
	λ^*		unpaired	paired
2	0.5	**GC**	0.2706	0.4363
4	0.3700	**GCXK**	0.4789	0.5088
		GCAU	0.4949	
6	0.3011	**GCAU**		0.4545

In figure 11a we present the dependence of $\overline{\lambda}_u$ and $\overline{\lambda}_p$ on the chain length n of the RNA molecules. The data clearly indicate that the probability of finding neutral neighbors approaches a constant in the limit of large molecules. The asymptotic values are tabulated in table 2 where those are also

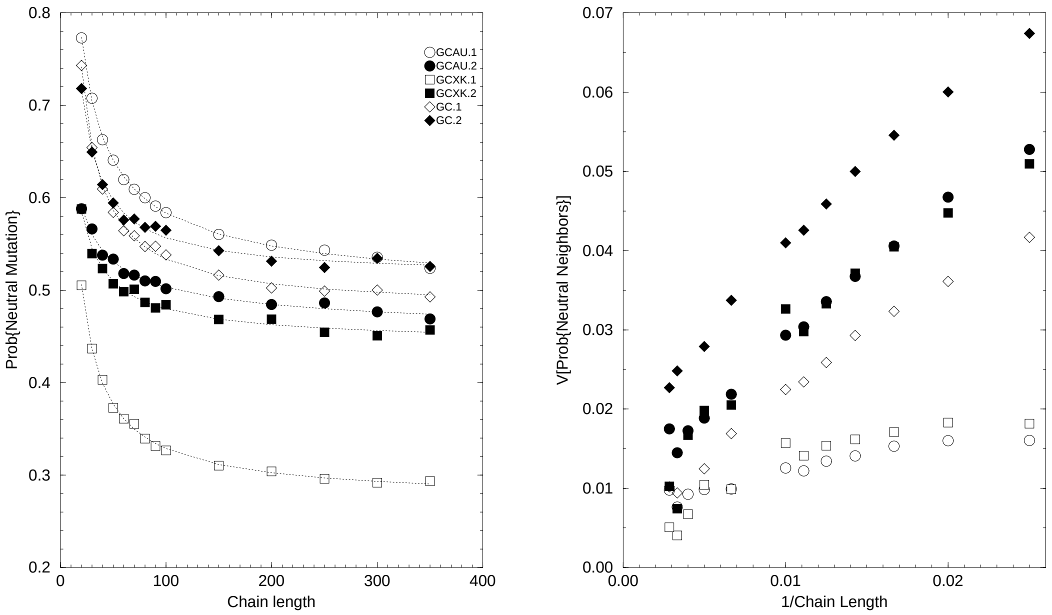


Figure 11: **a)** Frequency of neutral mutations, counted separately for single base exchanges in unpaired regions (open symbols) and base pair exchanges (full symbols) for different alphabets.
b) Variances in frequency for neutral mutations as a function of the inverse chain length.

compared with the theoretical threshold values for λ_u^* and λ_p^* respectively according to theorem 6 (p. 46) and theorem 7 (p. 49). In figure 11b we present the variance in the average frequencies of neutral neighbors for both the paired and the unpaired positions.

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

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

where $\{X_i\}$ is a partition of X and $\hat{Y}|_{X_i}$ denotes the restriction of the random variable to the subset X_i. In our situation the states of $\hat{Y}$ are the fractions of neutral neighbors of the unpaired and paired base pairs, respectively, and the partition of the sequence space consists of the preimages of the secondary structures. The empirical data in table 2 are consistent with the vanishing variance in the limes $n \nearrow \infty$, as implied by lemma 2.

The computational data presented above show that our model is in fact applicable to the combinatory map of RNA folding, as far as the *a priori* requirements of our approach are concerned. In the following sections we will discuss to what extent the properties of RNA are matched by the properties of our random graph models.

8.2. Shape Space Covering

The shape space covering theorem 8 (p. 50) 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 [27] has computed upper bounds, r_{upper} on these distances, see table 3.

Table 3. 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	15.6	12.5	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 [27]. The value $r_0 = (1 - \beta/\alpha^2)n_p(s)$ is lower bound on the shape space covering radius.

The data in table 3 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 for both model III and model IV. On the other hand, we expect the covering radius to increase only slowly when λ falls below its critical value. All we can say at this point is, that our model is consistent with the available data related to shape space covering.

Table 4. 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 [27].

The first part of theorem 8 can also be compared to computational data for RNA secondary structures. An bound on the distance between the networks of two different structures, the "closest approach" between the two networks are given in table 4. The data, produced by Hofacker [27], 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 accurate enough to decide whether the distance of closest approach in fact obeys theorem 8. Again our theory is consistent with the available numerical data within the statistical error bounds.

8.3. Percolating 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.

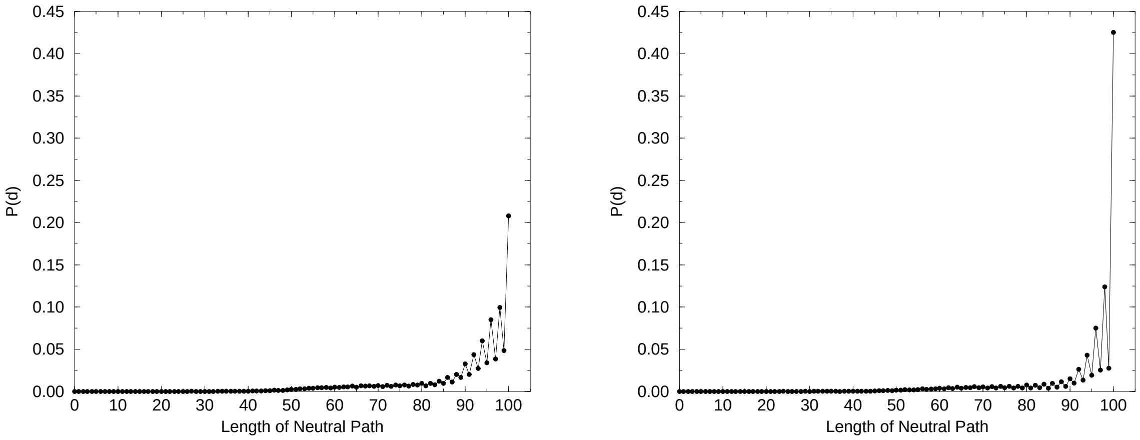


Figure 12: The probabilities of length of neutral paths in sequence space for $n = 100$. For the plots 1500 (randomly chosen) reference sequences were used.
a) AUGC alphabet, **b) GCXK** alphabet.

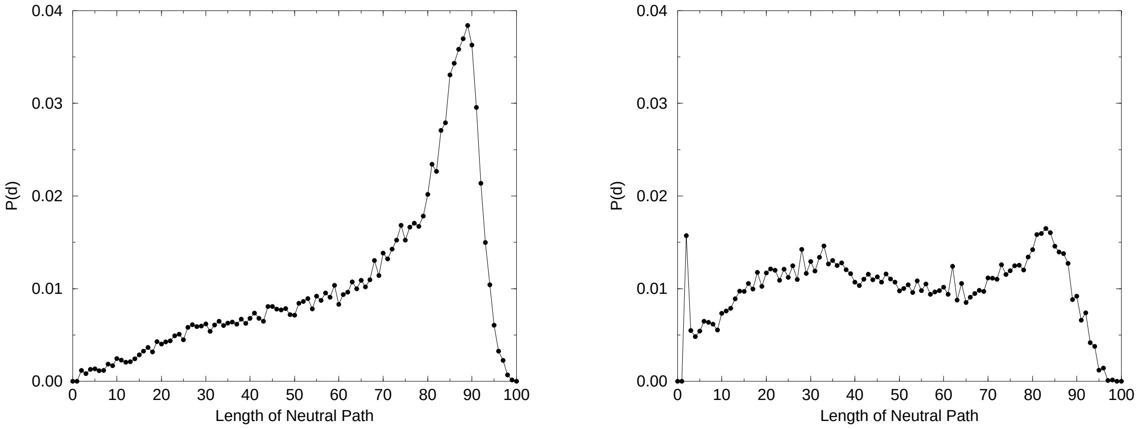


Figure 13: The probabilities of lengths of neutral paths in sequence space for $n = 100$. For the plots again 1500 (randomly chosen) reference sequences were used.
a) AU alphabet, **b) GC** alphabet.

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 [54], see figures 12 and 13. More than 20% of all path have length $\mathcal{L} = n$ for **AUGC** or **GCXK** sequences of length $n = 100$.

Neutral paths on the 2-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 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 are lots of short neutral walks terminating after just a few steps. For **AU** alphabets one inspects that the probability of having longer paths increases up to a length of 90. This observation can be explained by the fact that neutral networks with respect to frequent structures are supposed to be larger than those in the **GC** alphabet case. Over the **AU** alphabet one expects a lower number of secondary structures to be realized as minimum free energy structures since the **A-U** bond is characteristically weaker than the **G-C** bond.

These data indicate that neutral networks of the **AUGC** sequences percolate in general, while for **GC** this is not the case in most cases.

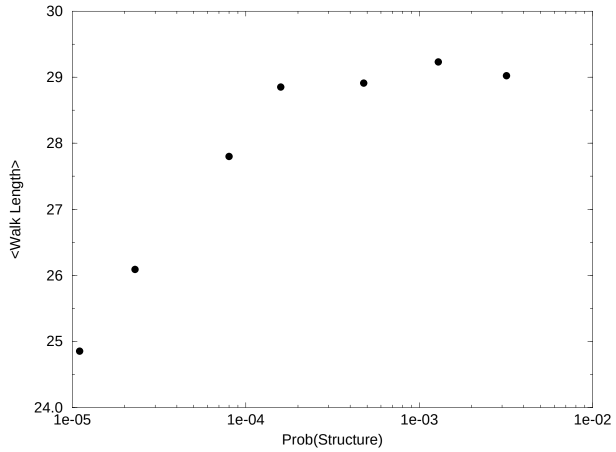


Figure 14: Average length of a neutral path for the **AUGC** alphabet as function of the abundance of the corresponding secondary structure. This is (good) lower bound on the diameter of the neutral network containing the path. Data are for **AUGC**-alphabet and chain length 30.

A more detailed analysis has been performed for the neutral networks of the **AUGC** alphabet [62], see figure 14. 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 14 are not known.

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 usual definition of percolation on regular, low-dimensional lattices [37, 59] is:

Definition 21. *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).

8.4. 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, [24].

Table 5. 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
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 15.

[†] 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 [24].

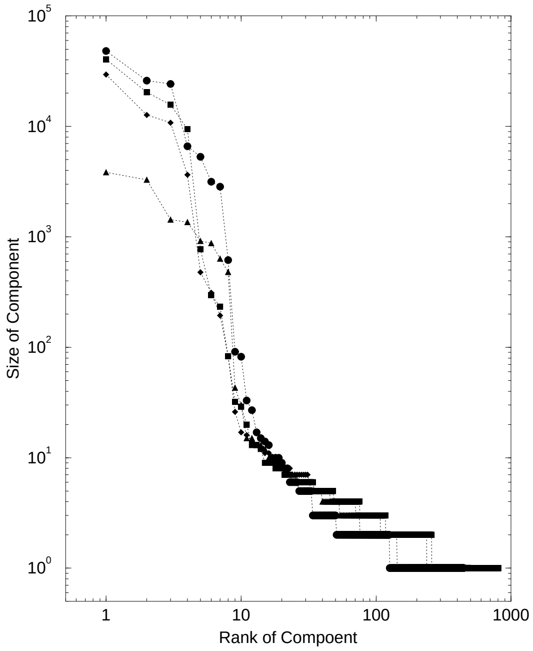


Figure 15: Sequence of Components for a few selected structures, see table 5. ●: structure #1984, □: #3030, ◇: #4723, and △: #13135. The numbers refer to the rank of the structure.

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 connectivity can be distinguished:

(1) Connected neutral networks are predicted by theorem 6 (p. 46) and theorem 7 (p. 49). In fact, the five most frequent secondary structures have connected networks, see table 5.

(2) Neutral networks consisting of a small number of components of almost equal size are ruled out by our random graph theory in the limes $n \nearrow \infty$. Nevertheless, they are found frequently in the $\mathbf{GC}_{30}$ case. There are the following possible explanations:

(i) The observations are simply a finite size effect

(ii) There exists a systematic "anisotropy" in the sequence to structure map favoring a particular ration of **G**/**C** content which depends on the structure or structural elements as e. g. loops. Clearly, in a large loop we expect a bias in the **G**/**C** content since further base pairs are forbidden. The approximate symmetry of the energy parameters implies then that one has to expect (at least) two components, see figure 16 for details.

We expect that this second effect is mainly responsible for the deviations from our random graphs models. Therefore the occurrence of several large components (of similar size) is due to structural elements and not ruled out in the long chain limit.

(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.

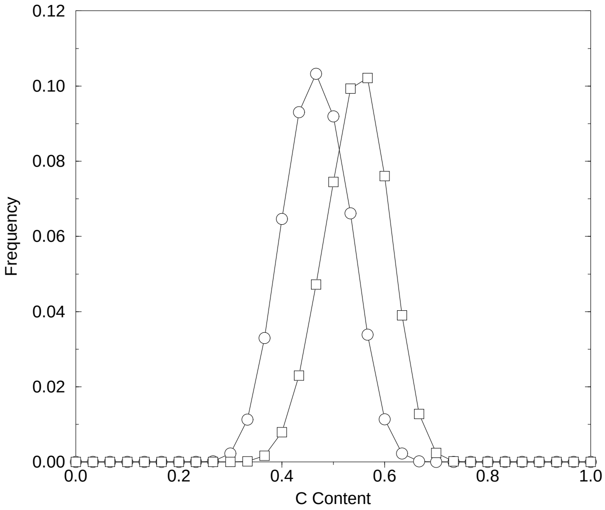


Figure 16: histogram of **GC** dependence of two large components in which the neutral network decomposes.

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 15.

8.5. The Quasispecies of RNA Secondary Structures and Error Thresholds

Eigen and Schuster's theory of the *molecular quasispecies* [10, 13, 12] describes the evolution of a population of haploid individuals in sequence space. Each string x replicates independently from all other members of the population with a sequence dependent replication rate A_x and a single digit replication accuracy $q = 1 - p$ (implying a probability for correctly replicating the entire sequence given by q^n, n being the sequence length). Analytical results for example for the frequency of certain mutants are available only for a few fitness landscapes. In particular the so called *single peak* landscape[11] has been studied in detail by many authors [60, 52, 56, 47, 12]. The main observation is here the existence of an *error threshold* $p^\dagger$ in terms of the single digit error

[11] i.e. a fitness assignment where one particular sequence the so called *master sequence* has superior fitness compared to all other sequences which have–using a mean field approach –all the same fitness

probability above which the population is essentially randomly distributed in sequence space. We show that single shape landscapes (p. 54) exhibit also error thresholds. This phenomenon has already been observed and intuitively interpreted by Fontana and Schuster [20, 19]. For single shape landscapes we observe a sharp transition from a population localized on the neutral network to a population that drifts randomly in sequence space. Above this threshold the information, manifested by the secondary structure, is destroyed. In fact p^* is a *phenotypic error threshold* as observed by computer simulations in [30] in difference to the *genotypic threshold* studied by Eigen and coworkers. In the following table we compare and summarize the main features of the genotypic and phenotypic error threshold:

Table 6. Comparison of Single Peak and Single Shape Landscape

	Single Peak Landscape	Single Shape Landscape$^\diamond$
Basic parameters	p, n, σ	p, σ, $n_u(s)$,$n_p(s)$, $\lambda_u(s)$, $\lambda_p(s)$
Partition for Quasispecies	Hamming classes with respect to the master sequence	Incompatible classes with respect to the neutral network
Threshold	genotypic threshold	phenotypic threshold
Threshold criterion*	$W_{\mu,\mu} \approx 1/\sigma$	$W_{\mu,\mu} \approx 1/\sigma$
$W_{\mu,\mu}$	$(1-p)^n$	$[1-(1-p)^{n_u}]\lambda_u(1-p)^{2n_p}+$ $(1-p)^{n_u}\lambda_p\,\Phi^\dagger(p)+$ $[1-(1-p)^{n_u}]\lambda_u\lambda_p\,\Phi(p)+$ $(1-p)^n$

$^\diamond$ with respect to a given RNA secondary structure s.

* neglecting back flow mutations and for infinite population size

† Φ is given by $\Phi(p) := [(\frac{p^2}{\kappa-1}+(1-p)^2)^{n_p}-(1-p)^{2n_p}]$.

The value of p^* depends crucially on the structure of the landscape and of the replication rate (or fitness) σ. We can localize the thresholds by mathematical modeling of the underlying stochastic process. The above results further give rise to interpreting a sequence to structure map as an abstract *coding*. Accordingly a biopolymer-structure is then an abstract "word" in this *code* (containing a certain information). Plainly, there is need for a variety of different "words" and our biological "words" have to be stable under the random action of the mutation group. The latter is observed in the constant fractions of neutral neighbors (see chapter 4). In this context it is interesting to study the dependence of the threshold of the basic parameters $\lambda_u, \lambda_p, n_u$ and n_p. These characterize "variability" and "stability" of the code.

However, the classical single peak landscape can be seen as a limiting case of our approach. One has a formal equivalence to Hamming classes: the *incompatible classes*[12]. Moreover the single shape landscape exhibits further phenomena (see the next section) like e. g. diffusion of the barycenter of the population.

[12] i.e. the classes of sequences with a certain number of incompatible base pairs (with respect to s)

Furthermore, fixing a sequence to structure map say, $f : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$, we can map the population $\mathbf{V}$ to a corresponding population of secondary structures, $f(\mathbf{V})$. Additionally choosing a metric d on the set of RNA secondary structures we obtain the metric space $(\mathcal{S}_n, d)$. Here, $f(\mathbf{V})$ forms the *quasispecies of RNA secondary structures.* Obviously the choice of the metric on RNA secondary structures is of central importance and we introduced in chapter 7 the metric $d^{(i)}$ that was defined on the corresponding involutions $\imath(s)$. This metrics focuses on the "contacts" induced by the structure and is free of any edit or cost function. In any case a metric on structures has to be seen context dependent.

8.6. Diffusion on Neutral Networks

Derrida and Peliti [9] have investigated the case of a "flat landscape" i.e $A_v = A$. 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,

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

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

$$D := \lim_{\tau \to 0} \frac{\partial \bar{\Delta}(\tau)}{\partial \tau}.$$

Extensive computer simulations of this model have been performed by Fontana and coworkers [20, 19]. Recently Huynen and coworkers [30] considered the evolution of a quasispecies on a landscape that is closely related to the combinatory map of secondary structure formation. They have chosen the structure of a 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 $p^\dagger$ is indistinguishable from 0. Nevertheless, the phenotype, that is the target structure, is conserved for moderate mutation rates. They are only lost in the simulations if the p exceeds the phenotypic error threshold p^*. Even below this critical value the population behaves similar to a population on a flat landscape. One can observe for example diffusion with a diffusion coefficient given by

$$D \approx A_{\text{target}}\, p\, n\, \lambda' / N$$

where λ' is the average fraction of neutral neighbors of a sequence in the population.

In chapter 6 we have considered a related system in which the fitness landscape is a single shape-landscape. Being neutral on the level of shapes and considering the long time limes for various λ-parameters the master-fraction of the population distributes nearly homogeneously. This has been observed even for small error-rates p. It has been proven that the population behaves like a fluid that *diffuses* on the neutral network. These observations support the neutral theory of Kimura: a *negative selection* [39][13] conserves a certain fraction of masters in the population. The latter searches by its non master offspring in the sequence space for better shapes. Evolutionary optimization is mainly obtained by the random walks of the master-fraction – positive selection occurs when an individual of the population has "found" a fitter shape. The results derived from the special case of binary alphabets with complementary base pairs as discussed in section 4 can easily be extended to alphabets with complementary base pairs of arbitrary length.

The distribution of pair distances obtained analytically for regular neutral networks has to be corrected to longer distances. This fact can be understood taking into account the localization effect described in [30]: The individuals of the population accumulate at vertices with higher degrees. Nevertheless the random walk ansatz is a good approximation for the density function of $\hat{d}^{\mu}_{\Gamma}$. It may surprise that there is no strong dependence of the average pair distances in the limes of infinite chain length on the error rate p. But in this context one has also to take into consideration the time scale in which the population reaches its stationary distribution. The time dependence can be observed in the diffusion-coefficient of the barycenter. It depends linearly on p and on the mean "fitness" $\overline{\sigma}$. Therefore the spread-out of the population occurs on an extremely short time scale, which is important since evolutionary progress takes place far from equilibrium [60]. It is known for example [11] that viruses replicate with an error-rate near their error threshold.

For moderate p values the population conserves the shape-information (guaranteed by $\mathbf{E}[\hat{X}_p] - \frac{\lfloor\tilde{\Gamma}\rfloor}{\alpha^n} \gg \sqrt{\mathbf{V}[\hat{X}_p]}$). By the non master offspring it is searched in sequence space for fitter shapes. In this context one can ask for the error rate $p^\star$ (i.e. the value with maximum number of non master-offspring in the long time limes). Surprisingly $p^\star$ does not depend strongly on the network parameters λ_u, λ_p (see fig. 16).

In other words during an evolutionary search in sequence space a population will replicate on different neutral networks (i.e. different shapes with different underlying neutral network-structures). The result reads that one error-rate $p^\star$ proves to maximize the non-master offspring in the long time limes for practically all those networks – a fundamental necessity for evolutionary adaption.

[13] recall that all sequences not located on the network have the fitness 1.

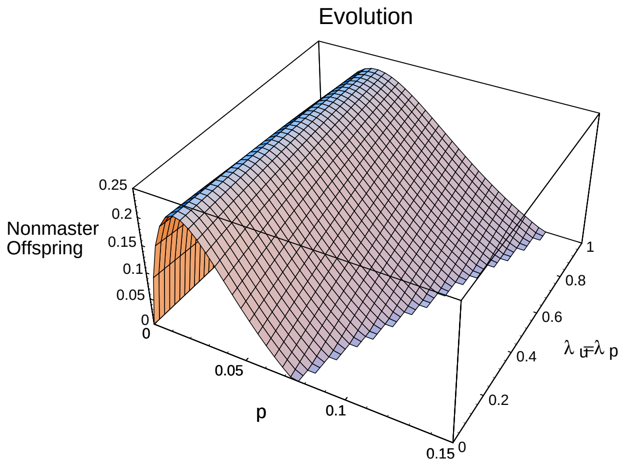


Figure 17: The surface $\mathbf{E}[\hat{X}_p]\; W^{\tilde{\Gamma}}_{\mu,\nu}$ as a function of p and $\lambda = \lambda_u = \lambda_p$ which expresses the order of the *non master offspring* of $\mathbf{V}_\mu$. For fixed $\lambda_u = \lambda_p$ these are unimodal curves. Their maximum is practically not affected by the λ parameter.

However, $p^\star$ may be far away from being "optimal" since we have to take into account the time behavior of the optimization process.

It seems hence, that the random graph approach to neutral networks is sufficiently powerful to explain essential features of the dynamics of a quasispecies on a landscape which can be decomposed as follows:

$$\boxed{\textbf{Genotypes}} \xrightarrow{\text{sequence to structure map}} \boxed{\textbf{Phenotypes}} \xrightarrow{\text{fittness assignment}} \mathbb{R}\,.$$

8.7. Metrics on RNA Secondary Structures

In the previous section we already applied the metric $d^{(i)}(s,s') = n - \Theta(\imath(s) \circ \imath(s'))$ to obtain the quasispecies of RNA secondary structures. $d^{(i)}$ can also be used to describe the transition probability between two neutral networks (corresponding to the secondary structures s, s') in sequence space. By this we mean the probability that a finite population (as introduced in chapter 5) reaches a new network [65] – a central question of evolutionary optimization. However, the structure of the intersection $\mathbf{C}[s] \cap \mathbf{C}[s']$ as introduced in chapter 4 is closely related to $d^{(i)}$. Moreover a natural metric on secondary structures allows to view a sequence to structure map as a map between *graphs*. From this point of view properties similar to continuity can be discussed and may give new perspectives in the study of sequence to structure maps. We finally remark that in chapter 7 we searched for metrics that can be obtained by considering the correla structure of the shapes.

9. Conclusions and Outlook

Let us summarize the main results shown in this thesis:

- For random induced subgraphs of configuration spaces there holds a density and a connectivity theorem. The corresponding threshold values for the above properties coincide.
- A.a.s. (asymptotically almost surely) the random induced subgraphs have a giant component.
- Neutral networks with respect to RNA secondary structures can be modeled as random induced subgraphs of configuration spaces. Here the above (abstract) results describe the structure of neutral networks.
- It holds an intersection theorem for each two sets of compatible sequences.
- Neutral networks induce single shape landscapes that exhibit phenotypic error thresholds extending the concepts of Eigen and Schuster.
- Finite populations diffuse on neutral networks with an error probability below the phenotypic error threshold.
- We present an algebraic representation of RNA secondary structures that can easily be extended to 3D-RNA or Proteins.

It turned out that the number of sequences, the number of shapes and the distribution of preimage sizes alone are not sufficient to construct a realistic model of a sequence to structure mapping. The missing ingredient is the correlation between structures of related sequences which is captured by the frequencies of neutral neighbors with respect to the base pairs and unpaired bases, λ_p and λ_u respectively. Studying sequence to structure mappings means to adopt a “new” viewpoint: instead of attacking the problem to fold a single sequence into a structure we consider the statistical properties of the complete mapping.

Random induced subgraphs of configuration spaces, obtained by independent edge and point selections, have been proven to be appropriate models for neutral networks of sequence to structure mappings. These random subgraphs exhibit threshold values for the density and connectivity property (which is closely related to a percolation phenomenon). RNA secondary structures require a refinement of the above models explicitly taking into account that unpaired and paired regions of the molecule have very different probabilities for neutral mutations. 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. The deviations observed in the sequence of components is a consequence of “biochemistry” i. e. caused by structural elements. Properties like density and connectivity are crucial for evolutionary optimization: by density one comes “close” to sequences folding into a particular structure and connectivity or giant components respectively guarantee that one can move through sequence space without loosing a given structure (i. e. conserving an eventually superior fitness). The sets of compatible sequences of any two secondary structures have a non-empty intersection implying that the neutral networks of any two secondary structures come close to each other. This allows in turn transitions between both networks and indicates how well secondary structures enable evolutionary optimization.

Our results indicate that optimization of structures by evolutionary trial and error strategies is much more effective than often suspected. In fact whole classes of sequence to secondary structure mappings, for example constructed by the random graph ansatz, are ideally suited for evolutionary adaption. Exploration of sequence space is easy because of vast neutral networks and shape space covering. Optimization is feasible since a sequence with the desired secondary structure is typically only a few point mutations away and a whole spectrum of neutral mutants searches for a better shape.

The random subgraph models provide moreover a tool for fast and yet realistic simulations of evolutionary adaptation since they give rise to realistic landscapes without requiring the time-consuming task of explicitly computing the sequence to structure mapping. We argue that estimating the frequency of neutral mutations, and even considering anisotropies related to the amino acid composition is a feasible task, while an ab-initio structure prediction will remain beyond our computational abilities in the near future. Consequently, the random graph models described here provide an indispensable tool for any simulations involving proteins.

At present we investigate the induced neutral networks of $\mathbf{C}^*$ random maps. We investigate here also properties “density” , “connectivity” and in particular study the sequence of components. Further the random graph approach and the representation of secondary structures as involutions have initiated the study of *transitions* between two neutral networks. The latter shall give further inside into aspects of the so called “neutral theory” of Kimura. Our theory makes feasible the study of “neutral evolution” basing on model landscapes on which we can formulate a rigorous stochastic formalism. New concepts, for example *random group theory*, could arise from this biological motivation namely viewing structures a abstract contact-matrices that can be embedded in permutation groups. However, a particular challenge for us is to prove the above results in the case of RNA-3D structures and we finally hope to be able to extend our approach to protein folding.

Appendix A: Integer Valued Random Variables

The sieve formula [3] (p.17) 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 13. *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 will be used frequently in the paper:

Corollary 10. *Let $\mu = \mu(n)$ be a bounded, non–negative function on $\mathbb{N}$ and assume a sequence of non–negative integer valued random variables $(\hat{X}_i)_{i\in\mathbb{N}}$ to be given. Suppose we have for arbitrary natural number r $\lim_{n\to\infty} \mathbf{E}[\hat{X}_n]_r - \mu^r = 0$. Then there holds the following convergence in distribution:*

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

where P_μ is the Poisson measure.

The following classical results is also used in the main text. A proof can be found, e.g., in [15].

Theorem 14. *[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 11. *If $x := \Delta/\sqrt{p(1-p)m} \to \infty$ 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: Simulation of finite Populations on Neutral Networks

"ansatz of Gillespie"

The time evolution of a spatially homogeneous mixture of chemically reacting molecules is usually calculated by solving a set of coupled ordinary differential equations. If there are N chemically active molecular species present, there will be N differential equations in the set.

The justification for using the stochastic approach, as opposed to the mathematically more simple deterministic approach, is that it takes *fluctuations and correlations* into account. It was demonstrated by Oppenheim et al. and proved by Kurtz that the stochastic formulation reduces to the deterministic formulation in the thermodynamic limit.

In the stochastic formulation reaction constants are not viewed as reaction rates but as reaction probabilities per unit time. The temporal behavior of a chemically reacting system takes the form of a Markovian random walk in the N-dimensional state space of the molecular populations of the N species. In the stochastic formulation of chemical kinetics the time evolution is analytically described by a single differential-difference equation for a grand probability function in which time and the N species' populations all appear as independent variables. The problem may be formulated as follows:

- There is given a volume V containing molecules of N chemically active species S_i ($i = 1, \dots, N$) and possibly molecules of several inert species.
- Let X_i be the current number of molecules of the species S_i in V with $i = 1, \dots, N$.
- The N species S_i can participate in M chemical reactions R_μ ($\mu = 1, \dots, M$), each characterized by a numerical *reaction parameter* c_μ which will be defined momentarily.
- A haploid replication process on a neutral network can be written as $\{R_{\mu_1}\}$: $S_i \to S_i + S_j$ and the deletion process as $\{R_{\mu_2}\}$: $S_i \to *$.
- The *fundamental hypothesis* of the stochastic formulation of chemical kinetics states that the reaction parameter c_μ can be defined as follows:
  $c_\mu\, \delta t \equiv$ average probability, to first order in δt, that a particular reaction R_μ appears in the next time interval δt.
- The principle task now is to develop a method for simulating the time evolution of the N quantities $\{X_i\}$, knowing only their initial values $\{X_i^{(0)}\}$, the forms of the M reactions $\{R_\mu\}$ and the values of the associated reaction parameters $\{c_\mu\}$.

Let $\mathcal{P}(X_1, X_2, \ldots, X_N; t)$ be the probability that there will be X_1 molecules of Specie S_1, X_2 molecules of Specie S_2, ..., X_N molecules of Specie S_N in the Volume V at time t.

The number X_i of S_i molecules found at time t will vary from run to run. But one may assume that in the limit for infinity many runs the values $X_i(t)$ approach to the average value and the variance of the values $X_i(t)$ is finite too.

The so-called master equation is the time evolution equation for the function $\mathcal{P}(X_1, \ldots, X_N; t)$. Often it turns out to be very fruitless to solve the master equation both analytically and numerically. That is why there is defined another quantity called the *reaction probability density function*, $P(\tau, \mu)$.

Definition 22. $P(\tau, \mu)d\tau \equiv$ *probability at time t that the next reaction in the volume V will occur in the differential time interval* $(t + \tau, t + \tau + d\tau)$ *and will be a* R_μ *reaction.*

Using the notations

- c_μ as reaction parameter characterizing the reaction R_μ. It is known by analytical calculation $\left[c_\mu = V^{-1}\, \pi\, d_{12}^2 \ exp\left(\frac{-u_\mu^*}{k\,T}\right) \sqrt{\frac{8\,k\,T}{\pi\, m_{12}}}\right]$ or experiments.
- h_μ as the number of distinct molecular reactant combinations for reactionR_μ found to be present in V at time t

it was shown in Gillespie [23] that there exists an exact expression for $P(\tau, \mu)$:

$$P(\tau, \mu) = h_\mu c_\mu \cdot \exp\left(-\sum_{\nu=1}^{M} h_\nu c_\nu \tau\right)$$

where $0 \leq \tau < \infty$, $\tau \in \mathbb{R}$, $1 \leq \mu \leq M$, $\mu \in \mathbb{N}$ and $P(\tau, \mu) = 0$ for all other τ, μ.

The simulation is done as follows:

Step 0: **Initialization:** Set $t = 0$, specify and store initial values for the N variables $X_1, \ldots, X_N$. Specify and store the values $c_1, \ldots, c_M$ for the set of M chemical reactions $\{R_\mu\}$. Calculate and store the M quantities $h_1 c_1, \ldots, h_M c_M$. Specify and store a series of "sampling times"$t_1 < t_2 < \ldots$ and a "stopping time"t_{stop}.

Step 1: Generate by suitable Monte Carlo techniques one random pair (τ, μ). (How to do this is shown below.)

Step 2: Using the numbers τ and μ generated in Step 1, advance t by τ and change the $\{X_i\}$ values of those species involved in reaction R_μ. Then recalculate the $h_\nu c_\nu$ quantities for those reactions R_ν whose reactants X_i-values have just been changed.

Step 3: If t has just been advanced through one of the sampling times t_i, read out the current molecular population values $X_1, \ldots, X_N$. If $t > t_{stop}$ or $h_\mu = 0$ for all μ terminate the calculation, otherwise return to Step 1.

By carrying out the above procedure from time 0 to time t, there is only obtained one possible realization of the stochastic process. In order to get a statistically complete picture of the temporal evolution of the system, there have to carry out several independent realizations, each starting with the same initial set of molecules and proceeding to the same time t.

We make use of the following procedure in order to carry out step 1 in the simulations.

Let $P(\tau, \mu)\, d\tau$ be the probability at time t that the next reaction in the fixed volume V will occur in the differential time interval $(t+\tau, t+\tau+d\tau)$ and will be a R_μ reaction. In terms of probability theory $P(\tau, \mu)$ is a joint probability density function on the space of the continuous variable τ and the discrete variable μ. Now $P(\tau, \mu)$ is written in the form $P(\tau, \mu) = P_1(\tau) \cdot P_2(\mu \,|\, \tau)$. $P_1(\tau)d\tau$ is the probability that the next reaction will occur between times $t+\tau$ and $t+\tau+d\tau$, irrespective of which reaction it might be. Further $P_2(\mu \,|\, \tau)$ is the probability that the next reaction will be a R_μ reaction, given that the next reaction occurs at time $t+\tau$.

By applying the addition theorem for probabilities we obtain $P_1(\tau) = \sum\limits_{\mu=1}^{M} P(\tau, \mu)$. Therefore it follows for $P_2(\mu \,|\, \tau)$ $P_2(\mu|\tau) = P(\tau, \mu) / \sum\limits_{\nu=1}^{M} P(\tau, \nu)$. Setting $a_\mu = h_\mu c_\mu$ and $a = \sum\limits_{\mu=1}^{M} a_\mu$ one finally gets $P_1(\tau) = a \cdot e^{-a\,\tau}$, $P_2(\mu \,|\, \tau) = a_\mu / a$. $P_2(\mu \,|\, \tau)$ is independent of τ. The idea is therefore

- to generate a random value τ according to $P_1(\tau) = a \cdot e^{-a\,\tau}$ and
- then to generate a random integer μ according to $P_2(\mu \,|\, \tau) = a_\mu / a$.

In other words, a random value τ can be generated according to $P_1(\tau)$ by simply taking a random number r_1 from the uniform distribution in the unit interval and setting $\tau = (1/a) \ln(1/r_1)$. Further a random integer μ can be obtained evaluating a number r_2 form the uniform distribution in the unit interval and taking μ as the integer fulfilling

$$\sum_{\nu=1}^{\mu-1} a_\nu < r_2\, a < \sum_{\nu=1}^{\mu} a_\nu \,.$$

Notation

$\mathbf{Map}(X,Y)$	Set of all maps $f : X \to Y$ (p.10)
$\hat{Z}_k(f)$	Random variable that counts all preimages of f having size k (p.10)
$B(k,n,p)$	Binomial distribution: $\binom{n}{k}\,p^k\,(1-p)^{n-k}$ (p.11)
$\mathrm{dis}(x_k,m,x)$	Set of all different distributions of x elements over m different cells with x_k cells containing k elements and no cell empty (p.11)
$\mathcal{Q}^n_\alpha$	The (Hamming) graph of all sequences over the alphabet $\mathcal{A}$ of length α with chain length n (p.12)
$\mathcal{S}_n$,$\mathcal{S}$	The set of all secondary structures of sequences of chain length n (p.13, see def.4 p.42)
$\imath(s)$	rank of a structure s
n	Parameter of system size, in particular chain length
$\mathcal{C}_n$	Configuration space (see def.1 p.15-16)
$\mathcal{C}[s]$	Graph of compatible sequences with respect to s (p.45)
$\mathrm{v}[G]$	Vertex set of the graph G (p.14)
$\mathrm{e}[G]$	Edge set of the graph G (p.14)
$\overline{\tau},\overline{\imath}$	Incidence maps w.r.t. an edge e (p.14)
$\boldsymbol{\mu}_p\{[G]\}$	$p^{\mid \mathrm{e}[G] \mid}\,(1-p)^{\mid \mathrm{e}[H] \mid - \mid \mathrm{e}[G] \mid}$
$\boldsymbol{\mu}_p\{G^*\}$	$\boldsymbol{\mu}_p\{G^*\} = \boldsymbol{\mu}_p\{[G]\}$
$H[V]$	Induced subgraph of the vertex set V in the finite graph H
$\mathcal{G}^{\mathrm{I}}(H)$	The set of all induced subgraphs Γ^{I} of the finite graph H for which there exists a nonempty graph $G^* < H$ such that $\Gamma^{\mathrm{I}} = H[\mathrm{v}[G^*]]$ (p.18)
$\mathcal{G}^{\mathrm{II}}(H)$	The set of all induced subgraphs of the finite graph H (p.19)
$\boldsymbol{\mu}_\lambda\{\Gamma\}$	$\lambda^{\mid \mathrm{v}[\Gamma] \mid}\,(1-\lambda)^{\mid \mathrm{v}[H] \mid - \mid \mathrm{v}[\Gamma] \mid}$ (p.19)
$\mathcal{B}_n(k,p)$	Discretized version of the Gaussian distribution with mean $p\,n$ and variance $p\,(1-p)\,n$ (p.20)
$\hat{\omega}_n$	Random variable that counts the order of a random graph
$\mathbf{C}[s]$	$\mathrm{v}[\mathcal{C}[s]]$ (p.20)
$\mid X \mid$	Cardinality of X as a set.
δ_v	number of adjacent vertices (with respect to the graph G) of $v \in \mathrm{v}[G]$ (p.14)
$(\Omega,\mathcal{A},\boldsymbol{\mu})$	Probability space consisting of point set, σ-field and (probability) measure (p.17)
$\hat{X}$	X is a random variable

$\hat{X}^j_{n,k}$	Random variable on random graphs Γ^j_n that counts the number of vertices with degree k (p.23)
$\overline{\mathrm{v}[\Gamma_n]}$	The set of all vertices that are either adjacent or contained in $\mathrm{v}[\Gamma_n]$
$\hat{Z}_n$	Random variable for random graphs of model II that counts the number of vertices in $\mathrm{v}[\mathcal{C}_n] \setminus \overline{\mathrm{v}[\Gamma_n]}$ (p.24)
$\mathbf{E}[\hat{X}]$	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 $\hat{X}$.
$\mathrm{Cov}[\hat{X},\hat{Y}]$	$\mathbf{E}[\hat{X}\,Y] - \mathbf{E}[\hat{X}]\,\mathbf{E}[\hat{Y}]$
$\mathcal{X},\mathcal{X}'$	Components of a random graph Γ_n
$\partial_{\Gamma_n} V$	The adjacent vertices in the graph Γ_n to a subset $V \subset \mathrm{v}[\Gamma_n]$.
$\overline{\mathcal{X}}$	$\mathrm{v}[\mathcal{X}] \cup \partial\mathcal{X}$
$B_r(v)$	$\{v' \in \mathrm{v}[\mathcal{C}] \mid d(v',v) \le r\}$, the ball with radius r and center v.
$d(\Gamma_1,\Gamma_2)$	The minimum distance between the graphs Γ_1 and Γ_2 considered as subgraphs of $\mathcal{Q}^n_\alpha$ (p.50)
n_u, n_p	The number of unpaired and paired bases of a certain secondary structure.
n, n_u, n_p	Chain length, number of unpaired bases and number of paired bases
$\mathcal{Q}^{n_u}_\alpha, \mathcal{Q}^{n_p}_\beta$	The projections on the unpaired and paired bases
Π	A pairing rule of an alphabet
$\Pi[s]$	The set of contacts of a RNA secondary structure s (p.43)
$\mathbf{V}$	The population
$\mathbf{V}_\mu, \mathbf{V}_\nu$	The master-fraction and the non master-fraction of $\mathbf{V}$
p^*_N	The critical mutation rate of a finite population of N strings replicating on a regular neutral network $\tilde{\Gamma}$
p^*_∞	The error threshold of a secondary structure
$p^\dagger$	The genotypic error threshold
$\hat{d}^\mu_{\Gamma_n}$	The random variable counting the distance of pairs in $\mathbf{V}_\mu$
$\hat{Z}_\mu, \hat{Z}_\nu$	The random variables that count the number of master-offspring and nonmaster-offspring
f_{Γ_n}	The fittness landscape induced by the neutral network Γ_n
M^μ	The barycenter of the master-fraction of the population
S_n	Symmetric group in n letters (p.44)
D_m	A dihedral group of order $2\,m$

$T(s)$	The transpositions corresponding to the base pairs of a RNA secondary structures
$d^{(i)}(s,s')$	The metric $d^{(i)}(s,s') = n - \Theta(\imath(s) \circ \imath(s'))$
$\mathbf{C}_i[s]$	The class of sequences in i-th incompatibel distance
Δ_i^u	The uniform nonmaster density with respect to classes of incompatible base pairs of a given secondary structure s
$S(s)$	$\langle T(s) \rangle$, the generated group of $T(s)$

Bibliography

[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] K. Chung. *Markov chains with stationary transition probabilities.* Springer-Verlag, Berlin, 1960.

[7] L. Demetrius. Random spin models and chemical kinetics. *J. Chem. Phys.*, 87(12):6939–6946, 1987.

[8] L. Demetrius, P. Schuster, and K. Sigmund. Polynucleotide evolution and branching processes. *Bull. Math. Biol.*, 47(2):239 – 262, 1985.

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

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

[11] M. Eigen, W. C. Gardiner Jr., C. K. Biebricher, and M. Gebinoga. The hypercycle. coupling of RNA and protein biosynthesis in the infection cycle of an RNA bacteriophage. *Biochemistry*, 30:11005 – 11018, 1991.

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

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

[14] Erdős. Graph theory and probability. *Canad.j.Math.*, 11:34–38, 1959.

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

[16] R. A. Fisher. *The genetical theory of natural selection.* Oxford: Clarendon Press., 1930.

[17] 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.

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

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

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

[21] 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.

[22] 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.

[23] D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Chem. Phys.*, 81:2340–2361, 1977.

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

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

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

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

[28] 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.

[29] J. Hofbauer and K. Sigmund. *Evolutionstheorie und dynamische Systeme.* Parey, 1984.

[30] M. Huynen, P. F. Stadler, and W. Fontana. Evolutionary dynamics of RNA and the neutral theory. *Nature*, 1994. submitted.

[31] J.L.Doob. *Stochastic Processes.* J.Wiley and Sons, New York, 1953.

[32] S. Karlin and H. M. Taylor. *A first course in stochastic processes.* Academic Press, second edition, 1975.

[33] S. A. Kauffman. The large scale structure and dynamics of gene control circuits. *J. Theor. Biol.*, 44:167–190, 1974.

[34] S. A. Kauffman. Emergent properties of random cellular automata. *Physica D*, 10:145–156, 1984.

[35] S. A. Kauffman. *The Origin of Order.* Oxford University Press, New York, Oxford, 1993.

[36] S. A. Kauffman and S. Johnsen. Co-evolution to the edge of chaos: Coupled fitness landscapes, poised states and co-evolutionary avalanches. In C. T. J. F. C. Langton and S. Rasmussen, editors, *Artificial Life II*, pages 325–369. Adison Wesley, Redwood City, 1991.

[37] H. Kesten. *Percolation Theory for Mathematics.* Birkhäuser, 1982.

[38] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624 – 626, 1968.

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

[40] C. E. Langton. *Artificial Life.* Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. VI. Addison Wesley (Redwood City), 1989.

[41] J. Lynch. Antichaos in a class of random boolean cellular automata. *Physica D*, 69:201–208, 1993.

[42] J. Lynch. A criterion for stability in random boolean cellular automata. *Ulam Quarterly*, 2:32–44, 1993.

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

[44] J. Maynard-Smith. Evolution and the theory of games. *Cambridge Univ. press: Cambridge U.K.*, 1982.

[45] J. Maynard-Smith and G. R. Price. The logic of animal conflict. *Nature*, 246:15–16, 1973.

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

[47] 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.

[48] 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.

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

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

[51] 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.

[52] P. Schuster. Optimization and complexity in molecular biology and physics. In P. J. Plath, editor, *Optimal Structures in Heterogenous Reaction Systems*, Synergetics. Springer Verlag, 1989.

[53] P. Schuster. Dynamics of autocatalytical reaction networks. In A. Perelson and S. Kauffman, editors, *Molecular Evolution on Rugged Landscapes*, volume IX, pages 281–306. SFI studies in the sciences of complexity, 1991.

[54] 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.

[55] P. Schuster and P. F. Stadler. Landscapes: Complex optimization problems and biomolecular structures. *Computers Chem.*, 18:295 – 324, 1994.

[56] P. Schuster and J. Swetina. Stationary mutant distributions and evolutionary optimization. *Bull. Math. Biol.*, 50:635, 1988.

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

[58] 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.

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

[60] J. Swetina and P. Schuster. Self-replication with errors – a model for polynucleotide replication. *Biophys.Chem.*, 16:329–345, 1982.

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

[62] 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.

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

[64] 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.

[65] J. Weber, C. Reidys, and P. Schuster. Evolutionary optimization on neutral networks of RNA secondary structures. in preparation, 1995.

[66] S. Wolfram. *Mathematica: a system for doing mathematics by computer.* Addison-Wesley, second edition, 1991.

[67] S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.

[68] S. Wright. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In D. F. Jones, editor, *int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

[69] S. Wright. Random drift and the shifting balance theory of evolution. In K. Kojima, editor, *Mathematical Topics in Population Genetics*, pages 1 – 31. Springer Verlag, Berlin, 1970.

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

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

Tabellarischer Lebenslauf

Name:	Christian Michael Reidys
Anschrift:	Ziegenhainer Str. 115 07745 Jena Telefon: 03641/6458, 6459; E-mail : duck@imb-jena.de
Geburtsdatum:	10.04.1966
Geburtsort:	Lippstadt
Familiestand:	ledig
Staatsangehörigkeit:	deutsch
Grundschule:	August 1972 - Juli 1976 Theodor-Heuss Schule, Hamm
Gymnasium:	August 1976 - Mai 1985 Hammonense Gymnasium, Hamm
Schulabschluss:	Allgemeine Hochschulreife, Zeugnis vom 13. Mai 1985
Zivildienst:	Juli 1985 - Februar 1987
Studium:	April 1987 - Dezember 1991 Ruprechts-Karls-Universität Heidelberg, Diplomstudiengang Mathematik mit Nebenfach theoretische Physik

Arbeitsverhältnisse:	Oktober 88 - März 90 Mathematik Korrektor für die Fernuniversität Hagen;
	Oktober 89 - Februar 93 wiss. Mitarbeit am Mathematikinstitut Heidelberg;
	Seit April 93 wiss. Mitarbeiter am IMB in der AG Schuster für Molekulare Evolutionsbiologie

Abschlussprüfung:	Diplom Mathematik, Dez. 1991
Examensnote:	sehr gut
Dissertation:	"Neutrale Netze von RNA Sekundärstrukturen", von April 1993 bis März 1995
Hobbys:	Schwimmen als Wettkampfsport, Surfen und Schach
Sonstige Kenntnisse:	Englisch Französisch (Fachsprache Mathematik)