Dynamics of Neutral Evolution

A case study on RNA secondary structures

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von

Diplom-Mathematikerin Jacqueline Weber
geboren am 5. September 1969 in Bergen

Gutachter

1. Prof. Dr. P. Schuster
2. Prof. Dr. K.P. Hadeler
3. PD Dr. habil. G. Jetschke

Tag des Rigorosums: 16. Dezember 1996

Tag der öffentlichen Verteidigung : 21. Januar 1997

Danksagung

An dieser Stelle möchte ich allen danken, die mir das Schreiben, der hier vorliegenden Arbeit, ermöglicht haben. Mein besonderer Dank gilt dabei

Prof. Dr. Peter Schuster für die Überlassung des Themas, seine Betreuung sowie für die Bereitstellung ausgezeichneter Arbeitsbedingungen am Institut für Molekulare Biotechnologie in Jena.

Dr. Christian Reidys für viele anregende wissenschaftliche Diskussionen und seine unermüdliche Hilfsbereitschaft.

Der gesamten Arbeitsgruppe ,,Molekulare Evolutionstheorie" – *Dr. Christian Forst, Stephan Kopp, Ulrike Göbel und Janosch Palinkas* – für die angenehme Arbeitsatmosphäre.

Dr. habil. Gottfried Jetschke für Unterstützung und Rat sowie die gemeinsamen Seminare mit seiner Arbeitsgruppe ,,Theoretische Ökologie".

Meinen Eltern für alles, was sie seit 27 Jahren für mich getan und mir gegeben haben.

Und *meinem Michel* ... .

Zusammenfassung

Das Modell der Faltung von RNA-Sequenzen konstanter Kettenlänge in Sekundärstrukturen minimaler freier Energie wird verwendet, um Effekte neutraler Evolution in endlichen Populationen zu studieren. Sequenzen (Genotypen), welche die gleiche Struktur (Phänotyp) ausbilden, werden als neutral bezeichnet. Der Kompatibilitätsbegriff bezüglich einer Struktur ist zentral für die Beschreibung der Sequenz–Sekundärstruktur–Abbildung. In der Menge kompatibler Sequenzen formen neutrale Sequenzen ausgedehnte Netzwerke, die in der Terminologie der Zufallsgraphen beschrieben werden können. Von eminenter Bedeutung für die evolutive Adaptation ist die Tatsache, daß Schnittmengen kompatibler Sequenzen für Paare von Sekundärstrukturen stets nichtleer sind und damit Sequenzen existieren, die Übergänge zu anderen Strukturen ermöglichen.

Fehlerhafte Replikation (Vererbung) unter Verwendung eines uniformen Fehlermodells sowie ein unspezifischer Verdünnungsfluß (Selektion) bilden die Grundlagen der in dieser Arbeit betrachteten Evolutionsprozesse. Drei Modelle werden untersucht, denen verschiedene Fitnesslandschaften und unterschiedlichen Anzahlen von ausgezeichneten Phänotypen zugrundeliegen. Das erste Modell betrachtet eine vollständig flache Landschaft und zwei fiktive Phänotypen. Eine Rate w wird eingeführt, die als mittlere Wahrscheinlichkeit interpretiert werden kann, bei einer Replikation eine Kopie zu erhalten, die der gleichen Phänotypgruppe angehört wie ihr Template. Sie wird phenotypische Fixationswahrscheinlichkeit genannt. Abhängig vom Wert w ändert sich der Charakter des evolutiven Prozesses. Das zweite Modell betrachtet ein ausgewähltes neutrales Netzwerk als ausgezeichneten und den Rest des Sequenzraumes als weniger fitten zweiten Phänotyp. Die Interpretation im Sinne der Zufallsgraphen erlaubt die Ableitung von Raten, die die Kopplung der beiden Phänotypen beschreiben. Auch diese unterliegen kritischen Werten, die das evolutive Verhalten entscheidend beeinflussen. Das dritte Modell, in dem zwei neutrale Netzwerke einen größeren Fitnesswert besitzen als der Rest des Sequenzraumes, gibt Einblick in die Mechanismen der neutralen Evolution und zeigt, daß neben Koexistenz beider und Aussterben eines Netzwerkes ein drittes Phänomen aufkommt, das durch abwechselnde Fixierung der beiden Strukturen in der Population gekennzeichnet ist.

1. Introduction

The understanding of the biological world surrounding us has undergone dramatic changes during the last century. Four concepts leading from the view of fixed, unchanging species to contemporary biology can be formulated (i) Darwin's idea of evolution and natural selection [7], (ii) the basic laws of heredity discovered by Mendel [51], (iii) Weismann's concept of the germ plasm from which organisms grow [78], and (iv) population genetics theory that provides insight into behavior of genes and population attributes under natural selection [5, 18].

The comprehension of the molecular basis and the mechanism of heredity has vastly increased since the discovery of the chemical structure of DNA by Watson and Crick in 1953 [77]. The fact that biological information is stored in genotypes (sequences of nucleotides) and that the mechanism of heredity is error prone gave rise to new aspects in evolutionary biology. Comparative studies on DNA/RNA and thus on protein sequences have exposed a high variability of nucleotide compositions in populations. In order to explain this effect a large degree of neutrality in these alterations was claimed. In view of nucleotide substitutions that do not have any effect on the protein sequence, and amino acid substitutions that result in functionally equivalent products (polymorphism) there are no doubts left about the existence of neutrality. Moreover it is surprising how little conservation there is at the sequence level [55].

The importance of non-adaptive gene interactions and of random drift of gene frequencies in finite populations was already stressed by Wright [79]. But the prime argument [43] leading to the *neutral theory* was that the amount of genetic substitutions estimated to have taken place during evolution could not be explained by processes caused by natural selection. It was claimed that the genetic load, i.e., the difference of maximum and average fitness value in a population scaled by the maximum value, that selective substitutions would imply would be tremendous [44, 45].

> ". . . a species consisting of a half million individuals, . . . even equating one year with one generation, the load per generation is roughly 30. This means that to maintain

> the same population number and still carry out mutant substitutions at the rate of one substitution every two years each parent must leave $e^{15} \approx 3.27 \times 10^6$ offspring for one offspring to survive and reproduce. "

Wright saw genetic drift merely as a process that could improve the "evolutionary search capacity" whereas Kimura proposed that the majority of evolutionary changes at the molecular level are caused by random drift and random fixation of selectively neutral or nearly neutral mutants on the level of genotypes rather than by positive Darwinian selection. As often documented in history the introduction of new theories usually splits people into at least two groups strictly supporting their point of view. By the same reason controversies between selectionists and neutralists arised. Often the neutral theory is called "non-Darwinian". Kimura himself has emphasized that his theory is not antagonistic to the view on evolution of form and function guided by Darwinian selection, but it brings a new and non-negligible aspect into discussion. It is really worth noticing that the greatest scholar of evolution, Charles Darwin, already saw the role of neutral variants [7]:

> "... This preservation of favorable individual differences and variations, and the destruction of those which are injurious, I have called Natural Selection, or the Survival of the Fittest. Variations neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, ..., or would ultimately become fixed, owing to the nature of the organism and the nature of the condition."

This clear recognition of selective neutrality and its consequences in evolution by Darwin is remarkable. What he could not be aware of are the extent of neutrality detected in molecular evolution [44] and the positive role it possibly plays in supporting adaptive selection through random drift.

Initiated by Kimura the effects of neutrality have been becoming an object of study in population genetics [5]. With this preparatory work and the arguments stated above we believe that the investigation of neutral evolution on the molecular level by making use of mathematical tools is an interesting field to deal with. Following the lines of Darwin and Kimura each evolutionary process carries dynamical and stochastic elements. Thus we claim that the theory of stochastic processes provides the proper mathematical framework of description. Each evolutionary path must be seen as a realization of one stochastic process where dynamics is caused by reproduction and natural selection whereas stochastic aspects enter by finite size of population, mutations that randomly introduce new genotypes and random extinction of phenotypes. Therefore a model needs to be set up that describes the behavior of a population under these circumstances.

Even at molecular level evolutionary processes are as complex as on the level of population genetics and population support dynamics. The decisive contribution that mathematics is able to make is that of providing models based on inherently drastic reductions. That means we can not expect that Nature necessarily follows all closely the models we construct but hope that they are adequately correct. To view evolutionary processes we must take into account that

1. we neither know the initial conditions nor the course of evolution itself, but only the results present in form of DNA and RNA sequences, and
2. the assignment of genotypes (sequences) to corresponding phenotypes (structure or function) is far from being understood and still a major task of current biological research.

Initially it is known that point mutations (single nucleotide exchanges in DNA) may have all kinds of effects ranging from drastic change in properties and functions to no change at all. The fact that point mutations do not alter the size of a genome allows to restrict ourselves to considerations of sequences of fixed chain length. The mapping of RNA sequences of chain length n into RNA secondary structures provides a powerful tool accessible to mathematical description. A convincing example that RNA secondary structures are worth looking at is the conservation of the tRNA clover leaf structure [35, 75]. Another supporting argument for the relevance of RNA secondary structures is provided by experiments performed on self replicating RNA molecules by Biebricher *et al.* [2]. Ma and Mathews [47] have shown that the inhibitory function of a small RNA preventing the activation of a special protein kinase depends on its structure.

RNA secondary structures are often defined in the context of minimum free energy structure fulfilling common thermodynamic condition of a molecular ground state. It has been shown that in case of small chain length minimum free energy structures match natural realized structures for RNA sequences quite well [6]. The relation between RNA sequences and their minimum free energy secondary structures have been studied intensively by exhaustive folding [30, 31] of all **GC** and **AU** sequences up to chain length 30. Sequences that adopt the same structure, i.e., they are neutral with respect to this structure, form extended networks in sequence space and facilitate optimization through adaptive walks and random drift in the absence of more fit genotypes. The evaluation of RNA secondary structures generates a toy landscape that is ideally suited to study neutral evolution. On landscapes of this type evolutionary optimization follows a combined mechanism of adaptive walk leading to minor peaks and random drift allowing to escape from evolutionary traps [62].

This thesis is devoted to study of the phenomenon of neutral evolution. Neutral evolution will be investigated by making use of the landscape provided by RNA sequences as genotypes and their corresponding secondary structures as phenotypes. In order to give a detailed analysis of the investigated dynamical models we pursue a combination of a detailed mathematical representation including analytical results and comparative computer simulations.

Following Dobzhansky [11] who stated: ”Nothing in biology makes sense except in the light of evolution” a brief view on evolutionary dynamics in biology as a sophisticated and complex phenomenon is given in chapter 2.

In chapter 3 we develop the mathematical framework for studying RNA folding landscapes. The concepts of compatible sequences, and neutral networks as randomly induced subgraphs are introduced. Section 3.3 is concerned with the geometric relationship between neutral networks in sequence space.

Chapter 4 deals with evolutionary dynamics. In order to work out aspects of neutral evolution finite populations of erroneously replicating strings are investigated in landscapes induced by neutral networks. Networks of predefined RNA secondary structures are either formed by sequences adopting this structure under minimum free energy conditions or they are constructed using the random graph model introduced in section 3.

Three landscapes are formulated for different numbers of specified phenotypes (secondary structures). We are interested in the fractions of a population covering particular neutral networks. The theory of stochastic processes is used to derive analytical expressions for stationary distributions. Critical phenotypic fixation probabilities are detected which alter the modalities of the distributions and affect the character of the evolutionary course.

In chapter 5 a number of computer experiments are performed in order to test the analytical expressions derived in section 4.3. For one pair of secondary structures two simulations are performed firstly using minimum free energy neutral networks and secondly randomly constructed neutral networks. Both evolutionary trajectories are compared with respect to the applicability of the random graph ansatz to the RNA folding model. The results of another computer experiment under minimum free energy conditions are used to demonstrate and visualize the mechanism of evolutionary searching.

Finally in chapter 6 a detailed discussion of the results obtained so far is presented.

2. Evolutionary Dynamics

Current biology is facing a grand synthesis of knowledge from three different disciplines: molecular biology, developmental biology, and evolutionary biology. A first step in this direction was already taken in the late sixties by the pioneering works of Sol Spiegelman [70]. At about the same time Manfred Eigen [12] conceived a kinetic theory of evolution at the molecular level. Since then the study of the evolution of molecules in laboratory systems has become a research area of its own.

The minimal conditions for a process of self-organization caused by natural selection were proposed by Dawkins [8] in terms of his *replicator concept*. He argued that units (replicators) which are capable of reproduction, inheritance of "genetic" information allowing variability, and interaction causing survival of replicators to be fitness dependent, will undergo evolution by natural selection.

Experiments with replicating molecules in the test tube furnished proof that Charles Darwin's principle of variation and natural selection is not a privilege of cellular life [1]: optimization of properties related to the *fitness* of replicating molecules is observed readily *in vitro* with *naked* ribonucleic acid (RNA) molecules in evolution experiments. Evolution characterized as Darwinian dynamics is often visualized as a hill climbing process on a "surface of selective value" [79] or a *fitness landscape* that assigns fitness values to genotypes or polynucleotide sequences (DNA or RNA). In terms of dynamical systems theory Darwinian dynamics is simple in the sense that it follows a gradient and eventually reaches a (local or global) fitness maximum. Given that the structure of a fitness landscape determines the evolutionary process, Darwinian dynamics was found to be just one feature of evolutionary systems. Others being, for example, suppression of optimization of individual fitness by (mutualistic) interaction through catalysis, predator prey or host parasite interactions. Evolutionary optimization needs not approach a steady state but may give rise to complex dynamical phenomena like oscillations, spatial pattern formation, deterministic chaos in space and time. Spatiotemporal patterns cover only one aspect of evolutionary phenomena - others being, for example, historical like the reconstruction of phylogenies or genetic like the fixation of alleles in populations [37]. A comprehensive model was introduced by Schuster *et al.* [67] that tries to cover most of these relevant features.

Following [67] there are three different abstract metric spaces being appropriate for illustrative projections of different complex evolutionary scenarios in order to elevate evolutionary aspects of interest:

1. the *sequence space* of genotypes being DNA or RNA sequences,
2. the *shape space* of phenotypes, and
3. the *concentration space* of biochemical reaction kinetics.

The sequence space is a metric space containing all sequences. The metric is given by the Hamming distance [33], that counts the number of positions in which two sequences differ. The sequence space of binary sequences is a hypercube of dimension n where n is the chain length of the genotype. In figure 1 the binary hypercube of dimension three is shown as well as an example for the neighborhood relation in the generalized hypercube of dimension two formed over the four letter alphabet **AUGC**.

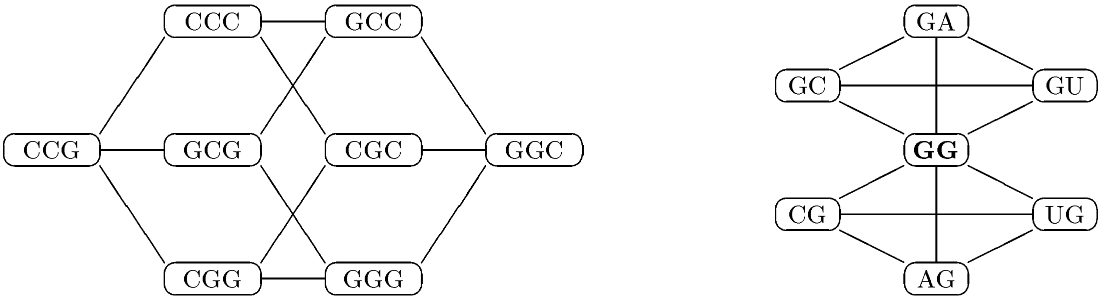


Figure 1: Sequence space. Sequences are presented by points in sequence space. Edges connect sequences of Hamming distance one, i.e. sequences that differ in a single position. The left part shows the binary hypercube of dimension 3. On the right hand side the one error class of sequence **GG** in the generalized hypercube formed over the four letter alphabet **AUGC** is presented.

The shape space is an abstract space covering all possible phenotypes under consideration. Phenotypes are formed by processing genotypes in a given context. Like sequence space, it is a metric space. But the definition of a meaningful distance in shape space is a task requiring information on properties and functions of interest: meaningful comparisons of active sites of enzymes require atomic resolution whereas studies of phylogenetic conservation of structures can be done much better on the coarse–grained level of *ribbon* or *wire* diagrams.

Concentration space, finally, is the conventional space in which chemical reaction kinetics or changes in populations take place. It is the space chemists and population geneticists are familiar with. Concentration space is restricted to the classes of genotypes actually present (support). It was formalized and put into precise mathematical terms by Feinberg [19].

Biological evolution is a phenomenon of high complexity. It can be understood and modeled more easily if it is partitioned into three simpler processes, each of them highlighting one particular aspect of evolution.

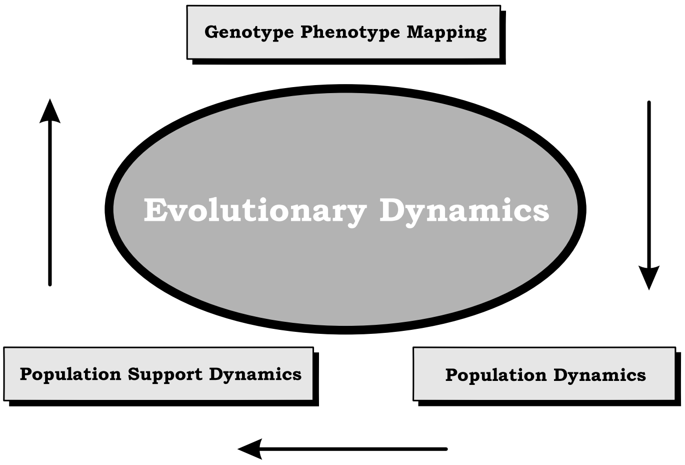


Figure 2: Evolutionary dynamics. The dynamics of evolution is partitioned into three simpler processes: (i) genotype-phenotype mapping, (i) population dynamics, and (iii) population support dynamics.

These three processes *genotype-phenotype mapping, population dynamics* and *support dynamics* (figure 2) are best visualized in the three metric spaces mentioned above.

The genotype phenotype mapping assigns a particular phenotype to a genotype. In molecular evolution this is tantamount to folding biopolymer sequences into structures. Accordingly the *shape space* is predestined to be used for the description. Precisely, it is the mapping from sequences into structures and further into functions that is relevant for evolution: genotype phenotype mapping provides the (kinetic and thermodynamic) parameters which enter population dynamics.

Population dynamics describes temporal evolution of the population variables (particle numbers, genotype frequencies, concentrations, etc.). It is properly described in the conventional *concentration space* of chemical reaction kinetics. The number of possible genotypes is huge and thus the majority of them will neither be materialized in an evolution experiment nor in nature. Only a small subset of all possible genotypes will be present at a given instant in the population. Whenever a new variant is coming into being by mutation or some genotype dies out the concentration space changes, a new variable describing the frequency of the new variant is added or the obsolete variable is removed, respectively.

Population support dynamics is a process to be visualized in *sequence space* since it deals with migrating sets of genotypes. Moreover, it provides the input for genotype phenotype mapping as it defines the areas in sequence space where new genotypes appear.

The three projections of evolutionary dynamics onto the three abstract spaces form a conceptual cycle in the sense that each of the three processes provides the input for the next one: genotype phenotype mapping provides the parameters (fitness, for example, being the most important of them) for population dynamics. Population dynamics deals with temporal alterations in concentrations and thus hands the information on arriving new and disappearing old genotypes over to the support dynamics. Support dynamics in turn transfers changes in the population support to genotype phenotype mapping in order to make the new parameters available for population dynamics.

3. Neutral Networks of RNA Secondary Structures

Conventional biophysics considers sequence structure relations of biopolymers primarily with respect to the folding problem: given a sequence; which structure does is form under specified experimental conditions. Such a condition is for example the thermodynamic equilibrium of minimum free energy structures. One biological important landscape that has received special attention during the last few years is induced by the "folding" of polynucleotides (RNA). While a prediction of true 3D structures is far beyond the possibilities of present-day computers, secondary structures, which are defined as a list of base pairs in the molecules, are readily accessible. A large body of computational data has been published [4, 21, 22, 25, 63, 73] on this example of a sequence to structure mapping. In fact secondary structures can be regarded as a simplified version of RNA phenotypes and at present this allows the most realistic modeling of genotype phenotype relations.

The evolution of RNA molecules in replication assays, viroids, and RNA viruses can be viewed as an adaptation process on a fitness landscape in the sense of Wright's imagination [79]. The scalar entities to the landscape are for example minimum free energies of secondary structure formations [21] or functional defined properties of the structure itself.

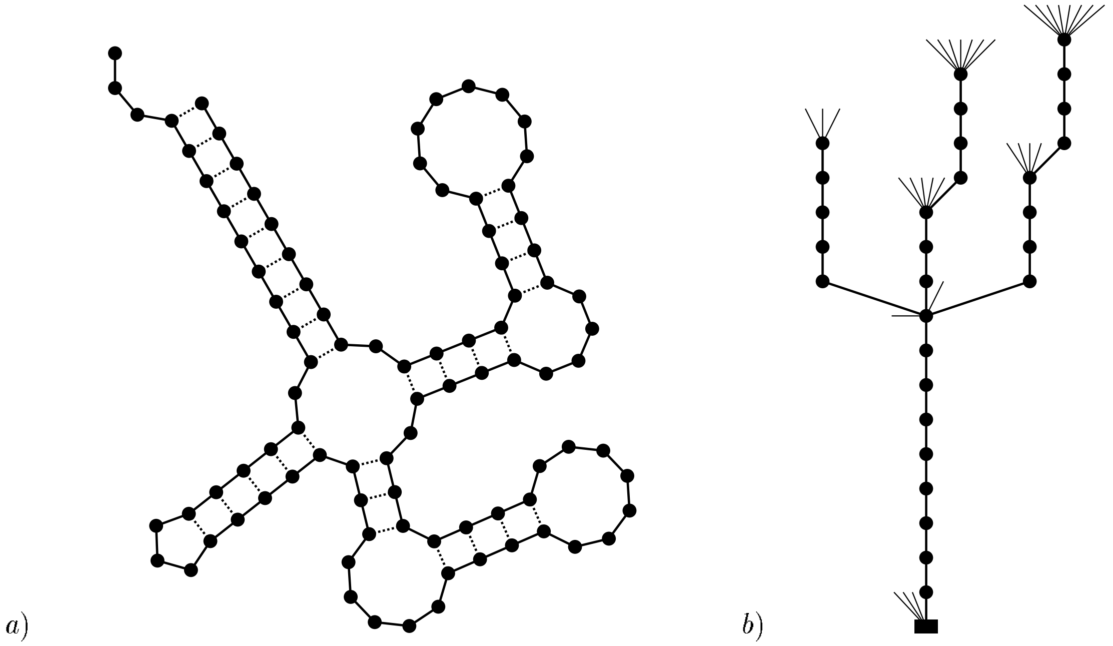
3.1. RNA Secondary Structures and Compatible Sequences

First we give the precise mathematical definition of a secondary structure. Using the terminology of graph theory [3] an RNA secondary structure is defined as a vertex labeled graph on n vertices with an adjacency matrix A fulfilling [76].

(1) $a_{i,i+1} = 1$ for $1 \leq i \leq n-1$;
(2) For each i there is at most one $k \neq i-1, i+1$ such that $a_{i,k} = 1$;
(3) If $a_{i,j} = a_{k,l} = 1$ and $i < k < j$ then $i < l < j$.

An edge (i,k), $|i-k| \neq 1$ is called a *bond* or *base pair*. A vertex i connected only to $i-1$ and $i+1$ it named *unpaired*. We will 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. Note that $n_u(s) + 2\,n_p(s) = n$ is the chain length of the molecule.

There are various but equivalent representations for RNA secondary structures (figure 3):



c) ...(((((((((.(((((...)))))(((.....((((.......))))))).((((....((((........)))))))).)))))))))

Figure 3: Different representations of one RNA secondary structure. Notation a) is common in biology and shows the secondary structure as a planar graph, b) is the corresponding tree, and c) gives the the linear string encoding for the structure.

a) In biology RNA secondary structures are commonly drawn as planar *secondary structure graphs*.

b) For some considerations it can be useful to translate the secondary structure into a rooted ordered tree by mapping base pairs to internal nodes and unpaired nucleotides to leaves [22, 25].

c) For computer handling the string representation is usually chosen. It is obtained by the following rules: (i) an unpaired vertex k is denoted by $x_k =$'.' whereas (ii) a pair $[i, j]$ with $i < j$ is represented by $x_i =$'(' and $x_j =$')' . Taking into account some biophysical constraints one can easily see that the set of all secondary structure is generated over an alphabet '.()' by the context free grammar: $G : S \rightarrow \text{‘...’}|\text{‘.’}S\,|\,S\text{‘.’}\,|\,SS\,|\,\text{‘(’}S\text{‘)’}$, [38].

d) A more mathematical view on RNA secondary structures interprets them as a permutation. The rules for this mapping are explained in detail in section 3.3 [58].

Similarities and dissimilarities of RNA secondary structures can be expressed by means of quantitative measures with metric properties. These measures depend on the authors' favored representations [46, 69]. Thus they compare the tree representation of RNA secondary structures with tree-editing whereas the compare the string representation with string-compare strategy.

A variety of algorithms [49, 50, 54, 81, 82], and different sets of thermodynamic parameters [26, 60, 74] have been used for the prediction of RNA secondary structures. Fortunately, it has been shown recently [73] that some relevant qualitative features and statistical properties of the sequence-structure mappings are largely independent of algorithm and parameter set. Statistical characteristics of RNA landscapes on the level of secondary structures are accessible by mathematical analysis and computer calculation. RNA landscapes belong to the same class as well known combinatorial optimization problems (traveling salesman problem) and simple spin class systems [63, 65]. The notion of a landscape has been extended to *combinatory maps*, thereby allowing for a direct statistical investigation of the sequence structure relationships of RNA at the level of secondary structures [21, 72]. There are intrinsic properties of this mapping that can be formulated in the following way

1. There are many more sequences than structures.
2. The frequency distribution of structures is sharply peaked. There are few common structures and many rare ones. The distribution follows Zipf's law [80].
3. Sequences folding into the same structure are distributed randomly on the set of "compatible" sequences. There exist neutral paths in sequence space along which structures remain unaffected by mutation.
4. Any desired secondary structure is formed by a sequence that can be found close to an arbitrary initial sequence.

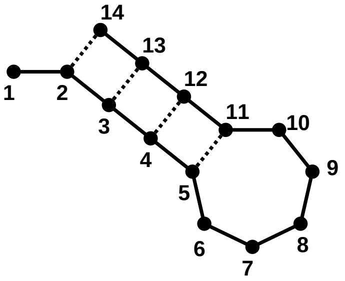
The shape space consisting of all secondary structure graphs provides an example for a level of coarse–graining that except of physical relevance is ideally suited for mathematical modeling.

RNA secondary structures are simple in the sense that they only distinguish unpaired and paired regions. Pairs are formed with respect to the underlying alphabet of nucleotides. Caused by biophysical and biochemical constraints, in general only some bases are able to pair, i.e. in the case of the biophysical alphabet (**A**,**U**,**G**,**C**) admissible pairs are only (**AU**,**UA**,**GC**,**CG**,**GU**,**UG**). Hence for an alphabet $\mathcal{A}$ a *pairing rule* Π on $\mathcal{A}$ is given as a set of pairs $\{[x,y]\} \subseteq \mathcal{A} \times \mathcal{A}$, such that $[x,y] \in \Pi$ implies $[y,x] \in \Pi$.

At this point we proceed with an analysis of RNA secondary structures that is free from chemical or physical restrictions. We will consider secondary structure over arbitrary alphabets with arbitrary pairing rules. Each secondary structure s is determined by the set of *contacts* of s:

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

A sequence x is said to be *compatible* to a structure s if and only if the nucleotides x_i and x_j form an admissible pair i.e. $[x_i, x_j] \in \Pi$, for each base pair $[i,j] \in \Pi(s)$. $\mathbf{C}[s]$ is the set of all sequences that are compatible with the structure s.



$$(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14})$$

$$\Downarrow$$

$$(x_1, x_6, x_7, x_8, x_9, x_{10}) \times ((x_2, x_{14}), (x_3, x_{13}), (x_4, x_{12}), (x_5, x_{11}))$$

Bases: $\alpha = 4$ (**A**,**U**,**G**,**C**)
Base pairs: $\beta = 6$ (**AU**,**UA**,**GC**,**CG**,**GU**,**UG**)

Figure 4: Partition of a sequence according to a structure. Natural RNA sequences are assembled from a four letter alphabet and form six base pairs

Remark: For the size of a compatible set we obtain $|\mathbf{C}[s]| = \alpha^{n_u}\,\beta^{n_p}$ (We denote the size of the alphabet $\mathcal{A}$ by α and the number of distinct base pairs by β).

Consider the sequence to structure mapping $f_n : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$, where $\mathcal{Q}_\alpha^n$ denotes the generalized hypercube of dimension n over an alphabet of size α. We know *a priori* that the preimage $f_n^{-1}(s)$ which consists of all sequences adopting the secondary structure s is contained in the set of *compatible sequences* $\mathbf{C}[s]$. We call the sequences $x \in f_n^{-1}(s)$ *neutral* with respect to s. For any sequence $x \in f_n^{-1}(s)$ *neutral neighbors* of x are those sequences $x' \in f_n^{-1}(s)$ in the hypercube $\mathcal{Q}_\alpha^n$ that differ in exactly one base. Consequently $\mathbf{C}[s]$ could be provided with a graph-structure in which in particular sequences of Hamming distance 1 are adjacent. However, for RNA secondary structures all neutral neighbors of a sequence x are located in the set $\mathbf{C}[f_n(x)]$. Unfortunately, the induced subgraph $\mathcal{Q}_\alpha^n[\mathbf{C}[f_n(x)]]$ is not connected – it decomposes into 'hyper-planes' defined by a particular choice of the base pairs. 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.) Therefore we introduce the graph $\mathcal{C}[s]$:

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

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

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 $[i, j] \in s$.

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 one and the same graph of compatible sequences.

3.2. Neutral Networks as Randomly Induced Subgraphs

In order to facilitate the understanding of the following section we shall state some basic facts of graph theory [3]:

- A *graph* is an ordered pair consisting of a *vertex set* v$[G]$ and an *edge set* e$[G]$.
- An edge is defined by an ordered or unordered pair of vertices. For our purposes we consider *undirected* graphs whose edges are *unordered* tuples.
- Two vertices v and v' are called *adjacent* if and only if $\{v, v'\} \in \mathrm{e}[G]$.
- A graph is called *finite* if and only if v$[G]$ and e$[G]$ is finite.
- A graph G' is a *subgraph* of G, if $\mathrm{v}[G'] \subseteq \mathrm{v}[G]$ and $\mathrm{e}[G'] \subseteq \mathrm{e}[G]$.
- Let $X \subset \mathrm{v}[G]$. The *induced subgraph* of X in G, $G[X]$, has the vertex set $\mathrm{v}[G[X]] = X$ and the edge set $\mathrm{e}[G[X]] \overset{\mathrm{def}}{=} \{\{v, v'\} \mid v, v' \in X \text{ and } \{v, v'\} \in \mathrm{v}[G]\}$.
- Two vertices v and v' are *connected* in G if a series of vertices $(v = v_1, v_2, \ldots, v_m = v')$ can be formed such that $\{v_i, v_{i+1}\} \in \mathrm{e}[G]$ for $i = 1, \ldots, m - 1$.

Suppose $f_n : \mathcal{Q}_\alpha^n \to \mathcal{S}_n$ is a prescribed sequence to structure mapping and $s \in \mathcal{S}_n$ a fixed RNA secondary structure then following Reidys *et al.* [59, 56] 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.,

$$\Gamma_n[s] \overset{\mathrm{def}}{=} \mathcal{C}[s][f_n^{-1}(s)] \quad .$$

It can be shown that neutral networks constructed as random graphs are in a certain sense dense and connected with probability one. In other words a "topology" of the network can be expected that is ideally suited for evolutionary adaption.

The major concepts entering into the construction of neutral networks have been established by Reidys *et al.* [56, 59]. We recall them briefly and state the results without proofs.

Let H be a finite graph. Then the set of all induced subgraphs in H is denoted by $\mathcal{G}(H)$. Further suppose $\lambda \in \mathbb{R}$ with $0 \leq \lambda \leq 1$ to be given and set for $G \in \mathcal{G}(H)$

$$\boldsymbol{\mu}_\lambda(G) \overset{\mathrm{def}}{=} \lambda^{|\mathrm{v}[G]|}(1 - \lambda)^{|\mathrm{v}[H]| - |\mathrm{v}[G]|} \quad .$$

Then evidently $\boldsymbol{\mu}_\lambda$ is a probability measure and together with $\mathcal{G}(H)$ we obtain a probability space

$$\Omega \overset{\text{def}}{=\!=} (\mathcal{G}(H), \mathcal{P}(\mathcal{G}(H)), \boldsymbol{\mu}_\lambda)$$

where $\mathcal{P}(X)$ denotes the power set of a set X.

Remark: Each graph $G \in \mathcal{G}(H)$ can be constructed by selecting each vertex $v \in \mathrm{v}[H]$ with the independent probability $0 \leq \lambda \leq 1$. This yields a set V_λ of vertices. Then G is the induced subgraph of V_λ in H, i.e., $G = H[V_\lambda]$, and G is called randomly induced subgraph

Let $G < H$ be a subgraph of H then the *boundary* of G is defined by

$$\partial G \overset{\text{def}}{=\!=} \{v \in \mathrm{v}[H] \setminus \mathrm{v}[G] \mid \exists v' \in \mathrm{v}[G] : (v, v') \in \mathrm{e}[H]\}$$

and the *closure* of G in H is given by $\overline{G} \overset{\text{def}}{=\!=} H[\partial G \cup \mathrm{v}[G]]$. A subgraph $G < H$ is called *dense* in H if and only if $\overline{G} = H$. Then for randomly induced subgraphs on generalized hypercubes mathematical explorations result in the following theorems.

Theorem 1. *Let $(Q_\alpha^n)_n$ be a sequence of generalized hypercubes and $\Gamma_n < Q_\alpha^n$ a randomly induced subgraph. Suppose $\lambda^* \overset{\text{def}}{=\!=} 1 - \sqrt[1-\alpha]{\alpha}$ then*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is dense}\} = \begin{cases} 1 & \text{for} \quad \lambda > \lambda^* \\ 0 & \text{for} \quad \lambda < \lambda^* \end{cases}$$

Theorem 2. *Let $(Q_\alpha^n)_n$ be a sequence of generalized hypercubes and $\Gamma_n < Q_\alpha^n$ randomly induced subgraphs. Suppose $\lambda^* \overset{\text{def}}{=\!=} 1 - \sqrt[1-\alpha]{\alpha}$ then*

$$\lim_{n\to\infty} \boldsymbol{\mu}_n\{\Gamma_n \text{ is connected}\} = \begin{cases} 1 & \text{for} \quad \lambda > \lambda^* \\ 0 & \text{for} \quad \lambda < \lambda^* \end{cases}$$

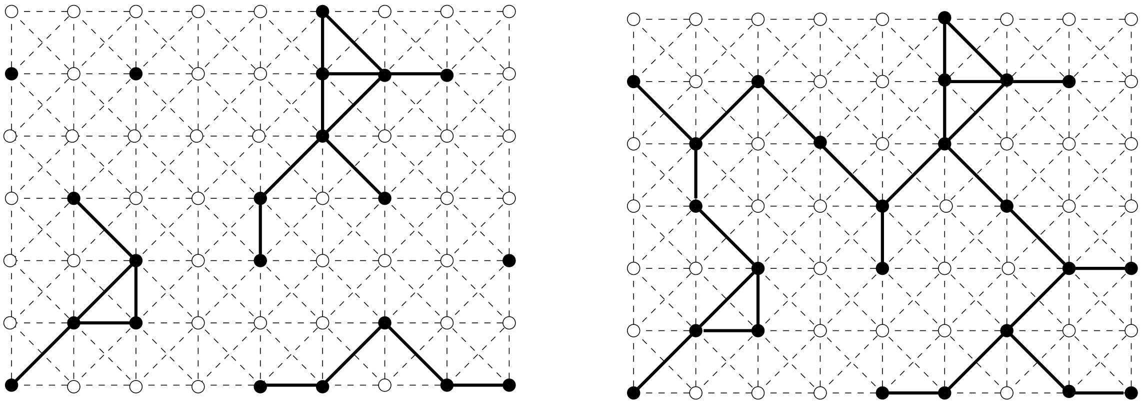


Figure 5: Connectivity and density of subgraphs. The vertices of the subgraph are indicated by black nodes. A couple of vertices is connected (solid line) if and only if it is connected in the underlying graph. On the left we present a graph that is dense but not connected at all. On the right side a dense and connected subgraph is shown.

Remark: An induced subgraph Γ_n whose vertex set is selected by an uniform probability λ from the vertex set of the generalized hypercube $\mathcal{Q}_\alpha^n$ is for $\lambda > 1 - \sqrt[1-\alpha]{\alpha}$ and infinite chain length almost surely dense and connected and almost surely non-dense and disconnected for $\lambda < 1 - \sqrt[1-\alpha]{\alpha}$.

In figure 5 we give two distinct examples of subgraphs in order to demonstrate the importance of density and connectivity. Note that although both properties for infinite chain length use to have the same critical probability λ, for finite chain length a subgraph does not need to comply both attributes.

Let us now return to RNA secondary structures. We proceed by considering randomly induced subgraphs of the graph product $\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ that are induced by certain subsets of vertices. *A priori* there is no reason why the probability of being ”neutral neighbor”[1] should be the same for both single base and base pair mutations. Therefore we construct randomly induced subgraphs in each of the generalized hypercubes $\mathcal{Q}_\alpha^{n_u}$, $\mathcal{Q}_\beta^{n_p}$.

Model I: *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* $G < \mathcal{C}[s]$ *such that*

---

[1] Here we mean by neutral neighbor an RNA sequence differing in exactly one base that has the same RNA secondary structure as the reference sequence.

$G = \mathcal{C}[s][V]$ is an induced subgraph of $\mathcal{C}[s]$ with $V \subset \mathrm{v}[\mathcal{C}[s]]$. Each set $V \subset \mathrm{v}[\mathcal{C}[s]]$ will be constructed in the following way

- *For a vertex $v = (v_u, v_p)$ with $v_u \in \mathcal{Q}_\alpha^{n_u}$ and $v_p \in \mathcal{Q}_\beta^{n_p}$ the coordinate v_u is selected with probability λ_u and v_p with probability λ_p.*
- *Finally a vertex $v = (v_u, v_p)$ is selected if either v_u or v_p have been chosen.*

Writing $\chi_{u,p} = \lambda_u + \lambda_p - \lambda_u\lambda_p$ we set

$$\mu_{n,\lambda_u,\lambda_p}(G) \overset{\text{def}}{=} \chi_{u,p}^{|\mathrm{v}[G]|}(1 - \chi_{u,p})^{\alpha^{n_u}\beta^{n_p} - |\mathrm{v}[G]|}$$

where $\chi_{u,p}$ is a probability measure on the set of all induced subgraphs. Then we define a neutral network $\Gamma_n^{\mathrm{I}}[s] < \mathcal{C}[s]$ to be an induced random subgraph of $\mathcal{C}[s]$ with underlying measure $\chi_{u,p}$.

Before we proceed with the fundamental result on model I we introduce some terminology:

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 Φ_x^Γ, Φ_y^Γ of Γ in $G_1 \times G_2$ are the induced subgraphs in G_2 and G_1:

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

It was shown in [56] that for randomly induced subgraphs $\Gamma_n^{\mathrm{I}}[s] < \mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ each fiber $\Phi_{v_2}^{\Gamma_n^{\mathrm{I}}[s]}$ is isomorphic to a random graph Γ_{n_u} and accordingly $\Phi_{v_1}^{\Gamma_n^{\mathrm{I}}[s]}$ is isomorphic to Γ_{n_p}. In other words we have

$$\forall (v_1, v_2) \in \mathrm{v}[\Gamma_n^{\mathrm{I}}[s]] : \quad \Phi_{v_2}^{\Gamma_n^{\mathrm{I}}[s]} \cong \Gamma_{n_u} \quad \text{and} \quad \Phi_{v_1}^{\Gamma_n^{\mathrm{I}}[s]} \cong \Gamma_{n_p} \,.$$

This finally leads to a sufficient criterion for density and connectivity of randomly induced subgraphs following model I.

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

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

In Model I point mutations and pair mutations are assumed to be completely uncorrelated. To introduce the other extreme model II will be developed.

Model II: *Let s be a secondary structure with corresponding graph of compatible sequences* $\mathcal{C}[s] = \mathcal{Q}^{n_u}_\alpha \times \mathcal{Q}^{n_p}_\beta$. *Subgraphs* Γ_{n_u} *and* Γ_{n_p} *are constructed randomly. Therefore a vertex set* V_{λ_u} *is created by selecting each vertex in* $\mathcal{Q}^{n_u}_\alpha$ *with probability* λ_u *and* V_{λ_u} *by selecting each vertex in* $\mathcal{Q}^{n_p}_\beta$ *with* λ_p. *Then* $\Gamma_{n_u} = \mathcal{Q}^{n_u}_\alpha[V_{\lambda_u}]$ *and* $\Gamma_{n_p} = \mathcal{Q}^{n_p}_\beta[V_{\lambda_p}]$. *Finally a neutral network* $\Gamma^{\mathrm{II}}_n[s] < \mathcal{C}[s]$ *is defined to be* $\Gamma^{\mathrm{II}}_n[s] \overset{\mathrm{def}}{=} \Gamma_{n_u} \times \Gamma_{n_p}$. *This is equal to* $\Gamma^{\mathrm{II}}_n[s] = \mathcal{C}[s][V_{\lambda_u} \times V_{\lambda_p}]$.

All subgraphs constructed in this way form a probability space with probability measure

$$\boldsymbol{\mu}_{\lambda_u,\lambda_p}(\Gamma^{\mathrm{II}}_n[s]) \overset{\mathrm{def}}{=} \boldsymbol{\mu}_{n_u,\lambda_u}(\Gamma_{n_u}) \times \boldsymbol{\mu}_{n_p,\lambda_p}(\Gamma_{n_p})\,.$$

where $\boldsymbol{\mu}_{n_u,\lambda_u}(\Gamma_{n_u}) = \lambda_u^{|\mathrm{v}[\Gamma_{n_u}]|}(1-\lambda_u)^{\mathrm{v}[\alpha^{n_u} - |\mathrm{v}[\Gamma_{n_u}]|}$ *and* $\boldsymbol{\mu}_{n_p,\lambda_p}$ *respectively.* We have $\Phi^{\Gamma^{\mathrm{II}}_n[s]}_{v_1} \cong \Gamma_{n_p}$ and $\Phi^{\Gamma^{\mathrm{II}}_n[s]}_{v_2} \cong \Gamma_{n_u}$ where we assume $\boldsymbol{\mu}_{n_u,\lambda_u}, \boldsymbol{\mu}_{n_p,\lambda_p}$ to be the underlying probability measures. The situation can summarized by the following diagram:

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

Since $\Phi^{\Gamma^{\mathrm{II}}_n[s]}_x \cong \Gamma_{n_p}$, $\Phi^{\Gamma^{\mathrm{II}}_n[s]}_y \cong \Gamma_{n_u}$ analogous to model I we derive a criterion for density and connectivity of a neutral networks that are randomly induced subgraphs $\Gamma^{\mathrm{II}}_n[s]$.

Corollary 2 *Suppose* $\lambda_u > 1 - \sqrt[\alpha-1]{\alpha^{-1}}$ *and* $\lambda_p > 1 - \sqrt[\beta-1]{\beta^{-1}}$, *then we have*

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

In order to connect the random graph theory with the combinatory map arising from folding RNA sequences into their secondary structure we need to estimate the fraction of neutral neighbors λ_u and λ_p. In general $\bar{\lambda} \stackrel{\text{def}}{=} |\Gamma_n[s]|/|\mathbf{C}[s]|$ is smaller than $\lambda_u\lambda_p$ since λ_u and λ_p are conditional probabilities, provided a sequence is folding into a specified structure what is the average fraction of neutral neighbors separately in the unpaired and paired regions. The effect is called *buffering* [39]. Therefore we have to estimate the fraction of neutral neighbors in terms of λ_u and λ_p as well as the fraction $\bar{\lambda}$ of neutral sequences in $\mathbf{C}[s]$.

3.3. Two Neutral Networks in Sequence Space

In the previous section a model was envisaged that describes neutral networks corresponding to RNA secondary structures and their intrinsic properties in terms of two parameters λ_u and λ_p. They simply reflect the average fraction of neutral neighbors in unpaired and paired regions of an arbitrary compatible sequence. Thereby we are encouraged to ask for the relationship between two and more neutral networks in sequence space which could play another decisive role in evolutionary adaptation mechanisms. We commence to represent an algebraic view on secondary structures that has been introduced by Reidys *et al.* [58, 59]:

- All RNA secondary structures adopted by sequences of chain length n can be mapped into elements of the symmetric group S_n. In other words an arbitrary secondary structure is equivalent to a permutation of n positions. The corresponding mapping is defined as follows

  $$\pi\,:\,\mathcal{S} \to S_n; \quad s \mapsto \pi(s) \stackrel{\text{def}}{=} \prod_{[i,k]\in\Pi(s)} (i,k)$$

  where (i,k) stands for a transposition in S_n and $[i,k]$ for a base pair of s. By $\Pi(s)$ we denote the set of all base pairs of s. Clearly the map π is an *embedding* of the set

$\mathcal{S}$ of secondary structures in the set S_n of permutations. Moreover it can be verified that $\pi(s)^2 = 1$, i.e., for any secondary structure the permutation assigned by π is an *involution*.

- In a natural way the *embedding* of $\mathcal{S}$ into the symmetric group gives rise to new metrics on the set of secondary structures. All distance functions on S_n can be applied to secondary structures as well. One commonly used is defined by the minimum number of transpositions that are necessary to convert $\pi(s)$ into $\pi(s')$. To our knowledge there is no other metric on RNA secondary structures that is equivalent to this one. However nothing new was found concerning statistical properties [58].
- As already mentioned permutations corresponding to secondary structures are *involutions*. A theorem of group theory [68] states that any two involutions generate a *dihedral group*, D_m. Therefore π in a natural way gives rise to the mapping

$$\jmath : \mathcal{S} \times \mathcal{S} \longrightarrow \{D_m < S_n\}; \qquad (s, s') \mapsto \jmath(s, s') \overset{\text{def}}{=\!=} \langle \pi(s), \pi(s') \rangle \,.$$

As the key result of the above considerations the following theorem can be stated:

Theorem 3. (Intersection–Theorem) *Let s and s' be two arbitrary secondary structures. Then*

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

Example: 1 *The sequence drawn in the middle gives an example for a sequence that is compatible with the structure to its left as well as to its right side.*

`((.((....)).)).` : **AGGAGGAUCCUUUUU** : `.(((((...)).)))`

A primary outcome that can be directly deduced from the existence of sequences being compatible with both structures is that any two neutral networks that are connected and dense come very close in sequence space (Hamming distance ≤ 4) [59]. We define $\mathbf{I}[s, s'] \overset{\text{def}}{=\!=} \mathbf{C}[s] \cap \mathbf{C}[s']$ and call $\mathbf{I}[s, s']$ the *intersection* or the *overlap* of s and s'. As already presumed we now claim that sequences on the *overlap* or at least 'close' to it are potential candidates for transitions between neutral networks of RNA secondary structures in the course of evolutionary adaptation.

Remark: Theorem 3 cannot be extended to three different structures.

If sequences on the intersection are as important as claimed one should properly ask for their number and how they are localized in sequence space. In order to answer these questions we firstly have to deepen our algebraic considerations:

The dihedral group given by $\langle \pi(s), \pi(s') \rangle$ is easily seen to be a *semi-direct product* of the form

$$\langle \pi(s) \circ \pi(s') \rangle \, .$$

The cyclic group $\langle \pi(s) \circ \pi(s') \rangle$ operates on the set of all 'positions' of the sequence $x = (x_1, \ldots, x_n)$. Any permutation induces a *cycle decomposition* and so does the generator $\pi(s) \circ \pi(s')$ of this group, i.e. it can be represented as a minimal product of disjoint cycles $\quad \pi(s) \circ \pi(s') = \prod_{i=1}^{m} \xi_i \quad$ where

$$\xi_{x_\ell} = \{\, (\pi(s) \circ \pi(s'))^i \, (x_\ell) \mid i = 0, ..., k\} \text{ with } k = \max\{j \in \mathbb{N} \mid (\pi(s) \circ \pi(s'))^j (x_\ell) \neq x_\ell \, \} \; .$$

In the sequel we omit the index x_ℓ of a cycle if confusion is not possible and write $\xi = \langle i_1, ..., i_k \rangle$ (where $\pi(s) \circ \pi(s')(i_l) = i_{l+1}$ if $l \leq k-1$ and $\pi(s) \circ \pi(s')(i_k) = i_1$). k is said to be the *length of the cycle* ξ.

Apart from ordering of its elements each cycle $\xi = \langle i_{l_1}, \ldots, i_{l_k} \rangle$ induced by $\pi(s) \circ \pi(s')$ of length greater or equal than two is equivalent to a set $z = (i_1, \ldots, i_k)$ such that $[i_l, i_{l+1}] \in \Pi[s] \cup \Pi[s']$ for $l = 1, \ldots, k-1$. In other words the entities of any cycle generated by $\pi(s) \circ \pi(s')$ can be reordered with respect to the pairing rules defined by structures s and s'. Such a rearranged set z will be called *orbit*.

Remark: We shortly want to present the idea of the rearrangement. Let i_1 be an unpaired position in structure s' and i_5 an unpaired position in s. Further on let $[i_1, i_2]$, $[i_3, i_4]$ form pairs in s and $[i_2, i_3]$, $[i_4, i_5]$ in s'. Then starting with position i_1 a chain can be built up

$$i_1 \overset{\pi(s)}{\longleftrightarrow} i_2 \overset{\pi(s')}{\longleftrightarrow} i_3 \overset{\pi(s)}{\longleftrightarrow} i_4 \overset{\pi(s')}{\longleftrightarrow} i_5 \; .$$

These five positions are included in a cycle ξ induced by $\pi(s) \circ \pi(s')$:

$$\xi = \langle i_1, i_3, i_5, i_4, i_2 \rangle$$

since $i_3 = \pi(s')(\pi(s)(i_1))$, $i_5 = \pi(s')(\pi(s)(i_3))$, $i_4 = \pi(s')(\pi(s)(i_5))$, etc. . The corresponding rearranged orbit is given by $z = (i_1, i_2, i_3, i_4, i_5)$ where by definition $[i_1, i_2] \in \Pi(s)$, $[i_2, i_3] \in \Pi(s')$, etc.

The orbit z is called *closed* if additionally $[i_1, i_k] \in \Pi[s] \cup \Pi[s']$ and *open* otherwise. Open orbits will be denoted by z^o whereas closed ones are denoted by z^c. For completeness we have to go over cycles of length one that are equivalent to fixed points of $\pi(s) \circ \pi(s')$. This is done in the following form: Let $\xi = \langle i \rangle$ and $\xi' = \langle i' \rangle$ be two cycles of length one then a orbit $z = (i, i')$ is defined if and only if $[i, i'] \in \Pi[s] \cap \Pi[s']$, i.e. $[i, i']$ form a common base pair of structures s and s', otherwise two orbits are derived with $z = (i)$ and $z' = (i')$.

The set of orbits corresponding to all cycles induced by $\pi(s) \circ \pi(s')$ defines the *orbit decomposition* Φ with respect to the pair of secondary structures s and s'. The following example shall demonstrate the procedure given above

Example: 2 *Let* $s =$ '`((.((....)).)).`' *and* $s' =$ '`.(((((...)).)))`'.

Then the corresponding permutations are given by

$$\pi(s) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 14 & 13 & 3 & 11 & 10 & 6 & 7 & 8 & 9 & 5 & 4 & 12 & 2 & 1 & 15 \end{pmatrix}$$

and

$$\pi(s') = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 1 & 15 & 14 & 13 & 11 & 10 & 7 & 8 & 9 & 6 & 5 & 12 & 4 & 3 & 2 \end{pmatrix} .$$

The product $\pi(s) \circ \pi(s')$ *is equal to*

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 3 & 4 & 14 & 5 & 6 & 10 & 7 & 8 & 9 & 11 & 13 & 12 & 15 & 1 & 2 \end{pmatrix}$$

which is equivalent to

$$\begin{pmatrix} 1 & 3 & 14 & 2 & 4 & 5 & 6 & 10 & 11 & 13 & 15 & 7 & 8 & 9 & 12 \\ 3 & 14 & 1 & 4 & 5 & 6 & 10 & 11 & 13 & 15 & 2 & 7 & 8 & 9 & 12 \end{pmatrix}$$

and this finally gives

$$\begin{array}{llcll} \xi_1 = & \langle 1, 3, 14 \rangle & \Rightarrow & z_1 = & (1, 14, 3) \\ \xi_2 = & \langle 2, 4, 5, 6, 10, 11, 13, 15 \rangle & \Rightarrow & z_2 = & (6, 10, 5, 11, 4, 13, 2, 15) \\ \xi_3 = & \langle 7 \rangle & \Rightarrow & z_3 = & (7) \\ \xi_4 = & \langle 8 \rangle & \Rightarrow & z_4 = & (8) \\ \xi_5 = & \langle 9 \rangle & \Rightarrow & z_5 = & (9) \\ \xi_6 = & \langle 12 \rangle & \Rightarrow & z_6 = & (12) \end{array}$$

In figure 6 we use the circle representation for RNA secondary structures [54] in order to visualize the construction of orbits corresponding to the pair of secondary structures (s, s').

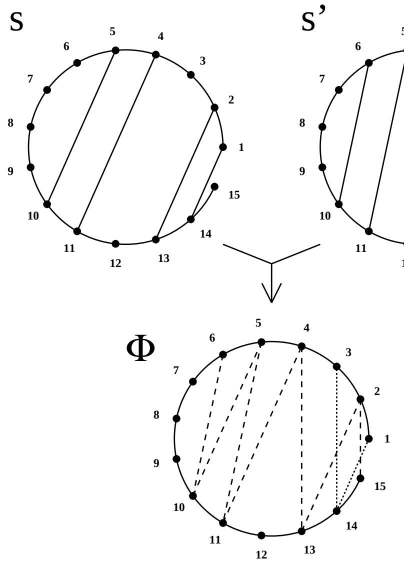


Figure 6: Circle representation for RNA secondary structures – a chord of the circle corresponds to a base pair. If both structures (s, s') are superimposed the chords induce paths in the circle which correspond to the orbits (dashed lines).

Now together with the orbit decomposition induced by s, s' we are able to give a rough approximation for the number of sequences being compatible with s and s'.

Corollary 3 *Let* $\mathcal{A}$ *be an alphabet of length* α *and* s *and* s' *be two secondary structures. Suppose that* $\mathcal{A}$ *admits at least one type of complementary base pair. Then* $|\mathbf{C}[s] \cap \mathbf{C}[s']| \geq \alpha^{|\Phi|}$ *where* Φ *is the set of orbits induced by* s *and* s'.

For any predefined pair of secondary structures (s, s') we now shall determine the size of the intersection in the case of the 'biophysical' (**AUGC**)–alphabet. For that reason we introduce alphabets of unpaired and paired bases $\mathcal{A}_1 \overset{\text{def}}{=\!=} (\mathbf{A},\mathbf{U},\mathbf{G},\mathbf{C})$ and $\mathcal{A}_2 \overset{\text{def}}{=\!=} (\mathbf{AU}, \mathbf{UA},\mathbf{UG},\mathbf{GU},\mathbf{GC},\mathbf{CG})$, respectively. Let $z^o = (i_1, \ldots, i_\nu)$ be an open orbit of length ν. Then we write N^o_ν for the number of possible chains $(\alpha_1, ..., \alpha_\nu)$, $\alpha_i \in \mathcal{A}_1$ such that $(\alpha_i \alpha_{i+1}) \in \mathcal{A}_2$ for $1 \leq i \leq \nu - 1$. Accordingly for a closed orbit $z^c = (i_1, ..., i_\nu)$ we set N^c_ν to be the number of chains $(\alpha_1, ..., \alpha_\nu)$, $\alpha_i \in \mathcal{A}_1$ such that $(\alpha_i \alpha_{i+1}) \in \mathcal{A}_2$ for $1 \leq i \leq \nu - 1$ and $(\alpha_1 \alpha_\nu) \in \mathcal{A}_2$. We immediately see that a sequence must cover all orbits induced by s, s' with sequences fulfilling the above given constraints in order to be compatible to a pair of RNA secondary structures s and s'.

Now the following lemma concerning N_ν^o and N_ν^c can be stated

Lemma 1 *Let z^o be an open orbit of length ν then*

$$N_\nu^o = \frac{2}{\sqrt{5}} \left[\left(\frac{2}{-1+\sqrt{5}} \right)^{\nu+2} - \left(\frac{2}{-1-\sqrt{5}} \right)^{\nu+2} \right] ; \quad \nu \geq 1 \quad .$$

And let z^c be a closed orbit of length $\nu = 2\mu$ then

$$N_{2\mu}^c = \frac{4}{\sqrt{5}} \left[\left(\frac{2}{-1+\sqrt{5}} \right)^{2\mu-1} - \left(\frac{2}{-1-\sqrt{5}} \right)^{2\mu-1} \right] ; \quad \mu \geq 1 \quad .$$

Proof: The proof is an application of the explicit formula for Fibonacci numbers $f_n = \frac{1}{\sqrt{5}} \left[\left(\frac{2}{-1+\sqrt{5}} \right)^{n} - \left(\frac{2}{-1-\sqrt{5}} \right)^{n} \right]$; for $n \geq 1$, $n \in \mathbb{N}$. ∎

Given a pair (s, s') of secondary structures and the corresponding orbit decomposition Φ the number of open orbits of length ν is denoted by n_ν^o and accordingly the number of closed orbits of length ν by n_ν^c. Now we are in the position to evaluate the exact cardinality of $\mathbf{I}[s, s']$ representing the number of sequences that are compatible with s and s' respectively.

Corollary 4 *Let $s, s' \in \mathcal{S}_n$ be two RNA secondary structures living on the 'biophysical' alphabet* **AUGC**. *Then the cardinality of their intersection $\mathbf{I}(s, s')$ is given by*

$$|\mathbf{I}(s, s')| = \prod_{\nu=1} (N_\nu^o)^{n_\nu^o} \cdot (N_{2\nu}^c)^{n_{2\nu}^c} = \prod_{\nu=1} 2^{n_{2\nu}^c} \, (N_\nu^o)^{n_\nu^o} \, (N_{2\nu-3}^o)^{n_{2\nu}^c}$$

where $\sum \nu \, (n_\nu^o + n_\nu^c) = n$ with $n_\nu^o = 0$ and $n_\nu^c = 0$ if no orbit of this kind is formed.

Example: 3 *For the pair of structures defined in the previous example we immediately derive a size of the intersection of* 281600 . *In this case the intersection covers more than* 1 *percent of* $\mathbf{C}[s]$ *and more than 3.5 percentage of* $\mathbf{C}[s']$.

The intersection corresponding to a pair of secondary structures in a natural way can be equipped with a canonical graph structure. For that reason we consider the induced subgraphs $\mathcal{C}[s][\mathbf{I}[s, s']]$ and $\mathcal{C}[s'][\mathbf{I}[s, s']]$ and define the intersection graph to be

$$\mathcal{I}[s,s'] \overset{\mathrm{def}}{=} (\mathbf{I}[s,s']\,,\, \mathrm{e}[\mathcal{C}[s][\mathbf{I}[s,s']]] \cup \mathrm{e}[\mathcal{C}[s'][\mathbf{I}[s,s']]]\,)\ .$$

Consequently two vertices $x, y \in \mathcal{I}[s,s']$ are adjacent if they are either adjacent in the subgraph $\mathcal{C}[s][\mathbf{I}[s,s']]$ or in $\mathcal{C}[s'][\mathbf{I}[s,s']]$. The orbit decomposition gives some more insight into the topology of $\mathcal{I}$. Adjacency of two sequences on the intersection graph is given if they either differ in a common unpaired position or in positions that form a common pair or in positions that are unpaired in one structure and paired in the other. One can verify that these positions are located in orbits of length one or two. Due to the number of orbits of length smaller or equal than two commensurate islands of the same size of intersection sequences are generated on $\mathcal{C}[s]$ and $\mathcal{C}[s']$ respectively. Orbits of length greater than two create paths in the sets of compatible sequences $\mathbf{C}[s]$ and $\mathbf{C}[s']$ that connect these islands. This may be described as follows: Let $z = (i_1, i_2, \ldots, i_\nu)$ be an orbit induced by the secondary structures s and s'. z is arranged in the following way $[i_k, i_{k+1}] \in \Pi[s] \cup \Pi[s']$ with $k = 1, \ldots, \nu - 1$. If we restrict ourselves to structure s then there exists a fixed $\ell \in \{0, 1\}$ such that $[i_{2k+\ell}, i_{2k+\ell+1}] \in \Pi[s]$ with $k = 0, 1, 2, \ldots$ all indices taken $(\mathrm{mod}\,\nu)$. Let $\alpha = (\alpha_1, \ldots, \alpha_\nu)$ and $\alpha' = (\alpha'_1, \ldots, \alpha'_\nu)$ be two different chains such that $(\alpha_j\alpha_{j+1}), (\alpha'_j\alpha'_{j+1}) \in \mathcal{A}_2$. Then we can transform α into α' by successive exchanges of pairs from $(\alpha_{2k+\ell}\alpha_{2k+\ell+1})$ to $(\alpha'_{2k+\ell}\alpha'_{2k+\ell+1})$ with $k = 0, 1, \ldots$ all indices taken $(\mathrm{mod}\,\nu)$. The situation is depicted in figure 7.

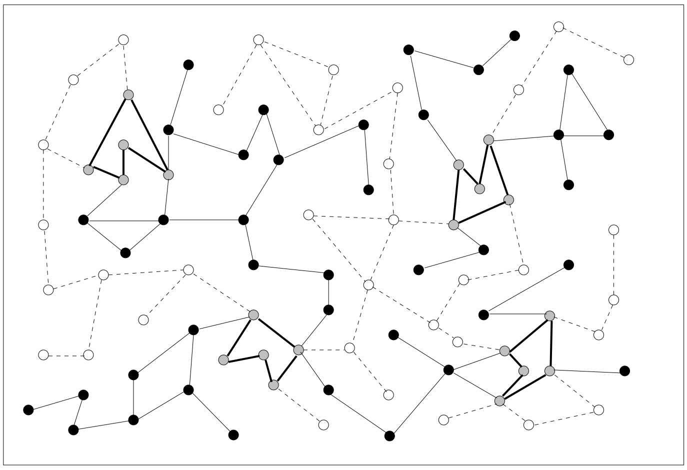


Figure 7: Embedding of the intersection (grey points) in $\mathcal{C}[s] \cup \mathcal{C}[s']$. The intersection graph decomposes into islands that are connected by paths in $\mathcal{C}[s]$ (white points) and $\mathcal{C}[s']$ (black points) respectively.

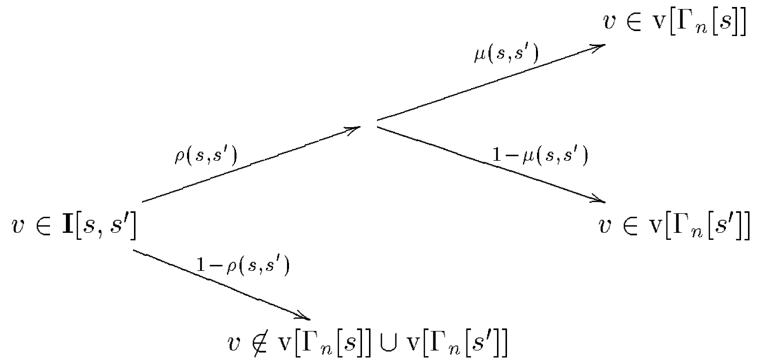
In section 3.2. two models for randomly constructing a single neutral network of an RNA secondary structure were introduced. These models need to be modified for two neutral networks corresponding to structures (s, s') thereby taking into account the existence of sequences being compatible with both structures. Therefore we propose to introduce a *joint neutrality parameter* $\rho(s, s')$ that gives the average probability of a sequence on the intersection to adopt one of the structures s or s', i.e.

$$\rho(s, s') \overset{\text{def}}{=} Prob\{v \in \Gamma_n[s] \cup \Gamma_n[s'] \mid v \in \mathbf{I}[s, s']\} \ .$$

The value of the conditional probability ρ is given by

$$\rho(s, s') = \frac{|\Gamma_n[s] \cap \mathbf{C}[s']|}{|\mathbf{I}[s, s']|} + \frac{|\Gamma_n[s'] \cap \mathbf{C}[s]|}{|\mathbf{I}[s, s']|} \ .$$

The sets $\Gamma_n[s] \cap \mathbf{C}[s']$ and $\Gamma_n[s'] \cap \mathbf{C}[s]$ cannot be described in terms of $\bar{\lambda}$ or $\bar{\lambda}'$ since more than one network covers the subset of sequences being compatible with structure but not adopting it. Therefore ρ needs to be considered as an independent and new parameter describing the interaction between two neutral networks. For any sequence $v \in \mathbf{I}[s, s']$ a random decision has to be realized according to the diagram:


$v \in \mathrm{v}[\Gamma_n[s]]$
$\mu(s,s')$
$\rho(s,s')$
$1-\mu(s,s')$
$v \in \mathbf{I}[s, s']$
$v \in \mathrm{v}[\Gamma_n[s']]$
$1-\rho(s,s')$
$v \notin \mathrm{v}[\Gamma_n[s]] \cup \mathrm{v}[\Gamma_n[s']]$

Furtheron the remaining sequences are handled according to the random model introduced in section 3.2. In the sequel we define $\mu(s, s') \overset{\text{def}}{=} \bar{\lambda}[s]/(\bar{\lambda}[s] + \bar{\lambda}[s'])$. This assumption leads to an extension in the case of two neutral networks which only requires one additional degree of freedom.

4. Dynamics on Neutral Networks

In molecular evolution the source of variation is the limited accuracy of replication. Copying errors, mutations, produce RNA sequences which differ from the parental template sequence. A change in genotype can result in a change in phenotype – the RNA structure. Mutations thus act on the nucleotides and provide the genetic reservoir from which better adapted variants are chosen. They fall into three classes

ACGAUGGGUUACC|G|AGGCAAGUCGUAG
Point mutation $\Downarrow$
ACGAUGGGUUACC|A|AGGCAAGUCGUAG

ACGAUG|GGUUACCG|AGGCAAGUCGUAG
Insertion $\Downarrow$
ACGAUG|GGUUACCG|GGUUACCG|AGGCAAGUCGUAG

ACGAUGGG|UUACCGAGGC|AAGUCGUAG
Deletion $\Downarrow$
ACGAUGGG|AAGUCGUAG

Figure 8: 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 lacking part of the template's sequence.

While insertions and deletions alter the size of the genome, the chain length is kept constant under point mutations. Mutational frequencies of all classes are phenotype dependent. Position dependence of point mutations are much weaker than that of insertions and deletions. Therefore a uniform error model of point mutations was conceived [12, 13, 14, 15, 16, 61, 64] and successfully applied to replication and mutation of RNA molecules *in vitro*, and of viroids and RNA viruses *in vivo*.

Mutations can be viewed as 'moves' in an abstract space of *configurations*. This provides a natural arrangement of the configurations in a geometrical context. Configurations that can be interconverted by a single move may be viewed as neighbors. Consequently the smallest number of moves necessary to convert two configurations into each other can be interpreted as their distance.

In this contribution we consider the most simple example of Darwinian evolution. A *population* $\mathbf{V}$ of replicating RNA sequences are competing for a common resource. We restrict ourselves to point mutations. This ensures that for a fixed chain length n the configuration space can be described as a generalized hypercube $\mathcal{Q}^n_\alpha$.

Definition 1 (population) *Let* $\mathcal{Q}^n_\alpha$ *be the generalized hypercube of all sequences of length* n *over the alphabet* α*. A population* $\mathbf{V}$ *is a (finite) family of vertices* $(v_i \in \mathcal{Q}^n_\alpha \mid i \in \{1,\ldots,N\})$.

Following the work of Eigen *et al.* [15] we consider a population of asexual replication strings (RNA sequences) of fixed length n evolving in a stirred flow reactor whose total RNA population fluctuates around a constant capacity N. The definition of the overall replication rate of a sequence together with the constrained population size specifies our selection criterion.

The minimal representation of the corresponding *reaction network* of polynucleotides I_k $(k = 1,\ldots,\nu)$ is given as follows:

$$\begin{array}{rcll} (A) + I_k & \xrightarrow{A_k Q_{ik}} & I_i + I_k & \qquad i,k = 1,\ldots\nu = \alpha^n \\ I_k & \xrightarrow{\Phi(t)} & * & \qquad k = 1,\ldots,\nu = \alpha^n \end{array}$$

where A_k is the replication rate of genotype I_k and Q is the matrix of mutation probabilities, Q_{ik} being the probability for a parent I_i to have an offspring of genotype I_k, and ν is the number of possible genotypes (2^n or 4^n). The total population size is kept constant over time by a flux $\Phi(t)$ compensating the production of new offsprings. The model mimics the asynchronous serial transfer. In the limit of infinite population size the time evolution in the flow reactor is described by the *quasi-species equation*

$$\dot{x}_i(t) = \sum_k Q_{ki} A_k(t) x_k(t) - x_i(t)\Phi(t), \qquad i = 1,\ldots,\nu = \alpha^n \tag{1}$$

where $x_i(t)$ denotes the concentration of genotype I_i at time t, and $\Phi(t) = \sum_k A_k(t)x_k(t)$ if $\sum_i x_i(t)$ is scaled to one.

Similar to the experiments in the laboratory the RNA populations we deal with should be small compared to the size of the sequence space. This forces a description in terms of stochastic chemical reaction kinetics. Two methods are appropriate to model the stochastic process taking place in a stirred flow reactor.

(1) Gillespie [28, 29] has described an algorithm for numerically simulating the time evolution of any spatially homogeneous mixture of chemically reacting molecules. A brief description of this method can be found in appendix B. The implementation for the flow reactor dynamics using this algorithm was realized by Fontana and Schuster [24]. For a population of RNA molecules evolving in a flow reactor the underlying stochastic process can be seen as a birth and death process. Birth events are realized by replication whereas death is realized by the unspecific dilution flux.

(2) While giving a complete description of the flow reactor dynamics the firstly mentioned stochastic formulation sometimes is suitable for a detailed mathematical analysis because of the fluctuating population size. Avoiding this the quasi-species model can be approximated by a generalized birth death process, following the lines of Nowak and Schuster [53] (for a detailed description see appendix B).

In order to make the two methods listed above distinguishable we shall call method (1) *flow reactor dynamics* and (2) the *replication deletion process*.

The set of replication rates $\{A_k\}_{k=1}^{\nu}$ considered as a function of the genotypes $\{I_k\}_{k=1}^{\nu}$ forms a fitness landscape $\tilde{f}$ over the sequence space [79]

$$\tilde{f} : \mathcal{Q}_\alpha^n \to \mathbb{R}^+; \qquad I_k \mapsto A_k \quad k = 1, \ldots, \nu = \alpha^n.$$

However the choice of A_k is somewhat arbitrary. Now the existence of neutral networks, i.e., of sequences that share the same phenotype, motivates to define the replication rates in the context of phenotypes such that

$$\phi : \mathcal{Q}_\alpha^n \to \mathcal{S}_n; \qquad f : \mathcal{S}_n \to \mathbb{R}^+$$

where the fitness of a genotype I_k is given by $f(\phi(I_k))$. Landscapes formed over extended neutral networks show a high degree of neutrality, i.e., a large number of sequences that have the same fitness value. They are called *smooth* and evolution taking place on them is called *neutral evolution*.

If the phenotype of an RNA sequence is supposed to be its secondary structure in a natural way a landscape is formed possessing a high degree of neutrality.
The next three sections we devote the study of neutral evolution. We will ask for the time evolution not for single genotypes but for the number of sequences sharing the same phenotype.

4.1. The Basic Model

Suppose there are only two phenotypes that also have the same fitness value, i.e., the landscape is completely flat and the sequences space decomposes into two disjoint classes of sequences. These classes will be denoted by G_1 and G_2. Let w_{ik} be the average probability to get a sequence on G_k by a replication reaction if its template was a sequence on G_i, $i,k = 1,2$. For the moment we assume $w_{22} = w_{11} = w$ and hence $w_{21} = w_{12} = 1-w$.

4.1.1. Constant Population Size

Now we study the evolutionary behavior of a replication deletion process acting on the bipartition of the sequence space. Let X_t be the random variable counting the number of sequences in a population $\mathbf{V}_t$ that belong to G_1. Because of symmetry we are allowed to restrict ourselves to G_1. We shall approximate the corresponding replication-deletion process by a birth-death process in continuous time [42, 53]. The infinitesimal transition probabilities are given by

$$P_{\ell,k}(h) := Pr\{X_{t+h} = k \mid X_t = \ell\} = \begin{cases} \lambda_\ell \cdot h + o(h) & k = \ell + 1 \\ \mu_\ell \cdot h + o(h) & k = \ell - 1 \\ 1 - (\lambda_\ell + \mu_\ell) \cdot h + o(h) & k = \ell \\ o(h) & k \neq \ell, \ell \pm 1 \end{cases}$$

with

$$\begin{array}{rcll} \lambda_\ell & = & \frac{N-\ell}{N(N-1)}(\ell\, w + (N-1-\ell)(1-w)) & \\ \mu_\ell & = & \frac{\ell}{N(N-1)}((\ell-1)(1-w) + (N-\ell)w) & \quad 0 \le \ell \le N \end{array}$$

where $P_{N,N+1}(h) = P_{0,-1}(h) = 0$ ensures reflecting boundary conditions. Sometimes it can be appropriate to consider a time scale depending on the mean fitness of the population [29]. This ansatz takes into account that a population with higher mean fitness is expected to replicate faster in time than a population with lower fitness value. For the example presented here the mean fitness is always equal to one and does not vary. Hence no transformation is necessary.

First we will state an ergodic theorem [20] that implies the existence of a stationary distribution for our birth death process.

Theorem 4. *Let X_t be a homogeneous Markov process with finitely many states $0, \ldots, N$. If there exists t^* with $0 < t^* < \infty$ such that $P_{i,k}(t^*) > 0$ for $0 \le i, k \le N$ then there exists the limit*

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

Using the Markovian property of our process we can verify by induction on $|i-k|$:

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

Whence the theorem can be applied. We derive the stationary distribution by use of the master equation[2]

$$p_k = \frac{\pi_k}{\sum_{i=0}^N \pi_i}, \quad \text{with} \quad \pi_0 = 1 \quad \text{and} \quad \pi_k = \frac{\mathcal{B}(a,N)}{\mathcal{B}(a,k)\cdot\mathcal{B}(a,N-k)} \tag{2}$$

where $a = (N-1)(1-w)/(-1+2w)$. It can be shown that $\pi_k = \pi_{N-k}$ for all $0 \le k \le N/2$. Furthermore the following lemma can be stated.

---

[2] $\mathcal{B}(x,\ell) \stackrel{\text{def}}{=} (\ell-1)!/[x(x+1)\ldots(x+\ell-1)], \quad x \in \mathbb{R},\ \ell \in \mathbb{N}$

Lemma 2 *There exists a critical value $w^* = N/(N+1)$ where the modality of the probability distribution changes.*

Proof: Using the symmetry property $\pi_k = \pi_{N-k}$ for all $0 \le k \le N/2$ we consider $\pi_k - \pi_{k+1}$ for any $k < (N-1)/2$

$$\pi_k - \pi_{k+1} = \pi_k \left(\frac{-(N-1-2k)(N-w\,(N+1))}{(k+1)(k(1-w)+(N-k-1)w)} \right).$$

Thus we find

$$\begin{array}{lll} \pi_k < \pi_{k+1} & \text{for} & w < N/(N+1) \\ \pi_k = \pi_{k+1} & \text{for} & w = N/(N+1) \\ \pi_k > \pi_{k+1} & \text{for} & w > N/(N+1) \qquad \forall\, k \le (N-1)/2 \end{array}$$

∎

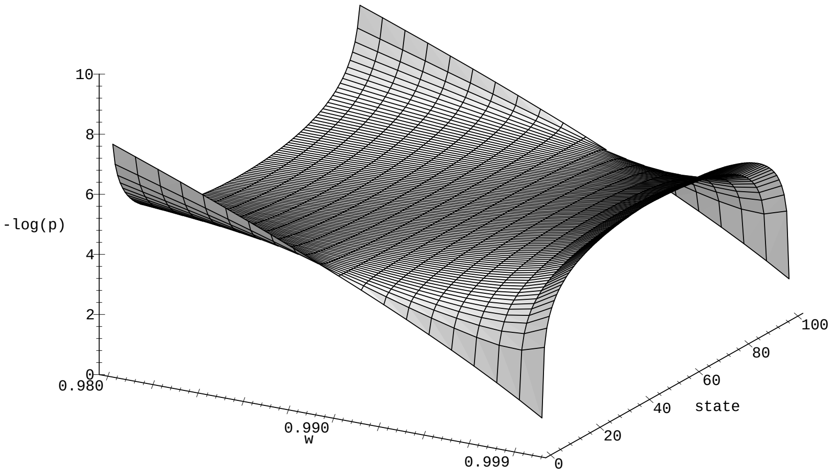


Figure 9: For increasing w we plot the negative logarithm of the probability distribution for the states 0,. . .,100.

For increasing w the stationary probability distribution goes from an unimodal distribution over an uniform in the critical value w^* to a bimodal distribution with peaks in states 0 and N. For population size $N = 100$ this is illustrated in figure 9.

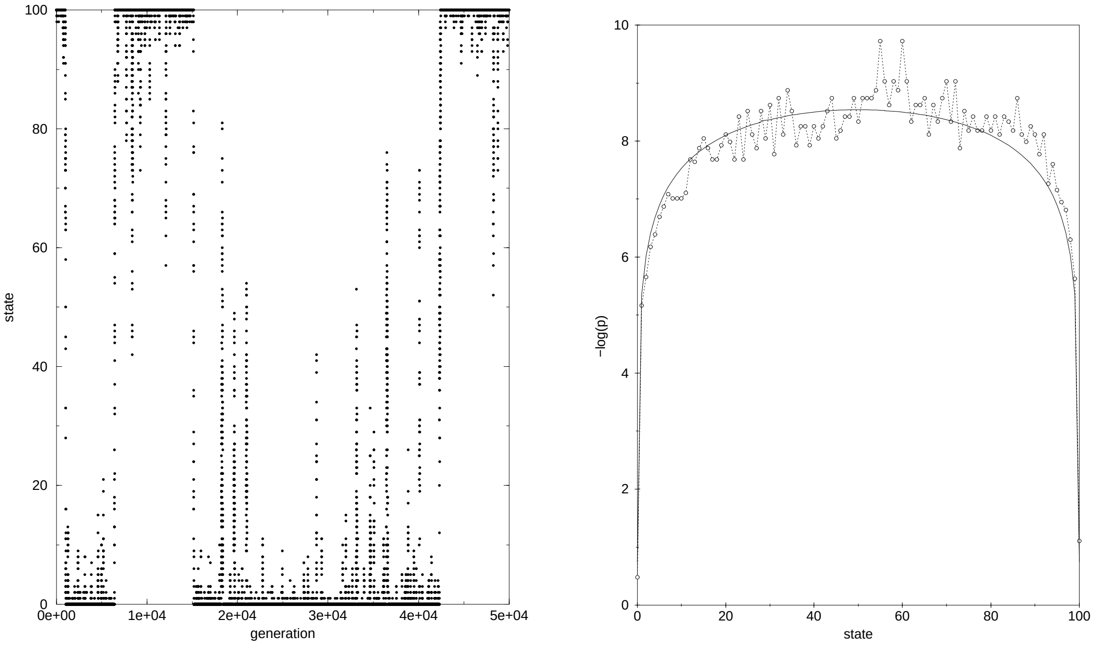


Figure 10: One experiment. For w=0.9999 the results corresponding to one realization of the stochastic process are presented. Starting with 100 elements on G_1 the experiment is executed $5 \cdot 10^6$ times. Assuming that 100 trials correspond to one generation on the left we plot the evolution of elements on G_1. On the right applying the ergodic property of the process the results are used to estimate a stationary distribution (circles) and this is compared with the analytical curve.

In figure 10 for $w = 0.9999$ we pursue one realization of the stochastic process. Starting with 100 elements on G_1 and executing the experiment $5 \cdot 10^6$ times after 100 trials the number of elements on G_1 is reported. Then the results of the experiment are used to estimate the stationary distribution which fits the analytical curve very well.

The parameter w is seen to act as a coupling constant between G_1 and G_2. If w is chosen above the critical value a strong decoupling of both sets is observed which is equivalent to an increasing fixation on each of them. This corresponds to a new emerging attribute of the evolutionary process that is characterized by sharp and fast transitions between the states zero and N (see figure 10). A tool properly provided by the theory of birth and death processes allows to describe this property. If state zero is considered to be absorbing, i.e., $\lambda_0 = 0$, a quantity ω_m giving the mean time when starting from state m of being absorbed in state 0 can be declared. Following [42] we derive an expression for

ω_m:

$$\omega_0 = 0; \quad \omega_m = N \cdot (N-1) \cdot \sum_{i=1}^{m} \frac{\sum_{k=i}^{N} \pi_k}{\pi_i \, i \left((1-2w)\, i + w\,(N+1) - 1 \right)}, \quad m = 1, \ldots, N$$

where all π_k's, $k = 1, \ldots, N$ are prescribed by equation (2).

Given a population that is completely located on G_1 one might ask for the mean time to lose all information G_1. The term ω_N, in a direct way, gives an expected value about the length of the corresponding time interval. For the example presented in figure 10 the length of this time interval is approximately 10197 generations. Figure 11 shows the dependence of ω_N on w. It can be seen that ω_N has a minimum w_{min} close to w^*. It is exponentially decreasing for $w < w_{min}$ and exponentially increasing for $w > w_{min}$.

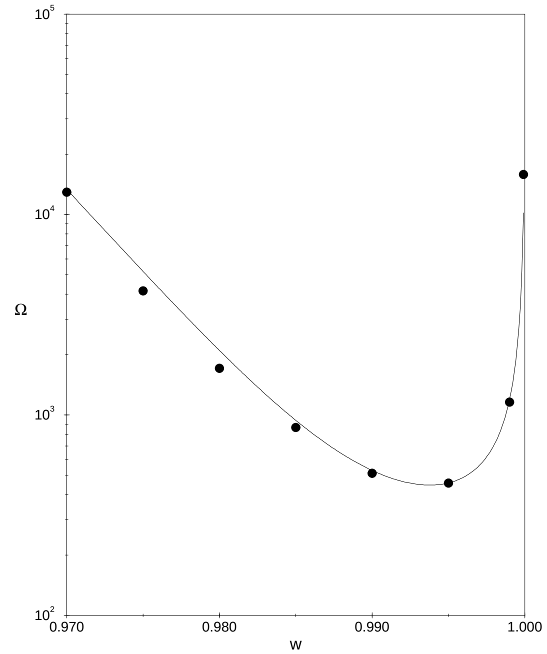


Figure 11: Mean time to extinction. For population size N=100 and increasing w the average number of generations to enter state 0 for the first time when starting with N elements $\Omega \overset{\text{def}}{=} \omega_N / N$ is plotted. The solid line corresponds to the analytical curve whereas the circles indicate experimentally determined values.

For completeness the case of unequal w_{11} and w_{22} needs to be considered. The results are displayed in figure 12. For reasons of symmetry we are allowed to restrict ourselves to $w_{11} < w_{22}$ and ask for the stationary distribution of G_1. For $w_{22} < w^*$ the unimodal distribution exhibits a local maximum greater than zero but smaller than $N/2$. Setting $w_{11} < w^* < w_{22}$ the state zero, i.e., extinction of G_1 becomes most probably. Again for

$w^* < w_{11} < w_{22}$, i.e., above the critical value, the stationary distribution has two local maxima in states zero and N but always we find p_0 to be greater than p_N.

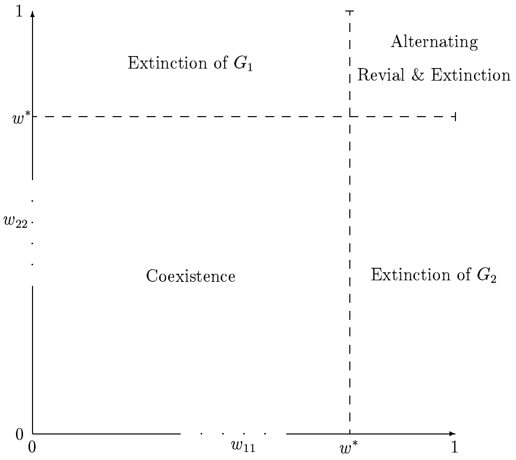


Figure 12: Different evolutionary scenarios depending on the tuple (w_{11}, w_{22})

4.1.2. Fluctuating Population Size

Under the same initial conditions we now study the stochastic process corresponding to the flow reactor dynamics. Taking into account fluctuating population size we have to consider the two-dimensional random variable (X_t, Y_t) where X_t is counting the number of sequences in a population belonging to G_1 and Y_t those belonging to G_2. Then the infinitesimal transition probabilities are defined by

$$P_{(x,y),(x',y')}(h) = \begin{cases} \lambda_x h + o(h) & x' = x+1, y' = y \\ \lambda_y h + o(h) & x' = x, y' = y+1 \\ \mu_x h + o(h) & x' = x-1, y' = y \\ \mu_y h + o(h) & x' = x, y' = y-1 \\ 1 - (\lambda_x + \lambda_y + \mu_x + \mu_y)h + o(h) & x = x', y = y' \\ 0 & x + y \le 0; x < 0; y < 0 \\ o(h) & \text{otherwise} \end{cases} \qquad x, y = 0, \ldots, \infty$$

with

$$\lambda_x(x,y) = \frac{xw + y(1-w)}{N(x+y)}, \; \lambda_y(x,y) = \frac{x(1-w) + yw}{N(x+y)}, \; \mu_x(x,y) = \frac{x}{N^2}, \; \mu_y(x,y) = \frac{y}{N^2}.$$

and $P_{(x,y),(x',y')}(h) \overset{\text{def}}{=} Prob\{(X_{t+h}, Y_{t+h}) = (x', y') | (X_t, Y_t) = (x, y)\}$.

The stationary distribution of the random variable (X_t, Y_t) is fully determined by the relation

$$P\{(X,Y) = (x,y)\} = P(X+Y = x+y) \cdot P(X = x | X+Y = x+y) \tag{3}$$

We immediately see that the random variable $Z_t \overset{\text{def}}{=} X_t + Y_t$ follows a birth death process because X_t and Y_t are independent random variables at time t and therefore the infinitesimal rates are additive. Thus we find for $z = 0, 1, \ldots$:

$$P(Z_{t+h} = z' \,|\, Z_t = z) = \begin{cases} N^{-1}h + o(h) & z' = z+1 \\ zN^{-2}h + o(h) & z' = z-1 \\ 1 - (N+z)N^{-2}h + o(h) & z = z' \\ o(h) & \text{otherwise} \end{cases}$$

Consequently the stationary probability distribution for the random variable Z_t can be determined [42] and is seen to follow a Poisson distribution

$$P(Z = z) = \frac{N^z}{z!}\,\mathrm{e}^{-N}; \quad z = 0, 1, \ldots \quad . \tag{4}$$

Now it remains to show that $\lim\limits_{t\to\infty} P(X_t = x | Z_t = z) = P(X = x | Z = z)$ exists. Let $h' = \sqrt{h}$ be a small time increment. Then the probability that there is a birth event for random variable X_t under the condition that Z_t remains constant is

$$\begin{array}{rl} & (\lambda_x(x, z-x)h' + o(h')) \cdot (\mu_y(x+1, z-x-1)h' + o(h')) \\ + & (\mu_y(x, z-x)h' + o(h')) \cdot (\lambda_x(x, z-x-1)h' + o(h')) \\ = & (\lambda_x(x, z-x)\mu_y(x+1, z-x-1) + \lambda_x(x, z-x-1)\mu_y(x, z-x))h + o(h) \end{array} \tag{5}$$

The same argument runs for a conditional death event of X_t. Thus for $0 \le x \le z$ the infinitesimal birth and death rates are given by

$$\begin{aligned} \lambda(x|z) &= \tfrac{z-x}{N^3 z}(x(-1+2w) + z(1-w)) + \tfrac{z-x}{N^3(z-1)}(x(-1+2w) + (z-1)(1-w)); \\ \mu(x|z) &= \tfrac{x}{N^3 z}(-x(-1+2w) + zw) + \tfrac{x}{N^3(z-1)}(-(x-1)(-1+2w) + (z-1)w); \end{aligned} \tag{6}$$

Consequently by solving the corresponding master equation [42] we get the conditional stationary distribution

$$p(x|z) = \frac{\pi(x|z)}{\sum\limits_{x=0}^{z} \pi(x|z)} \quad \text{with} \quad \pi(0|z) = 1; \qquad \pi(x|z) = \frac{\mathcal{B}(a', z)}{\mathcal{B}(a', x) \cdot \mathcal{B}(a', z-x)} \tag{7}$$

where $a' = 2(1-w)z(z-1)/((-1+2w)(2z-1))$.

Hence the left side of equation (4) is fully determined. And finally the stationary probability for the random variable X_t counting the number of elements belonging to G_1 becomes

$$P(X=x) = \sum_{z=x}^{\infty} \frac{N^z}{z!} \mathrm{e}^{-N} p(x|z) \tag{8}$$

For increasing w and an average population size of 100 we present in figure 13 the stationary distribution for random variable X_t.

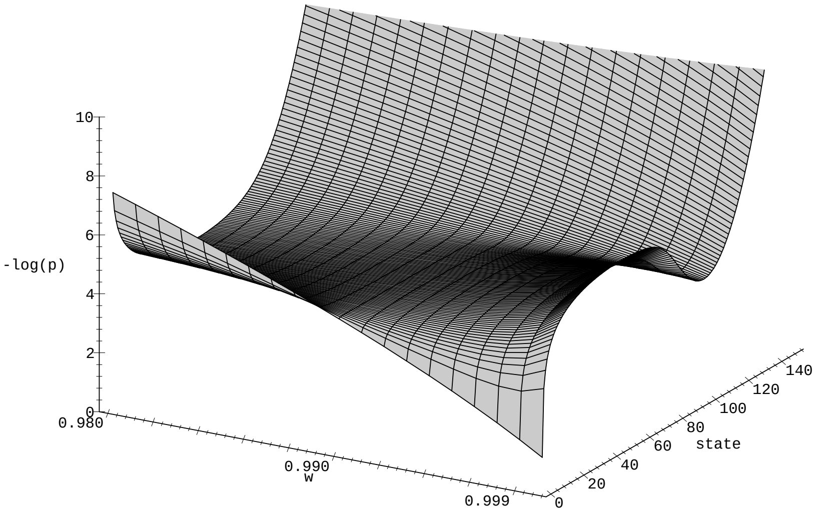


Figure 13: For increasing w and $N = 100$ we plot the negative logarithm of the stationary probability distribution for random variable X_t.

It is much harder to give an analytical value for the critical rate w where the modality of the probability distribution changes. The conditional probability distribution $p(x|z)$ shows the same behavior and its modality changes for $w = (2z^2-1)/(2(z^2+z-1))$. The population is biased in the value N and therefore it is appropriate to assume that the critical value is located at $w^* = (2N^2-1)/(2(N^2+N-1)) \approx N/(N+1)$.

Remark: Let $\{G_i\}_{i=1}^m$ be a disjoint partition of the hypercube with respect to m different phenotypes. Let further w_{ik} with $i, k = 1, \ldots, m$ be the average probability by a replication reaction to get an element on G_k if its template was chosen from G_i. Then w_{ii} is called *phenotypic fixation probability* and w_{ik} with $i \neq k$ is called *phenotypic transition probability.*

4.2. One Neutral Network – Single Shape Landscape

The counterpart of the single peaked landscape [16] is the so called *single shape landscape* on the level of RNA secondary structures. A neutral network $\Gamma_n[s]$ defined by the *shape* s induces a landscape on the complete sequence space $\mathcal{Q}_\alpha^n$:

$$f_s(x) := \begin{cases} 1 & \text{iff} \quad x \notin \Gamma_n[s] \\ \sigma > 1 & \text{iff} \quad x \in \Gamma_n[s] \end{cases}$$

A single RNA secondary structure is fixed and all sequences folding into this structure have a superior fitness σ compared to all other sequences.

A detailed analytical description of the evolution of a finite population on a single shape landscape was already presented by Reidys *et al.* [56, 57]. f_s always induces a bipartition of a population $\mathbf{V}$ in $\mathcal{Q}_\alpha^n$:

$$\mathbf{V}_\mu \;:=\; \{v \in \mathbf{V} \mid \; v \in \mathrm{v}[\Gamma_n[s]]\,\}; \qquad \mathbf{V}_\nu \;:=\; \{v \in \mathbf{V} \mid \; v \notin \mathrm{v}[\Gamma_n[s]]\}$$

The elements of $\mathbf{V}_\mu$ are called *masters* because of their superior fitness and those of $\mathbf{V}_\nu$ are called *non-masters.*

A crucial point that was already stressed by Nowak and Schuster [53] is seen to be the formulation of the phenotypic fixation probabilities $W_{\nu\nu}$ and $W_{\mu\mu}$.

4.2.1. Phenotypic Fixation Probabilities

We first introduce some terminology

- An alphabet $\mathcal{A}$ is called a $\star$–alphabet iff

  (i) $\mathcal{A}$ consists of complementary bases i.e., $\mathcal{A} = \{A_1, A_1^c, \dots, A_m, A_m^c\}$ where $|\mathcal{A}| = \alpha = 2m$, and

  (ii) the induced pair alphabet $\mathcal{B}$ (length β) has the form $\mathcal{B} = \{(A_1, A_1^c), \dots, (A_m, A_m^c)\}$, whence $\beta = \alpha$.

  Examples for $\star$–alphabets are (**G,C**) and (**G,C,X,K**).
- Let $\Gamma_n[s]$ be a neutral network with respect to the secondary structure s and let $v = (v_1, \dots, v_n)$ be a sequence. Then we define the *incompatibility distance* $d(\Gamma_n[s], v)$ by

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

  where Π is the pairing rule of the underlying alphabet and $\Pi(s)$ the set of contacts of s [57, 56].
- By setting $\mathcal{E}_k := \{v \mid d(\Gamma_n[s], v) = k\}$ a natural decomposition of the hypercube with respect to a secondary structure s is given that in a formal way corresponds to the different Hamming classes studied in the case of the single peaked landscape.
- By

  $$C_i[s] := \{\, v \mid v \notin \mathrm{v}[\Gamma_n[s]], \quad d(\Gamma_n[s], v) = i \,\}, \quad i = 1, \dots, n_p$$

  we define the i-th *incompatible class* with respect to the secondary structure s. And consequently the density of the i-th incompatible class is given by

  $$\Delta_i := |C_i[s]|/(|\mathcal{Q}_\alpha^n| - |\Gamma_n[s]| = \begin{cases} \binom{n_p}{i}\, \alpha^{n_u + n_p} (\alpha - 1)^i / (\alpha^n - |\Gamma_n[s]|); & 1 \le i \le n_p \\ (\alpha^{n_u + n_p} - |\Gamma_n[s]|)/(\alpha^n - |\Gamma_n[s]|); & i = 0 \end{cases}$$

Lemma 3 *Suppose* $\Gamma_n[s] < \mathcal{Q}_\alpha^n$ *is a fixed neutral network and* $\mathcal{A}$ *is a* $\star$*-alphabet. Suppose that we have a random mapping* $v = (x_1, \ldots, x_n) \mapsto v' = (x_1', \ldots, x_n')$, $v, v' \in v[\mathcal{Q}_\alpha^n]$ *that is defined as follows:* $x_i = x_i'$ *with probability* $1-p$ *and* $x_i \neq x_i'$ *with uniform probability* p*. Then*

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

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

A proof of the above lemma was given by Reidys *et al.* [56]. Note that $W_{\nu\nu} = 1 - W_{\nu\mu}$. In contrast to [56] we found for $W_{\nu\mu}$ the following formula

$$\begin{aligned} W_{\nu\mu}(p) = \ & \bar{\lambda}\Big\{\Delta_0\Big[-(1-p)^n + \Big(\tfrac{p^2}{\alpha-1} + (1-p)^2\Big)^{n_p}\Big] + \\ & \sum_{i=1}^{n_p} \Delta_i \Big(\tfrac{p\,(1-p)}{\alpha-1} + \tfrac{(\alpha-2)\,p^2}{(\alpha-1)^2}\Big)^i \Big(\tfrac{p^2}{\alpha-1} + (1-p)^2\Big)^{n_p-i}\Big\} \end{aligned}$$

Proof: An arbitrary sequence in incompatible class $C_i[s]$ can be arranged as follows

$$\underbrace{x_1, x_2, \ldots, x_{n_u}}_{n_u}\ \underbrace{(y_1, y_1'), \ldots, (y_i, y_i')}_{i}\ \underbrace{(z_1, z_1'), \ldots, (z_{n_p-i}, z_{n_p-i}')}_{n_p-i}$$

An incompatible base pair has to undergo mutations to become compatible. This can be realized in two different ways. The first is to mutate one position and to leave the others unchanged. The probability for this to happen and to obtain a compatible base pair is $p(1-p)/(\alpha-1)$. The second way is to mutate in both positions. The probability to occur and to obtain a compatible base pair is $p^2(\alpha-2)/(\alpha-1)^2$. Therefore i incompatible base pairs have the probability

$$\sum_{i_o=0}^{i} \binom{i}{i_o} \left(\frac{p\,(1-p)}{\alpha-1}\right)^{i_o} \left(\frac{p^2(\alpha-2)}{(\alpha-1)^2}\right)^{i-i_o} = \left(\frac{p\,(1-p)}{\alpha-1} + \frac{p^2(\alpha-2)}{(\alpha-1)^2}\right)^i$$

to become compatible with the target structure. For a compatible base pair that undergoes mutation the probability to become compatible again is $p^2/(\alpha-1) + (1-p)^2$.

Suppose there is no incompatible base pair. Then at least one mutation is required. For this we obtain the probability

$$\begin{aligned}
&[1-(1-p)^{n_u}]\,(1-p)^{2n_p} + [1-(1-p)^{n_u}]\left\{\left[\frac{p^2}{\alpha-1}+(1-p)^2\right]^{n_p} - (1-p)^{2n_p}\right\}\\
&+(1-p)^{n_u}\left\{\left[\frac{p^2}{\alpha-1}+(1-p)^2\right]^{n_p} - (1-p)^{2n_p}\right\}\\
&= -(1-p)^{n} + \left[\frac{p^2}{\alpha-1}+(1-p)^2\right]^{n_p}
\end{aligned}$$

This completes the proof of the lemma. ■

In the sequel we write for short w_{ik} for the $W_{\cdot,\cdot}$ with $i,k=0,1$, where 0 corresponds to a non master, and 1 to a master on $\Gamma_n[s]$. Obviously for $i=0,1$ holds $w_{i0}+w_{i1}=1$.

4.2.2. Stochastic Approach

Following Reidys *et al.* [57] we study the random random variable X_t that counts the number of master sequences in a population $\mathbf{V}_t$ evolving in a single shape landscape. The replication–deletion process is approximated by a birth death process. The ansatz of the birth and death rates is completely analogous to that of Nowak and Schuster [53].

$$P_{\ell,k} := Pr\{X_{t+h}=k \,|\, X_t=\ell\} = \begin{cases} \lambda_\ell\cdot h + o(h) & k=\ell+1\\ \mu_\ell\cdot h+o(h) & k=\ell-1\\ 1-(\lambda_\ell+\mu_\ell)\cdot h+o(h) & k=\ell\\ o(h) & k\neq \ell,\ell\pm 1\end{cases}$$

with

$$\begin{aligned}
\lambda_\ell &= \frac{N-\ell}{(N-1)(N+(\sigma-1)\ell)}(\sigma\, w_{11}\,\ell + w_{01}(N-\ell-1))\\
\mu_\ell &= \frac{\ell}{(N-1)(N+(\sigma-1)\ell)}(\sigma\, w_{10}\,(\ell-1)+w_{00}\,(N-\ell)) \qquad 0\le\ell\le N
\end{aligned}$$

Note that again 0 and N are reflecting boundaries. With the same arguments as in section 4.1. the existence of the stationary probability distribution can be shown. Hence by making use of the master equation one can determine $p_k=\pi_k/(\sum_{i=0}^N \pi_i)$ with

$$\pi_0 = 1 \quad\text{and}\quad \pi_k = \left(1+(\sigma-1)\frac{k}{N}\right)\left(\frac{\Lambda_{11}}{\sigma-1-\Lambda_{11}}\right)^k \frac{\mathcal{B}(a_1,N)}{\mathcal{B}(a_2,k)\cdot\mathcal{B}(a_1,N-k)} \tag{9}$$

where $a_1 = -\sigma w_{10}(N-1)/\lambda_{10}$ and $a_2 = w_{01}(N-1)/\Lambda_{11}$ and $\Lambda_{ik} := \sigma\, w_{ik} - w_{0k},\ i,k = 0,1$ and $\Lambda_{11}, \Lambda_{10} \neq 0$.

For certain ranges of the parameter the distribution is unimodal. But there are other shapes, and we can detect two critical values altering the modality of the distribution. The first one is $w^*_{00} = \frac{N-1+\sigma}{N+\sigma}$ and the second $w^*_{11} = \frac{(N-1)\sigma+1}{N\sigma+1}$.

Proof: We consider the difference $\pi_{k+1} - \pi_k = \pi_k \zeta(k)$ with $\zeta(k) := \frac{\lambda_k}{\mu_{k+1}} - 1$. We find $\zeta(k) \geq 0$ for $k \geq 0$ if and only if

$$k\,(\sigma w_{11} - (1 - w_{00})) + (1 - w_{00})(N-1) \geq \frac{(k+1)(N + (\sigma-1)k)(N - 1 + (\sigma-1)k)}{N\,(N + (\sigma-1)k + \sigma)}$$

For fixed w_{00} and w_{11} the linear function on the left hand side and the monotonously increasing parabolic function can have at most 2 points of intersection. We consider the inequality at the points $k = 0$ and $k = N-1$ and find the following scenarios

1. $\zeta(0) > 0 \quad \Leftrightarrow \quad w_{00} < w^*_{00} = \frac{N+\sigma-1}{N+\sigma}$ and $\zeta(N-1) > 0 \quad \Leftrightarrow \quad w_{11} > w^*_{11} = \frac{(N-1)\sigma+1)}{N\sigma+1}$

   No intersection point exists and state N is the global maximum of the stationary probability distribution.

2. $\zeta(0) > 0 \quad \Leftrightarrow \quad w_{00} < w^*_{00}$ and $\zeta(N-1) < 0 \quad \Leftrightarrow \quad w_{11} < w^*_{11}$

   There exists one intersection point. We get one global maximum of the stationary probability distribution that is located between state 0 and N.

3. $\zeta(0) < 0 \quad \Leftrightarrow \quad w_{00} > w^*_{00}$ and $\zeta(N-1) > 0 \quad \Leftrightarrow \quad w_{11} > w^*_{11}$

   State 0 and N become local maxima of the distribution. State N becomes the global maximum if the following inequality holds

   $$w_{11} > \max\left\{ \frac{((N-1)(\sigma-1)+N)(N+\sigma-1)(1-w_{00})}{\sigma w_{00} + ((N-1)(\sigma-1)+N)(N+\sigma-1)(1-w_{00})}, w^*_{11} \right\}.$$

4. $\zeta(0) < 0 \quad \Leftrightarrow \quad w_{00} > w^*_{00}$ and $\zeta(N-1) < 0 \quad \Leftrightarrow \quad w_{11} < w^*_{11}$

   There exists a bifurcation value $\bar{w}_{11}(w_{00})$ where the distribution switches from two maxima in 0 and a value between 1 and N to a single maximum in state zero.

$$\bar{w}_{11}(w_{00}) = \min_{0<k<N-1} \frac{(k+1)(N+(\sigma-1)k)(N-1+(\sigma-1)k)}{\sigma N(N+(\sigma-1)k+\sigma)\,k} - \frac{(N-k-1)(1-w_{00})}{\sigma\,k}$$

∎

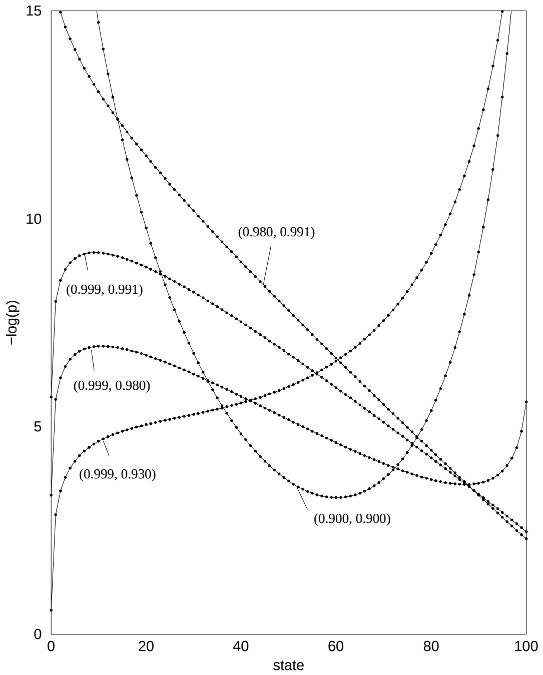


Figure 14: Different modalities of the stationary probability distribution. For $N = 100$ and superior fitness $\sigma = 1.1$ depending on the tuple (w_{00}, w_{11}) we present the distributions for the number of masters.

In the case of infinite population size when a completely deterministic ansatz can be applied an error threshold p^* fulfilling $W_{\mu\mu}(p^*) = 1/\sigma$ was shown to exist where in the long time limit the information of the master network is lost [57]. For a finite population size such a threshold can be defined in consistency with Nowak and Schuster [53]. For $w_{00} \approx 1$ we derive the value p^* such that

$$W_{\mu\mu}(p^*) = \min_{0\leq k\leq N} \frac{(k+1)(N+(\sigma-1)k)(N-1+(\sigma-1)k)}{\sigma N(N+(\sigma-1)k+\sigma)\,k}$$

which coincides with the threshold value where for the first time stationary probability p_0 becomes greater than p_i $(i = 1,\ldots,N)$. It can be verified that $\lim_{N\to\infty} W_{\mu\mu}(p^*) = 1/\sigma$.

4.3. Two Neutral Networks – Double Shape Landscape

In section 3.3 we have elaborated that any two dense and connected neutral networks of RNA secondary structures come very close in sequence space since there always exists a set of sequences being compatible with both structures. Any evolutionary event leading to a better adapted species requires a change in phenotype – secondary structure. A change in secondary structure is equivalent to a transition from one neutral network to another one. It turns out that sequences belonging to the intersection play a crucial role in evolutionary optimization.

Beside the aspect of better adaptation evolutionary search can take place over a set of neutral or relatively neutral phenotypes. One may ask what happens if more than one secondary structure has the same fitness value.

For our considerations we will restrict to one pair of secondary structures. On the molecular level we show that there is a nontrivial *coevolution* of two phenotypes in the space of secondary structures. This phenomenon gives new insights in *neutral evolution*.

We define an artificial landscape as follows: Let $\Gamma_n[s]$ and $\Gamma_n[s']$ two neutral networks. They induce a fitness function

$$f_{s,s'}(v) := \begin{cases} \sigma; & v \in \mathrm{v}[\Gamma_n[s]] \\ \sigma; & v \in \mathrm{v}[\Gamma_n[s']] \\ 1; & v \in \mathrm{v}[\mathcal{Q}_\alpha^n] \setminus (\mathrm{v}[\Gamma_n[s]] \cup \mathrm{v}[\Gamma_n[s']]) \end{cases}$$

where $\sigma > 1$. We call this landscape *double shape landscape*.

The fitness function $f_{s,s'}$ implies a disjoint partition of a population $\mathbf{V}$ of the form:

$$\begin{aligned} \mathbf{V}_{\mu_1} &:= \{v \in \mathbf{V} \mid \ v \in \mathrm{v}[\Gamma_n[s]]\,\} \\ \mathbf{V}_{\mu_2} &:= \{v \in \mathbf{V} \mid \ v \in \mathrm{v}[\Gamma_n[s']]\,\} \\ \mathbf{V}_{\nu} &:= \{v \in \mathbf{V} \mid \ v \notin \mathrm{v}[\Gamma_n[s]] \wedge v \notin \mathrm{v}[\Gamma_n[s']]\,\} \end{aligned}$$

whence $\mathbf{V} = \mathbf{V}_{\mu_1} \cup \mathbf{V}_{\mu_2} \cup \mathbf{V}_{\nu}$. We call the elements of $\mathbf{V}_{\mu_1}$ and $\mathbf{V}_{\mu_2}$ *masters* because of their superior fitness and those of $\mathbf{V}_{\nu}$ *non masters*.

Before we study the time evolution of a population $\mathbf{V}$ in a double shape landscape we have to determine the phenotypic fixation and transition probabilities.

4.3.1. Phenotypic Fixation and Transition Probabilities

First we have to recall and introduce some notations

- 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
$$d(\Gamma_n[s], v) := |\{[v_i, v_k] \mid [v_i, v_k] \notin \Pi \vee [i,k] \in \Pi(s)\}|$$
where Π is the pairing rule of the underlying alphabet and $\Pi(s)$ the set of contacts of s [56].
- By setting $\mathcal{E}_{ik}(s,s') := \{v \,|\, d(\Gamma_n[s], v) = \wedge\, d(\Gamma_n[s'], v) = k\}$ a natural decomposition of the hypercube with respect to two secondary structures s and s' is given.

We can formulate the following lemma.

Lemma 4 *Let s and s' be two secondary structures, $\mathcal{A}$ be a $\star$-alphabet. Denote the number of common pairs in s and s' by n_0. Then*

$$Prob\{\sigma \in \mathcal{E}_{0k}(s,s') \,|\, \sigma \in \mathbf{C}[s]\} = \binom{n'_p - n_0}{k} \left(\frac{\alpha - 1}{\alpha}\right)^k \left(\frac{1}{\alpha}\right)^{n'_p - n_0 - k} .$$

For any two neutral networks $\Gamma_n[s], \Gamma_n[s']$ we define

$$C_{i,s'}[s] = \{v \,|\, v \in \Gamma_n[s] \,\wedge\, d(\Gamma_n[s'], v) = i\}.$$

Then $\Delta_{i,s'}[s] = |C_{i,s'}[s]|/|\Gamma_n[s]|$ can be called the *density of distance class* $\mathcal{E}_{0i}(s,s')$ in $\Gamma_n[s]$. Provided $\Gamma_n[s]$ and $\Gamma_n[s']$ are regular networks it is determined by

$$\Delta_{i,s'}[s] = (\alpha - 1)^i \left(\frac{1}{\alpha}\right)^{n'_p - n_0} \binom{n'_p - n_0}{i}, \quad i = 0, \ldots, n'_p - n_0 \qquad (**)$$

Finally we can state the following lemma

Lemma 5 *Suppose* $\Gamma_n[s] < \mathcal{Q}_\alpha^n$ *and* $\Gamma_n[s'] < \mathcal{Q}_\alpha^n$ *are two fixed neutral networks and* $\mathcal{A}$ *is a* $\star$*-alphabet. Suppose that we have a random mapping* $v = (x_1, \ldots, x_n) \mapsto v' = (x'_1, \ldots, x'_n)$, $v, v' \in v[\mathcal{Q}_\alpha^n]$ *that is defined as follows:* $x_i = x'_i$ *with probability* $1 - p$ *and* $x_i \neq x'_i$ *with uniform probability* p*. Then the phenotypic fixation probability on* $\Gamma_n[s]$ *is given by*

a)

$$\begin{aligned} W_{\mu_1\mu_1}(p) = \quad & (1-p)^n + [(1-\pi)\lambda_u + \pi\rho\lambda]\,[1-(1-p)^{n_u}]\,(1-p)^{2n_p} \\ & +[(1-\pi)\lambda_p + \pi\rho\lambda](1-p)^{n_u}\left[\left(\frac{p^2}{\alpha-1}+(1-p)^2\right)^{n_p} - (1-p)^{2n_p}\right] \\ & +[(1-\pi)\bar\lambda + \pi\rho\lambda]\,(1-(1-p)^{n_u})\left[\left(\frac{p^2}{\alpha-1}+(1-p)^2\right)^{n_p} - (1-p)^{2n_p}\right] \end{aligned}$$

with $\lambda = \bar\lambda/(\bar\lambda + \bar\lambda')$ *and* $\pi = (1/\alpha)^{n'_p - n_0}$.

b) and the transition probability from $\Gamma_n[s]$ *to* $\Gamma_n[s']$ *by*

$$\begin{aligned} W_{\mu_1\mu_2}(p) = \quad & ((1-\pi')\bar\lambda' + \rho\,\pi'\lambda')\left\{\Delta_{0,s'}[s]\left[-(1-p)^n + \left(\frac{p^2}{\alpha-1}+(1-p)^2\right)^{n'_p}\right]\right. \\ & \left. + \sum_{i=1}^{n'_p - n_0} \Delta_{i,s'}[s]\left(\frac{p\,(1-p)}{\alpha-1} + \frac{p^2\,(\alpha-2)}{(\alpha-1)^2}\right)^i \left(\frac{p^2}{\alpha-1}+(1-p)^2\right)^{n'_p - i}\right\} \end{aligned}$$

with $\lambda' = 1 - \lambda$ *and* $\pi' = (1/\alpha)^{n_p - n_0}$.

Proof:

a) Denoting an error in the unpaired positions by (–,.) and in the paired positions by (.,–) one has to distinguish between four types of errors. The probability for (+,+) is simply $(1-p)^n$. For (–,+) we get

$$(1-p)^{2n_p} \sum_{j=1}^{n_u} \binom{n_u}{j} p^j (1-p)^{n_u - j} = (1-p)^{2n_p}[1-(1-p)^{n_u}]$$

The probability for the new element to be a μ_1 is $(1-\pi)\lambda_u$ for non intersection sequences and $\pi\rho\lambda$ for intersection elements. Similar arguments run for (+,–) and (+,+) and the formulae can be computed.

b) An arbitrary sequence in $\Gamma_n[s]$ and distance class $\mathcal{E}_{0i}$ with $1 \leq i \leq n'_p$ can be arranged in the form

$$\underbrace{x_1, x_2, \ldots, x_{n'_u}}_{n'_u}\ \underbrace{(y_1, y'_1), \ldots, (y_i, y'_i)}_{i}\ \underbrace{(z_1, z'_1), \ldots, (z_{n'_p - i}, z'_{n'_p - i})}_{n'_p - i}$$

where we have i incompatible pairs and $n'_p - i$ compatible pairs.

An incompatible base pair has to undergo mutations to become compatible. This can be realized in two different ways. The first is to mutate one position and to keep the other unmutated. The probability to happen and to obtain a compatible base pair is $p\,(1-p)/(\alpha-1)$. The second way is to mutate in both positions. The probability to happen and to obtain a compatible base pair is $p^2(\alpha-2)/(\alpha-1)^2$. Therefore i incompatible base pairs have the probability to become compatible

$$\sum_{i_o=0}^{i} \binom{i}{i_o} \left(\frac{p\,(1-p)}{\alpha-1}\right)^{i_o} \left(\frac{p^2(\alpha-2)}{(\alpha-1)^2}\right)^{i-i_o} = \left(\frac{p\,(1-p)}{\alpha-1} + \frac{p^2(\alpha-2)}{(\alpha-1)^2}\right)^{i}$$

Suppose there is no incompatible base pair. Then at least one mutation is required. For this we obtain the probability

$$\begin{aligned}
&[1-(1-p)^{n'_u}]\,(1-p)^{2n'_p} + [1-(1-p)^{n'_u}] \left\{ \left[\frac{p^2}{\alpha-1} + (1-p)^2\right]^{n'_p} - (1-p)^{2n'_p} \right\} \\
&+(1-p)^{n'_u} \left\{ \left[\frac{p^2}{\alpha-1} + (1-p)^2\right]^{n'_p} - (1-p)^{2n'_p} \right\} \\
&= -(1-p)^{n} + \left[\frac{p^2}{\alpha-1} + (1-p)^2\right]^{n'_p}
\end{aligned}$$

Taking into account that common base pairs are always compatible the sum only runs from 1 to $n'_p - n_0$. The probability for the new elements to be a μ_2 is $(1-\pi')\bar{\lambda}'$ for non intersection and $\pi'\lambda'$ for intersection sequences.

∎

Let

$$C_{ik}[s,s'] := \{v \mid v \notin \mathrm{v}[\Gamma_n[s]] \wedge v \notin \mathrm{v}[\Gamma_n[s']], \quad d(\Gamma_n[s],v) = i,\ d(\Gamma_n[s'],v) = k\}$$

be the (i,k) th *incompatible class*.

For regular networks $\Gamma_n[s]$ and $\Gamma_n[s']$ holds

$$|C_{ik}[s,s']| = \begin{cases} \binom{n_p-n_0}{i}\binom{n'_p-n_0}{k}\alpha^{n+n_0-n_p-n'_p}(\alpha-1)^{i+k} & \text{iff} \quad i,k \neq 0 \\ \binom{n_p-n_0}{i}\alpha^{n+n_0-n_p-n'_p}(\alpha-1)^{i} - |\Gamma_{i0}[s']| & \text{iff} \quad k=0; i\neq 0 \\ \binom{n'_p-n_0}{k}\alpha^{n+n_0-n_p-n'_p}(\alpha-1)^{k} - |\Gamma_{0k}[s]| & \text{iff} \quad i=0; k\neq 0 \\ \alpha^{n+n_0-n_p-n'_p} - |\Gamma_{00}[s]| - |\Gamma_{00}[s']| & \text{iff} \quad i=0; k=0 \end{cases}$$

where $\Gamma_{ik}[s] := \mathcal{E}_{ik}(s,s') \cap \mathrm{v}[\Gamma_n[s]]$. Now consequently $\Delta_{ik}[s,s'] = |C_{ik}[s,s']|/(|\mathcal{Q}^n_\alpha| - |\Gamma_n[s]| - |\Gamma_n[s']|)$.

Lemma 6 *Suppose $\Gamma_n[s]$ and $\Gamma_n[s']$ to be two fixed neutral networks and $\mathcal{A}$ is a $\star$–alphabet. Suppose that we have a random mapping $v = (x_1, \dots, x_n) \mapsto v' = (x'_1, \dots, x'_n)$, $v, v' \in v[\mathcal{Q}^n_\alpha]$ that is defined as follows: $x_i = x'_i$ with probability $1 - p$ and $x_i \neq x'_i$ with uniform probability p. Then corresponding to Lemma 5 we find*

c)

$$
\begin{aligned}
W_{\nu\mu_1}(p) = \quad & ((1-\pi)\bar{\lambda} + \pi\rho\lambda) \sum_{k=0}^{n'_p - n_0} \left\{ \Delta_{0k}[s,s'] \left[-(1-p)^n + \left(\frac{p^2}{\alpha - 1} + (1-p)^2 \right)^{n_p} \right] \right. \\
& \left. + \sum_{i=1}^{n_p - n_0} \Delta_{ik}[s,s'] \left(\frac{p\,(1-p)}{\alpha - 1} + \frac{p^2(\alpha-2)}{(\alpha-1)^2} \right)^i \left(\frac{p^2}{\alpha - 1} + (1-p)^2 \right)^{n_p - i} \right\}
\end{aligned}
$$

with $\lambda = \bar{\lambda}/(\bar{\lambda} + \bar{\lambda}')$.

Proof: The proof runs with the same arguments as *b)* in the previous lemma. ∎

In the sequel we write for short w_{ik} for the $W_{\cdot,\cdot}$ with $i, k = 0, 1, 2$, where 0 corresponds to a non master, 1 to a master on $\Gamma_n[s]$ and 2 to a master on $\Gamma_n[s']$. Obviously we have $w_{i0} + w_{i1} + w_{i2} = 1$ for $i = 0, 1, 2$.

4.3.2. Deterministic Approach

The evolutionary dynamics of a population of erroneously replicating strings can be described by a system of ordinary differential equations if the population size N is assumed to be sufficiently large. Let the underlying dynamics be determined by a double shape landscape. Thus we shall not be interested in the concentration of a single sequence but in the concentration of masters on the dominating networks and non-masters respectively. If the rates derived in the last section give a good approximation for mutation events and their results, the over-all system of ordinary differential equations can be reduced to a set of three coupled ODE's given by:

$$\dot{x}_i = w_{0i}\, x_0 + \sigma w_{1i}\, x_1 + \sigma w_{2i}\, x_2 - x_i\, \Phi(t), \qquad i = 0, 1, 2 \tag{10}$$

where x_0 denotes the concentration of non-masters, x_1 that of masters on network $\Gamma_n[s]$, and x_2 that of masters on network $\Gamma_n[s']$. $\Phi(t)$ is the non-specific dilution flux keeping the concentration in the flow reactor constant. It is given by $x_0 + \sigma(x_1 + x_2)$ if $\sum x_i = 1$. Note that the system above corresponds to a system of replication mutation equations [71].

Substituting $x_0 = 1 - x_1 - x_2$, $z = x_1 + x_2$ and $x_1 = x$ the system (10) turns out to be equivalent to:

$$\begin{aligned} \dot{x} &= (-1 + \sigma w_{11} - \sigma w_{21})x - (\sigma - 1)x\,z + (-w_{01} + \sigma w_{21})z + w_{01} \\ \dot{z} &= -(\sigma - 1)z^2 + (-2 + \sigma + w_{00} - \sigma w_{20})z + \sigma(-w_{10} + w_{20})x + (1 - w_{00}) \end{aligned} \tag{11}$$

The range of meaningful or acceptable solutions is $\mathcal{D} = \{(x, z)\,|\,0 \le x \le z \le 1\}$. The boundary of the domain $\mathcal{D}$ is

$$\partial\mathcal{D} = \{x = 0, 0 \le z \le 1\} \cup \{0 \le x \le 1, z = 1\} \cup \{x = z, 0 \le x \le 1\} \quad .$$

System (11) possess three fixed points but it is known that acceptable stationary solutions have to comply with the relation $0 \le x_s \le z_s \le 1$. Before considering the general case we start with investigating degenerated cases where one or more of the rates w_{ik}, $i, k = 0, 1, 2$ vanish.

1. Let $w_{01} = w_{02} = w_{12} = w_{21} = 0$. We will call this constellation *double decoupled case* because back-flow from the non-masters to the networks and flow between the networks is eliminated. Three fixed points can be derived for the system being

	fixed point	**eigenvalues**
1.	$(0, 0)$	$\lambda_1 = -(1 - \sigma\, w_{11}), \quad \lambda_2 = -(1 - \sigma\, w_{22})$
2.	$\left(0, \frac{-(1 - \sigma\, w_{22})}{\sigma - 1}\right)$	$\lambda_1 = \frac{w_{11} - w_{22}}{w_{22}}, \quad \lambda_2 = \frac{1 - \sigma\, w_{22}}{\sigma w_{22}}$
3.	$\left(\frac{-(1 - \sigma\, w_{11})}{\sigma - 1}, \frac{-(1 - \sigma\, w_{11})}{\sigma - 1}\right)$	$\lambda_1 = \frac{1 - \sigma\, w_{11}}{\sigma w_{11}}, \quad \lambda_2 = \frac{-(w_{11} - w_{22})}{w_{11}}$

Denoting an acceptable fixed point by a plus `(+)` and a non-acceptable one by a minus `(-)` and furthermore asymptotically stability by a plus `(+)` and instability by a minus

(-) we get for the triple of fixed points

relation	**acceptability**	**stability**
$w_{11} < w_{22} < 1/\sigma$	(+,-,-)	(+, ,)
$w_{22} < w_{11} < 1/\sigma$	(+,-,-)	(+, ,)
$1/\sigma < w_{11} < w_{22}$	(+,+,+)	(-,+,-)
$w_{11} < 1/\sigma < w_{22}$	(+,+,-)	(-,+,)
$1/\sigma < w_{22} < w_{11}$	(+,+,+)	(-,-,+)
$w_{22} < 1/\sigma < w_{11}$	(+,-,+)	(-, ,+)

The double decoupled case never tends to coexistence of both masters. The dominating sequences on the networks go extinct if their corresponding rates w_{ii} are below the value $1/\sigma$. Otherwise the network having the higher rate will survive and the other one will go extinct. If one relation given above is fulfilled by the equal sign a bifurcation occurs and nothing can be said about stability without considering terms of higher order.

2. Let $w_{01} = w_{02} = 0$ and $w_{10} = w_{20}$. Back-flow from the non-masters to the networks is eliminated and the flow from the master networks to the non-masters is assumed to be of the same magnitude. Then two fixed points can be found for the system being

	fixed point	**eigenvalues**
1.	$(0,0)$	$\lambda_1 = (\sigma - 1) - \sigma w_{10} - \sigma(w_{12} + w_{21})$, $\lambda_2 = (\sigma - 1) - \sigma w_{10}$
2.	$\left(\frac{(\sigma - 1 - \sigma w_{10}) w_{21}}{(\sigma - 1)(w_{12} + w_{21})}, \frac{\sigma - 1 - \sigma w_{10}}{\sigma - 1}\right)$	$\lambda_1 = \frac{-(w_{12}+w_{21})}{1-w_{10}}$, $\lambda_2 = \frac{-(\sigma-1)+\sigma w_{10}}{\sigma(1-w_{10})}$

For $0 \leq w_{10} < (\sigma - 1)/\sigma$ fixed point 2. is acceptable and asymptotically stable. Hence the system tends to a coexisting state of masters on $\Gamma_n[s]$ and $\Gamma_n[s']$. $w_{10} = (\sigma - 1)/\sigma$ is a bifurcation value and nothing can be said about stability without considering terms of higher derivation. Finally for $w_{10} > (\sigma - 1)/\sigma$ fixed point 1. becomes asymptotically stable, i.e., the masters go extinct.

3. Let $w_{01} = w_{21} = 0$. According to point (1.) we will call this constellation the *single*

decoupled case. Under this circumstances we find the following fixed points

	fixed point
1.	$\left(0,\ \dfrac{-1-w_{02}+\sigma\, w_{22}-\sqrt{(-1-w_{02}+\sigma\, w_{22})^2+4(\sigma-1)w_{02}}}{2\,(\sigma-1)}\right)$
2.	$\left(0,\ \dfrac{-1-w_{02}+\sigma\, w_{22}+\sqrt{(-1-w_{02}+\sigma\, w_{22})^2+4(\sigma-1)w_{02}}}{2\,(\sigma-1)}\right)$
3.	$\left(\dfrac{-w_{20}(\sigma w_{11}-1)+(1-w_{11})(\sigma w_{11}-w_{00})}{(\sigma-1)(w_{10}-w_{20})},\ \dfrac{-(1-\sigma w_{11})}{\sigma-1}\right)$

Fixed point 1. has a component smaller than zero and therefore it is non-acceptable whereas fixed point 2. is always acceptable. It becomes asymptotically stable if one of the following relations is fulfilled:

$$(i)\quad w_{11} < 1/\sigma \quad \text{or}$$

$$(ii)\quad w_{11} > 1/\sigma \wedge w_{00} > \sigma(2w_{11}-w_{22}) \quad \text{or}$$

$$(iii)\quad w_{11} > 1/\sigma \wedge w_{00} < \sigma(2w_{11}-w_{22}) \wedge w_{20} < \frac{(1-w_{11})(\sigma w_{11}-w_{00})}{\sigma w_{11}-1} \quad .$$

Further on fixed point 3. becomes an acceptable solution if and only if $w_{11} > 1/\sigma$ is fulfilled and additionally one of the following inequalities holds

$$(i)'\quad w_{10} < \frac{(1-w_{11})(\sigma w_{11}-w_{00})}{\sigma w_{11}-1} < w_{20} \quad \text{or}$$

$$(ii)'\quad w_{20} < \frac{(1-w_{11})(\sigma w_{11}-w_{00})}{\sigma w_{11}-1} < w_{10} \quad .$$

It is found to be asymptotically stable if and only if $w_{10} < w_{20}$ and $w_{00} < \sigma(2w_{11}-w_{22})$. Despite neglecting back-flow to the dominant network $\Gamma_n[s]$ it is able to coexist with the other network if its rate w_{11} is greater than the reciprocal value of its fitness.

Investigation of the isoclines corresponding to system (11)

$$x_h(z) = \frac{w_{01}(1-z)+\sigma w_{21} z}{1-z+\sigma(w_{21}-w_{11}+z)};\ x_v(z) = \frac{(\sigma-1)z(1-z)+(1-w_{00})(1-z)-\sigma w_{20} z}{\sigma(w_{10}-w_{20})}$$

shows that there always exists an acceptable and stable fixed point. It is a single one as long as no rate w_{ik} becomes zero. In this case the fixed points outside the acceptable range may move to the boundary ∂D. As shown for these *decoupled cases* always exists a single asymptotically stable fixed point. In figure 15 we show the graphs of the isoclines x_h, x_v for $w_{11} = 0.5, w_{22} = 0.6, w_{01} = w_{02} = 0.0005$ and $w_{12} = w_{21} = 0.05$. Additionally for the same parameters and increasing $w_{12} = w_{21} = 0.0001$ to 0.4 we monitor one fixed point outside and the course of the stable fixed point inside the acceptable range.

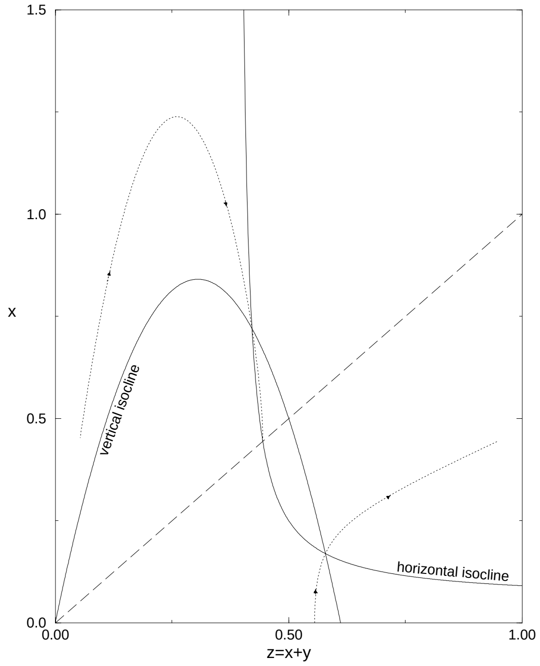


Figure 15: Isoclines and fixed points. For $w_{11} = 0.5, w_{22} = 0.6, w_{01} = w_{02} = 0.0005$ the course of the isocline x_h and x_v is shown for $w_{12} = w_{21} = 0.05$. For $w_{12} = w_{21} = 0.0001$ to 0.4 the dotted lines indicate (i) one fixed point outside and (ii) the acceptable stable fixed point inside. The long-dashed line gives the upper boundary of the domain of meaningful solutions.

4.3.3. Stochastic Approach

In this section we intend to study the variables X_t and Y_t counting the number of elements on $\Gamma_n[s]$, $\Gamma_n[s']$ for any given time t. Therefore we shall embed X_t and Y_t in the context of probability theory. A probability space is defined to be a triple $(\Omega, \mathcal{A}, \mu)$ where in our case $\Omega = \{(x,y) \,|\, x, y \in \{0, \ldots, N\},\, x + y \leq N\}$ is the set of elementary events, $\mathcal{A}$ is a σ–algebra on Ω and μ a probability measure. X_t and Y_t are two dependent random variables defined on Ω in the following way

$$\begin{array}{lll} X_t:\ \Omega \to \mathbb{N}, & X_t^{-1}(x) \ = A_x = & \{(x,i) \mid i = 0,\dots,N-x\} \\ Y_t:\ \Omega \to \mathbb{N}, & Y_t^{-1}(y) \ = A_y = & \{(i,y) \mid i = 0,\dots,N-y\} \end{array}$$

such that $\mu(A_x) = P(X_t = x)$ and $\mu(A_y) = P(Y_t = y)$. Applying the theory of conditional probability it holds

$$P(X_t = x,\, Y_t = y) = P(X_t = x \mid X_t + Y_t = x+y) \cdot P(X_t + Y_t = x + y). \qquad (12)$$

Note that $\mu((x,y)) = P(X_t = x,\, Y_t = y)$. Because of the σ–additivity of probability measures we obtain $P(X_t = x) = \sum_{i=0}^{N-x} P(X_t = x, Y_t = i)$ and $P(Y_t = y) = \sum_{i=0}^{N-y} P(X_t = i, Y_t = y)$. For short we write

$$P(x,y;t) = P(X_t = x,\, Y_t = y),$$

$$P_X(x|z;t) = P(X_t = x|X_t + Y_t = z) \quad \text{and} \quad P_Y(y|z;t) = P(Y_t = y|X_t + Y_t = z).$$

We now shall approximate the replication deletion process acting on a double shape landscape by a homogeneous Markov process in continuous time with finitely many states.

Let $P_{(x,y)(x',y')}(h) = P((X_{t+h}, Y_{t+h}) = (x',y')|(X_t, Y_t) = (x,y))$ then the transition probabilities are given by

$$\begin{cases} 0 & 0 < x, y; x+y > N \\ \alpha(x,y)\, h + o(h) & (x',y') = (x+1,y) \\ \beta(x,y)\, h + o(h) & (x',y') = (x+1,y-1) \\ \gamma(x,y)\, h + o(h) & (x',y') = (x,y+1) \\ \delta(x,y)\, h + o(h) & (x',y') = (x,y-1) \\ \varepsilon(x,y)\, h + o(h) & (x',y') = (x-1,y+1) \\ \varphi(x,y)\, h + o(h) & (x',y') = (x-1,y) \\ 1 - (\alpha+\beta+\gamma+\delta+\varepsilon+\varphi)(x,y)\, h + o(h) & (x',y') = (x,y) \\ o(h) & \text{otherwise} \end{cases} \qquad (13)$$

with

$$\begin{array}{rcl} \alpha(x,y) & = & \dfrac{N-x-y}{(N-1)(N+(x+y)(\sigma-1))}(\sigma\, w_{11}x + \sigma\, w_{21}y + w_{01}(N-x-y-1)) \\ \beta(x,y) & = & \dfrac{y}{(N-1)(N+(x+y)(\sigma-1))}(\sigma\, w_{11}x + \sigma\, w_{21}(y-1) + w_{01}(N-x-y)) \\ \gamma(x,y) & = & \dfrac{N-x-y}{(N-1)(N+(x+y)(\sigma-1))}(\sigma\, w_{12}x + \sigma\, w_{22}\, y + w_{02}(N-x-y-1)) \\ \delta(x,y) & = & \dfrac{y}{(N-1)(N+(x+y)(\sigma-1))}(\sigma\, w_{10}x + \sigma\, w_{20}(y-1) + w_{00}(N-x-y)) \\ \varepsilon(x,y) & = & \dfrac{x}{(N-1)(N+(x+y)(\sigma-1))}(\sigma w_{12}(x-1) + \sigma\, w_{22}y + w_{02}(N-x-y)) \\ \varphi(x,y) & = & \dfrac{x}{(N-1)(N+(x+y)(\sigma-1))}(\sigma\, w_{10}(x-1) + \sigma\, w_{20}y + w_{00}(N-x-y)) \end{array}$$

Figure 16 illustrates the state space where the random process is living in and its accepted infinitesimal transitions.

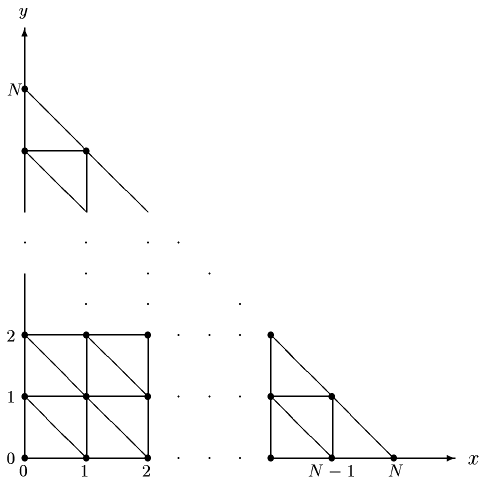


Figure 16: State space and accepted infinitesimal transitions. States of the stochastic process are tuples of natural numbers (x, y) fulfilling $0 \leq x, y \leq N$ and $x + y \leq N$. The grid indicates acceptable infinitesimal transitions.

Figure 16 intuitively makes clear that for any two states $z = (x, y)$ and $z' = (x', y')$ there exists a time t^* such that $P_{(z,z')}(t^*) > 0$. Precisely by induction on $|x - x'|$ and $|y - y'|$ it be verified that:

1.) if $x - x' > 0$, $y - y' > 0$ then

$$P_{z,z'}(h\,(x - x' + y - y')) \geq \prod_{\ell=0}^{x-x'-1} P_{(x-\ell,y)(x-\ell-1,y)}(h) \cdot \prod_{\ell=0}^{y-y'-1} P_{(x',y-\ell)(x',y-\ell-1)}(h) > 0$$

2.) if $x - x' < 0$, $y - y' < 0$ then

$$P_{z,z'}(h\,(x' - x + y' - y)) \geq \prod_{\ell=0}^{x'-x-1} P_{(x+\ell,y)(x+\ell+1,y)}(h) \cdot \prod_{\ell=0}^{y'-y-1} P_{(x',y+\ell)(x',y+\ell+1)}(h) > 0$$

3. if $x - x' < 0$, $y - y' > 0$ then

$$P_{z,z'}(h\,(x' - x + y - y')) \geq \prod_{\ell=0}^{x'-x-1} P_{(x+\ell,y)(x+\ell+1,y)}(h) \cdot \prod_{\ell=0}^{y-y'-1} P_{(x',y-\ell)(x',y-\ell-1)}(h) > 0$$

4.) if $x - x' > 0$, $y - y' < 0$ then

$$P_{z,z'}(h\,(x - x' + y' - y)) \geq \prod_{\ell=0}^{x-x'-1} P_{(x-\ell,y)(x-\ell-1,y)}(h) \cdot \prod_{\ell=0}^{y'-y-1} P_{(x',y+\ell)(x',y+\ell+1)}(h) > 0.$$

From the ergodic theorem 4 stated in section 4.1 the existence of $P(x,y) \overset{\text{def}}{=} \lim_{t\to\infty} P(x,y;t)$ can be deduced. Hence the following relation can be derived from equation (12)

$$P_X(x|x+y) = P(x,y)/\sum_{i=0}^{x+y} P(i,x+y-i) \tag{14}$$

For further considerations it turns out to be useful to substitute $z \overset{\text{def}}{=} x+y$ and to introduce

$$\bar{P}(x,z) := P(x,z-x);\ \bar{\alpha}(x,z) := \alpha(x,z-x), \ldots, \bar{\varphi}(x,z) := \varphi(x,z-x) \qquad .$$

Then for $0 \le x \le z \le N$ the stationary master equation corresponding the defined Markov process can be written as

$$\bar{\mathcal{J}}(x,z) - \bar{\mathcal{J}}(x-1,z-1) + \bar{\mathcal{H}}(x,z) - \bar{\mathcal{H}}(x-1,z) + \bar{\mathcal{K}}(x,z) - \bar{\mathcal{K}}(x,z-1) = 0 \tag{15}$$

with

$$\begin{aligned}
\bar{\mathcal{J}}(x,z) &:= \bar{P}(x+1,z+1)\,\bar{\varphi}(x+1,z+1) - \bar{P}(x,z)\,\bar{\alpha}(x,z),\\
\bar{\mathcal{H}}(x,z) &:= \bar{P}(x+1,z)\,\bar{\varepsilon}(x+1,z) - \bar{P}(x,z)\,\bar{\beta}(x,z),\\
\bar{\mathcal{K}}(x,z) &:= \bar{P}(x,z+1)\,\bar{\delta}(x,z+1) - \bar{P}(x,z)\,\bar{\gamma}(x,z).
\end{aligned}$$

Summing the set of equations (15) in an appropriate way yields

$$\sum_{x=0}^{z} \bar{\mathcal{J}}(x,z) + \bar{\mathcal{K}}(x,z) = 0 \tag{16}$$

which is equivalent to

$$\sum_{x=0}^{z+1} \bar{P}(x,z+1)\,(\bar{\varphi}(x,z+1) + \bar{\delta}(x,z+1)) = \sum_{x=0}^{z} \bar{P}(x,z)\,(\bar{\alpha}(x,z) + \bar{\gamma}(x,z)) \qquad . \tag{17}$$

Considering the coefficients of equation (17) we get

$$\bar{\varphi}(x,z) + \bar{\delta}(x,z) = g_1(z)\,x + g_2(z) \quad \text{and} \quad \bar{\alpha}(x,z) + \bar{\gamma}(x,z) = f_1(z)\,x + f_2(z)$$

where

$$g_1(z) = \frac{(z-1)(\Lambda_{10}-\Lambda_{20})}{(N-1)(N+z(\sigma-1))}, \quad g_2(z) = \frac{z(z-1)\Lambda_{20} + w_{00}(N-1)\,z}{(N-1)(N+z(\sigma-1))}$$

$$f_1(z) = \frac{(N-z)(\Lambda_{20}-\Lambda_{10})}{(N-1)(N+z(\sigma-1))}, \quad f_2(z) = \frac{(N-z)(z(-\Lambda_{20}+\sigma-1)+(N-1)(1-w_{00})}{(N-1)(N+z(\sigma-1))}$$

with $\Lambda_{ik} := \sigma\, w_{ik} - w_{0k},\ i,k = 0,1,2.$

Hence equation (17) becomes

$$g_1(z+1)\, Q_M(z+1) + g_2(z+1)\, Q(z+1) = f_1(z)\, Q_M(z) + f_2(z)\, Q(z) \tag{18}$$

with $Q_M(z) := \sum_{x=0}^{z} x\, \bar{P}(x,z)$ and $Q(z) := \sum_{x=0}^{z} \bar{P}(x,z)$. Note that $g_1(.)$ and $f_1(.)$ vanish if $w_{10} = w_{20}$.

Using the relation given by equation (14) the term $Q_M(z)$ can be written as

$$Q_M(z) = Q(z) \sum_{x=0}^{z} x\, P_X(x \mid z) = Q(z)\, \tilde{Q}(z).$$

Applying this and equation (18) we find a recursion formula for $Q(.)$ of the following form

$$Q(z+1) = Q(z)\, \frac{f_1(z)\tilde{Q}(z) + f_2(z)}{g_1(z+1)\tilde{Q}(z+1) + g_2(z+1)}. \tag{19}$$

defined for $z = 0, \ldots, N-1$ and initial value $Q(0)$ chosen in that way that $\sum\limits_{z=0}^{N} Q(z) = 1$.

The solution for $Q(z)$ can be easily determined in the special case $w_{10} = w_{20}$. It has already been mentioned that under these circumstances $f_1(z)$ and $g_1(z)$ vanish. So we get

$$Q(0) = 1/(\sum_{i=0}^{N} \varpi(i)), \quad Q(z) = \varpi(z)/(\sum_{i=0}^{N} \varpi(i))$$

with

$$\varpi(0) = 1; \quad \varpi(z) = \left(1 + \frac{z}{N}(\sigma - 1)\right) \left(\frac{\Lambda_{10} - \sigma + 1}{\Lambda_{10}}\right)^{z} \frac{\mathcal{B}(b_2, N)}{\mathcal{B}(b_1, z) \cdot \mathcal{B}(b_2, N - z)} \tag{20}$$

where $b_1 = (N-1)w_{00}/(\sigma - 1 - \Lambda_{10})$ and $b_2 = -\sigma(N-1)w_{10}/\Lambda_{10}$.

Finally we are interested in the conditional probability $P_X(x \mid z)$. By definition of the process the random variable X_t follows an ordinary birth and death process if the number of masters is required to be constant. The birth rate is given by

$$\lambda(x \mid z) = \bar{\beta}(x,z) = \frac{z - x}{(N-1)(N + (\sigma - 1)z)}((\Lambda_{11} - \Lambda_{21})\, x + \Lambda_{21}(z-1) + (N-1)w_{01})$$

and the death rate by

$$\mu(x \mid z) = \bar{\varepsilon}(x, z) = \frac{x}{(N-1)(N+(\sigma-1)z)}((\Lambda_{12}-\Lambda_{22})(x-1)+\Lambda_{22}(z-1)+(N-1)w_{02}).$$

The stationary solution of this process can be completely determined

$$P_X(0 \mid z) = 1/\sum_{k=0}^{z} \pi(k \mid z) \quad \text{and} \quad P_X(x \mid z) = \pi(x \mid z)/\sum_{k=0}^{z} \pi(k \mid z). \tag{21}$$

with

$$\pi(0 \mid z) = 1 \quad \text{and} \quad \pi(k \mid z) = \left(\frac{\Lambda_{11}-\Lambda_{21}}{\Lambda_{22}-\Lambda_{12}}\right)^k \frac{\mathcal{B}(c_2, N)}{\mathcal{B}(c_1, k)\cdot\mathcal{B}(c_2, N-k)} \tag{22}$$

with $c_1 = \dfrac{\Lambda_{21}(z-1)+(N-1)w_{01}}{\Lambda_{11}-\Lambda_{21}}$ and $c_2 = \dfrac{\Lambda_{12}(z-1)+(N-1)w_{02}}{\Lambda_{22}-\Lambda_{12}}$ for $k = 1, \ldots, z$.

Given the stationary probability $P_X(x|z)$ for all $0 \leq x \leq z \leq N$ the function $\tilde{Q} \overset{\text{def}}{=\!=} \sum_{x=0}^{z} x P_X(x|z)$ is completely determined. Thus, applying recursion formula (19), the values $Q(z)$ can be derived. And finally putting things together we obtain the stationary probability distribution $P(x, y)$ from $P(x, y) = Q(x+y)P_X(x|x+y)$.

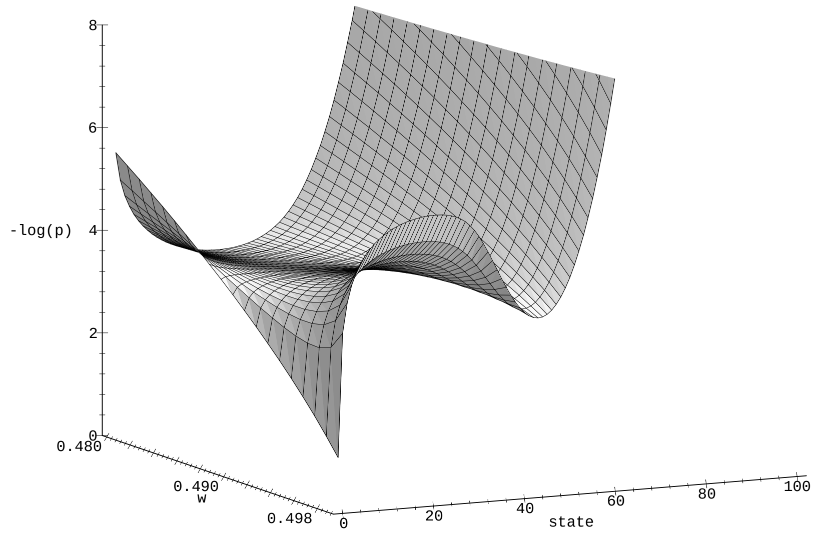


Figure 17: For population size 100 and mean values $w_{00} = 0.999$, $w_{02} = w_{01}$ and $w_{20} = w_{10} = 0.5$ the value $w = w_{22} = w_{11}$ is varied. The figure shows for increasing w the stationary distribution of X_t – the number of masters on $\Gamma_n[s]$.

In order to be consistent with section 4.1 and 4.2 we shall ask for the stationary distribution of the number of masters on network $\Gamma_n[s]$ (or $\Gamma_n[s']$ respectively). We immediately see that it is given by $P(X = x) = \sum_{z=x} Q(z) P_X(x|z)$. The system depends on several parameters. By changing only one of them its particular effects on the distribution can be detected. One example is presented in figure 17. Here the population size is $N = 100$. We fix the values: $w_{00} = 0.999$, $w_{02} = w_{01}$ and $w_{20} = w_{10} = 0.5$. Now we plot $P(X = x)$ depending on $w = w_{22} = w_{11}$ for increasing w. Again we observe the existence of a critical value for w – the phenotypic fixation probability – where the modality of the distribution changes from an unimodal to a bimodal one. Being confronted with an ensemble of parameters it is impossible to derive analytical expressions for these bifurcation values. But in principle the evolution of the master subpopulation can be explained in terms of the simple model developed in section 4.1. Thereafter first of all the probabilities of phenotypic fixation, $W_{\mu_i \mu_i}$, and secondly the probabilities of transitions between phenotypes, $W_{\mu_i \mu_k}$, are crucial for survival, extinction and revival of a master subpopulation.

Let $U(x)$ be a positive real valued function defined by $U(x) \overset{\text{def}}{=} -\log\left(P(X = x)\right)$ with $x = 0, 1, \ldots, N$. Then $U(x)$ can be interpreted as a potential function [27, 32]. A particle – representing the state of the neutral network – is moving in the potential driven by some random forces. An unimodal curve can be imagined as a basin. A particle on a point inside the basin is going to fall down and rest at the bottom – the minimum of the potential. From any point of inside the basin the particle returns to this point. We can say that this is the stable equilibrium position of the particle. In the case when the function $U(x)$ takes the form of a double well potential the particle might first be trapped in one local minimum. In order to go to the other local minimum it has to pass the barrier of the local maximum before. Both minima are stable equilibrium positions whereas the local maximum is an unstable equilibrium position of the particle. Clearly the higher the barrier the more time is needed for crossing. Thus from the height of the local maximum and the depth of the local minima it is possible to predict the evolutionary behavior of the process without performing it.

5. Neutral Evolution – Computational Results

The regularity of sequence to structure mappings as expressed by the properties of neutral networks has stimulated the study of neutral evolution on the level of RNA secondary structures. Computer simulations were performed to investigate several effects of neutral evolution [40]. We now study the dynamic behavior of populations evolving in various double shape landscapes. This might be considered as a shift in investigating neutral evolution to a more complex level. Neutrality on the level of shapes is combined with neutrality on the level of fitness.

For a comprehensive study we select three different ordered pairs of structures (s, s') from the shape space containing all RNA secondary structures. Evidently each pair is expected to induce a different kind of double-shape landscape.

	s	s'
$A.$	`........(((((((((....)))))))))`	`.........(((((((((...)))))))))`
$B.$	`........(((((((((....)))))))))`	`((((((((...))))))))((((...))))`
$C.$	`.......((((((((.......))))))))`	`((((((((.......)))))))).......`

Lateron these pairs will be simply presented by their capital letters given in the table.

We commence to apply the mathematical framework developed in 3.3 to these three pairs. First we consider the mapping of RNA secondary structures of chain length n into elements of the symmetric group S_n. For a fixed structure s its involution is denoted by $\pi(s)$. Then by writing $\pi(s)$ and $\pi(s') \in S_{30}$ as products of transpositions we get for

$A.$

$$\begin{aligned}\pi(s) &= \texttt{(9,30)(10,29)(11,28)(12,27)(13,26)(14,25)(15,24)(16,23)(17,22)}\\ \pi(s') &= \texttt{(10,30)(11,29)(12,28)(13,27)(14,26)(15,25)(16,24)(17,23)(18,22)}\end{aligned}$$

$B.$

$$\begin{aligned}\pi(s) &= \texttt{(9,30)(10,29)(11,28)(12,27)(13,26)(14,25)(15,24)(16,23)(17,22)}\\ \pi(s') &= \texttt{(1,19)(2,18)(3,17)(4,16)(5,15)(6,14)(7,13)(8,12)(20,30)(21,29)}\\ &\ \texttt{(22,28)(23,27)}\end{aligned}$$

C.

$\pi(s)$ = `(8,30)(9,29)(10,28)(11,27)(12,26)(13,25)(14,24)(15,23)`
$\pi(s')$ = `(1,23)(2,22)(3,21)(4,20)(5,19)(6,18)(7,17)(8,16)`

Any two involutions constitute a dihedral group. The operation of the group generated by $\pi(s), \pi(s')$ leads to a cycle decomposition that can be reordered with respect to the structures s, s'. Finally each pair of structures is associated with an orbit decomposition

A. $\langle \pi(s), \pi(s') \rangle$:

`(1) (2) (3) (4) (5) (6) (7) (8) (19) (20) (21)`
`(9,30,10,29,11,28,12,27,13,26,14,25,15,24,16,23,17,22,18)`

B. $\langle \pi(s), \pi(s') \rangle$:

`(1,19)(2,18)(5,15,24)(6,14,25)(7,13,26)(9,30,20)(10,29,21)`
`(3,17,22,28,11)(4,16,23,27,12,8)`

C. $\langle \pi(s), \pi(s') \rangle$:

`(2,22)(3,21)(4,20)(5,19)(6,18)(7,17)(9,29)(10,28)(11,27)`
`(12,26)(13,25)(14,24)(1,23,15)(16,8,30).`

These decompositions are seen to be different and consequently size and topology of the corresponding overlaps between the sets of compatible sequences are distinct. On the account that sequences on the overlap play a decisive role in evolutionary optimization we shall study their organization in more detail. Note that only for pairs A and C the interacting sets of compatible sequences have equal size for both structures. For pair B the cardinality of the corresponding sets of compatible sequences is larger for structure s than for s'.

Structure pair A is characterized by a large number of common unpaired positions. (11 orbits of length 1) and a long orbit of length 19. It can be shown that the graph of the overlap decomposes into large islands that can be connected by long paths in $\mathcal{C}[s]$ and $\mathcal{C}[s']$ respectively. Positions covered by orbits of length one are completely variable, i.e. a mutation of a sequence from the overlap in these positions yields a sequence being compatible with both structures, the intersection is not left. Eleven positions fulfill this condition. The remaining nineteen positions are restricted by the pairing rule of the (**AUGC**) alphabet. In the course of evolutionary optimization a frequent interaction between the neutral networks is expected if the population is located in a number of sequences on the overlap.

Structure pair B possesses orbits of length two up to six. That means by mutating a sequence on the overlap in a single position we always leave the intersection. There are only two orbits of length two and hence by a single pair exchange it is likely to leave the intersection, too. The overlap subgraph decomposes into a large number of isolated small islands. The different sizes of the sets of compatible sequences for s and s' implies that sequences on the overlap are more frequent in $\mathcal{C}[s']$ than in $\mathcal{C}[s]$. Therefore a population in evolutionary optimization is more likely to be located on the neutral network generated by s than on those of s'.

Structure pair C offers a large number of open orbits of length two (12) and no orbit of length one. That means by a single base exchange in a sequence on the overlap it is left, but by a pair exchange it is likely to stay on the intersection. The induced subgraph of the overlap consists of large islands that can be connected by short paths in $\mathcal{C}[s]$ and $\mathcal{C}[s']$ respectively. In evolutionary adaptation processes we expect a strong interaction of the neutral networks generated by s and s' if a number of sequences of the population is located on the overlap.

The next two sections are devoted to computer simulations of evolutionary dynamics determined by double shape landscapes induced by the three different pairs of structures. In general the model consists of a population of RNA sequences which replicate in a stirred flow reactor. When a sequence undergoes a replication, each base is copied with fidelity $1-p$. To ensure that an offspring, after replication, in terms of Hamming distance is close to its parent, depending also on the chain length, the single digit accuracy $1-p$ has to be chosen sufficiently large. Therefore in general we set p to 0.03. Finally a selection pressure is ensured by an unspecific dilution flux.

5.1. Neutral Evolution Based On RNA Folding

In this section we are concerned with the combinatory map $\varphi : \mathcal{Q}_\alpha^n \to S_n$ induced by the minimum free energy folding [36]. That means all RNA secondary structures under considerations are minimum free energy structures. Since only two neutral networks are of interest in the case of double-shape landscapes not all sequences need to be folded in an evolutionary run but only those being compatible with one or the other network.
First we mention some technical details that the following simulations have in common:

- The underlying sequence space is the 30-dimensional hypercube formed over the alphabet $\mathcal{A}$ =(**G**,**C**). The pairing rules are determined by $\mathcal{B}$ =(**GC**,**CG**).
- Since in general RNA population manageable in the laboratory are tiny compared to the size of sequence space ($2^{30} \approx 10^9$) the computer simulations are started with an initial population consisting of $N = 100$ sequences divided into 50 copies of a random sequence on network $\Gamma_n[s]$ and 50 copies of another random sequence on network $\Gamma_n[s']$.
- The superior fitness σ of master sequences is set to be 10 as opposed to 1 as minor fitness.
- When a sequence undergoes a replication each base is copied with single digit accuracy 0.97. This ensures point and pair mutations to be most likely whereas simultaneous exchanges of more than three nucleotides are improbable.

Simulations

Two evolutionary experiments based on (i) the flow reactor dynamics and (ii) on the replication-deletion process are executed on double-shape landscapes predefined by the structure pairs A, B and C. After each generation the number of sequences occupying network $\Gamma_n[s]$ and $\Gamma_n[s']$, respectively, is checked and stored. A simulation is terminated after 10^5 generations. On account of the ergodic property of the processes the stationary probability distribution for the number of sequences on the master networks can be evaluated by averaging over time. Figure 17, 18, and 19 display the results we received for structure pairs A, B, and C. Each figure is subdivided into two parts: the l.h.s. shows the results for network $\Gamma_n[s]$ whereas the r.h.s. displays those for $\Gamma_n[s']$. Circles coincide with the values we received from the flow reactor experiment whereas squares with those of the replication-deletion process. The solid line always presents the analytical curve for the actually investigated network whereas the dashed line corresponds to the other not depicted network.

In order to evaluate the analytical expression for the stationary probability distribution the phenotypic fixation and transition probabilities $W_{\mu_1\mu_1}, W_{\mu_1\mu_2}$, etc. need to be estimated. Except for the joint neutrality parameter ρ all input parameters $(\lambda_u[s], \lambda_u[s']$ etc.) had been determined by exhaustive enumeration [30, 31] and thus they are as accurate as possible. Of course in order to skip such an expensive procedure a statistical approach is sufficient. The joint neutrality parameter ρ is determined in this way. The values of the fixation and transition probabilities providing the input for the analytical curve evaluation are tabulated in appendix A.

Remark: The determined theoretical curves are based on the replication-deletion process and therefore expected to fit the numerical results of the replication process slightly better than those of the real flow reactor dynamic where fluctuating population size additionally comes into play.

A.

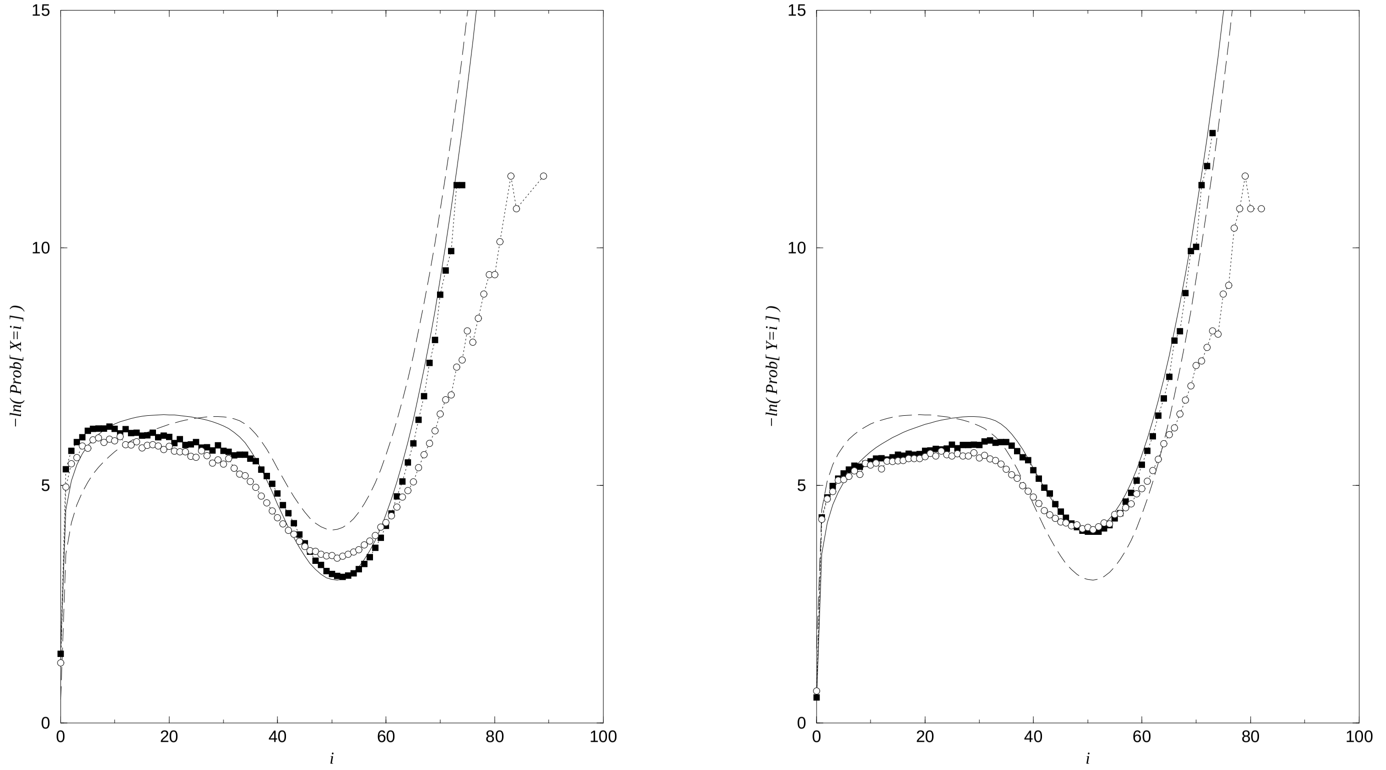


Figure 18: Stationary probability distribution. Two evolutionary experiments based on (i) the flow reactor dynamics (circles) and (ii) the replication-deletion process (squares) are performed. The underlying double shape landscape is induced by structure pair A and minimum free energy folding conditions. For an average population size of $N = 100$ the random variables X and Y count the number of masters on the neutral networks corresponding to s and s' respectively. We plot the negative logarithm of the stationary probability distribution for X and Y. The left part of the figure presents the results obtained for master sequences on the network corresponding to s whereas the right part shows the results for s'. The solid line always gives the analytical curve for the actually investigated network and for comparison the dashed line gives that of the non depicted one. (The underlying alphabet is (**G**,**C**). The simulations are started with 50 copies of one random sequence on network $\Gamma_n[s]$ and 50 copies of another sequence on $\Gamma_n[s']$. The averaging was done after 10^5 generations.)

B.

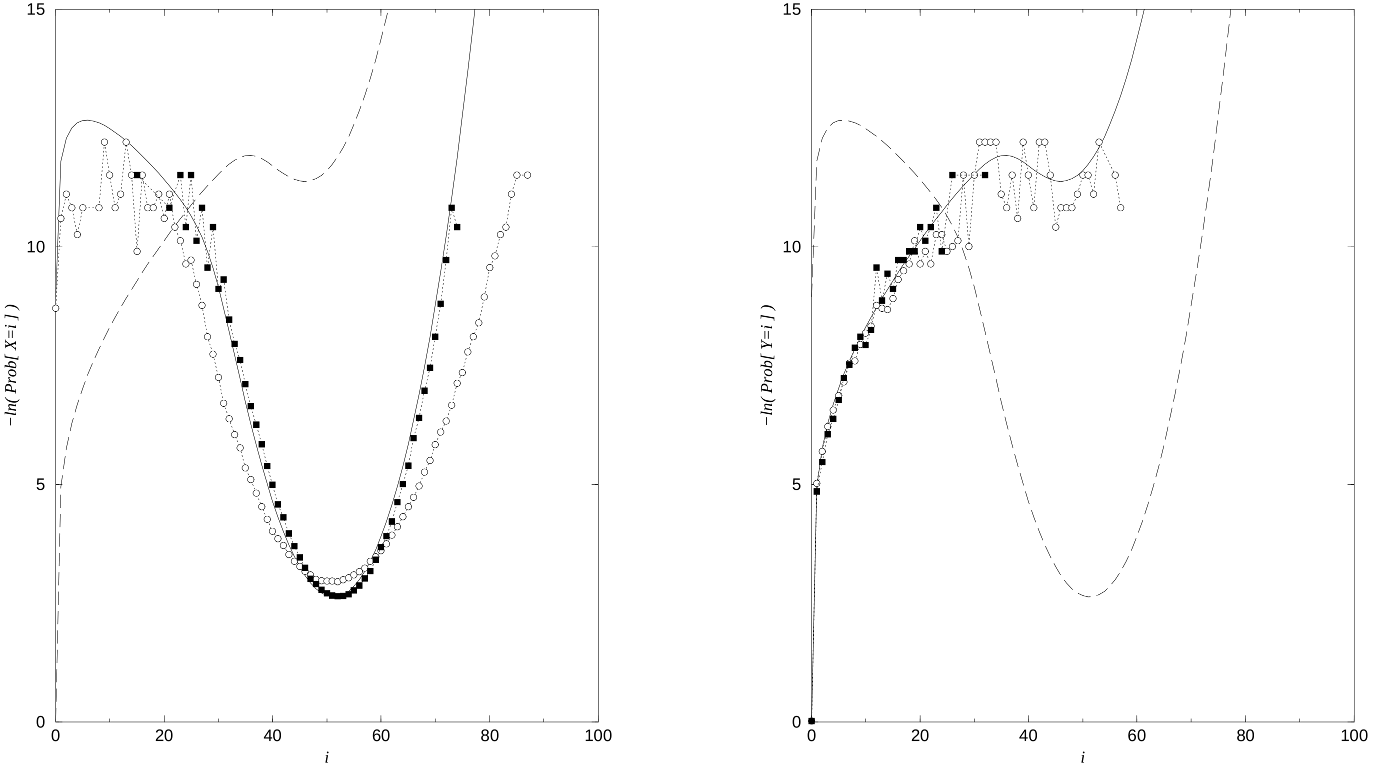


Figure 19: Stationary probability distribution. Two evolutionary experiments based on (i) the flow reactor dynamics (circles) and (ii) the replication-deletion process (squares) are performed. The underlying double shape landscape is induced by structure pair B and minimum free energy folding conditions. For an average population size of $N = 100$ the random variables X and Y count the number of masters on the neutral networks corresponding to s and s' respectively. We plot the negative logarithm of the stationary probability distribution for X and Y. The left part of the figure presents the results obtained for master sequences on the network corresponding to s whereas the right part shows the results for s'. The solid line always gives the analytical curve for the actually investigated network and for comparison the dashed line gives that of the non depicted one. (The underlying alphabet is (**G,C**). The simulations are started with 50 copies of one random sequence on network $\Gamma_n[s]$ and 50 copies of another sequence on $\Gamma_n[s']$. The averaging was done after 10^5 generations.)

C.

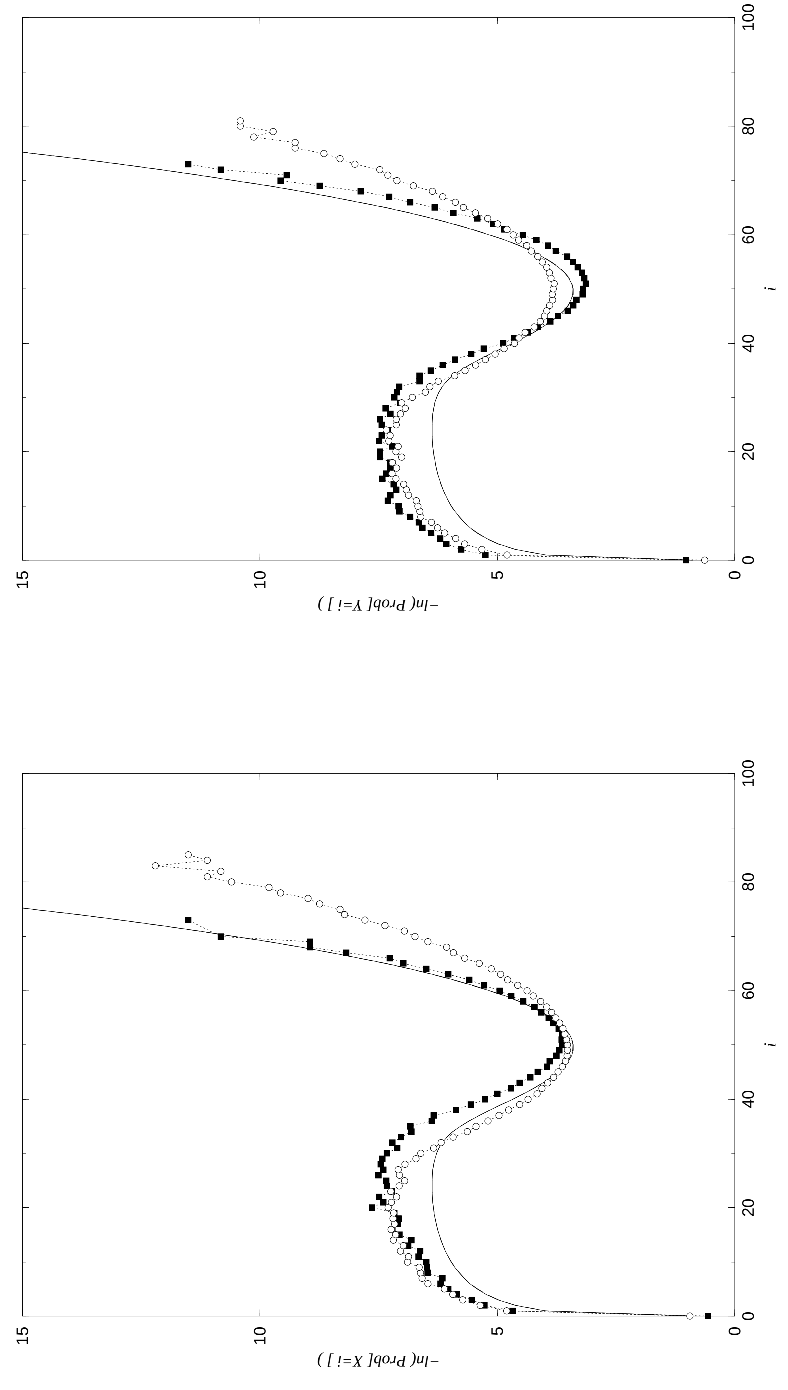


Figure 20: Stationary probability distribution. Two evolutionary experiments based on (i) the flow reactor dynamics (circles) and (ii) the replication-deletion process (squares) are performed. The underlying double shape landscape is induced by structure pair A and minimum free energy folding conditions. For an average population size of $N = 100$ the random variables X and Y count the number of masters on the neutral networks corresponding to s and s' respectively. We plot the negative logarithm of the stationary probability distribution for X and Y. The left part of the figure presents the results obtained for master sequences on the network corresponding to s whereas the right part shows the results for s'. The solid line always gives the analytical curve for the actually investigated network and for comparison the dashed line give that of the non depicted one. (The underlying alphabet is **(G,C)**. The simulations are started with 50 copies of one random sequence on network $\Gamma_n[s]$ and 50 copies of another sequence on $\Gamma_n[s']$. The averaging was done after 10^5 generations.)

5.2. Neutral Evolution Based On Random Graphs

Do randomly constructed neutral networks of RNA secondary structures behave in the same manner as those obtained by minimum free energy folding? In this section we shall verify this question for coevolution of two neutral networks.

In principle the application of the random graph models discussed in section 3.3 for computer simulations can be organized in two ways:

1. Two networks are constructed by making use of model I or II (section 3.1) and following the scheme (section 3.3) to deal with sequences on the overlap. Finally one need to store both networks that are random but fixed for the following simulation. The advantage of this construction is that the average fractions of neutral neighbors and the degree of neutrality can be exactly determined. A big disadvantage of this is the immense effort to store both networks on disk and finally to access to the data. Therefore we will use the second way.
2. The networks are constructed dynamically. The parameters λ_u and λ_p as well as the joint neutrality parameter ρ are defined. A new mutant entering the population is checked for compatibility with $\Gamma_n[s]$ and $\Gamma_n[s']$ respectively. A sequence which is compatible with structure s is chosen to be a member of $\Gamma_n[s]$ with probability $\lambda_u[s]\lambda_p[s]$ if it is not on the overlap and with $\rho{\cdot}(\lambda_u[s]\lambda_p[s])/(\lambda_u[s]\lambda_p[s]{+}\lambda_u[s']\lambda_p[s'])$ if it is found on the intersection. The construction for the second network runs analogically. The advantage of this procedure is that it is very fast. The disadvantage is that the construction is not unique. A sequence eventually lost and found for the second time might be first chosen to be on one network and the second time to be a non-master.

All simulations that have been performed on randomly constructed neutral networks are performed by using ansatz 2. We now proceed mentioning some technical details that the following simulations have in common.

- The underlying sequence space in which the structure pairs A,B,C will be explored is the hypercube of dimension 30 formed over the alphabet $\mathcal{A}$ =(**G, C**) with corresponding pairing alphabet $\mathcal{B}$ =(**GC**,**CG**).

- We consider an initial population consisting of $N = 100$ sequences divided into 50 copies of a random sequence on network $\Gamma_n[s]$ and 50 copies of a random sequence on network $\Gamma_n[s']$. Both random sequences are demanded to be non-intersection elements.
- The parameters λ_u and λ_p are set to 0.5 for s and s'. The joint neutrality ρ is chosen to be 0.25.
- The superior fitness σ of all master sequences is set to be 10 as opposed to 1 as minor fitness for non-master sequences.

Simulations

Two evolutionary experiments are executed based on (i) the flow reactor dynamics and (ii) the replication-deletion process. The underlying fitness landscapes are induced by the pairs of secondary structures A,B,C. The number of sequences occupying network $\Gamma_n[s]$ and $\Gamma_n[s']$ respectively is checked and stored after each generation. After having stopped a simulation at generation 10^5 the stationary probability distribution is evaluated by averaging over time. Figure 20, 21, and 22 present the outcomes for A,B, and C respectively. Each figure is subdivided into two part: the left sub-figure shows the results for network $\Gamma_n[s]$ whereas the right one monitors those for $\Gamma_n[s']$. Circles present the values we received from the flow reactor dynamics whereas squares correspond to those of the replication deletion process. The solid line pictures the analytical curve for the actually reported network whereas the dashed line corresponds to the network not reported. Since all parameters describing the topology of the networks had chosen to be equal the theoretical curves for the stationary probability distributions coincide for both networks in case of structure pair A and C. Hence for them the dashed line is covered by the solid line.

A.

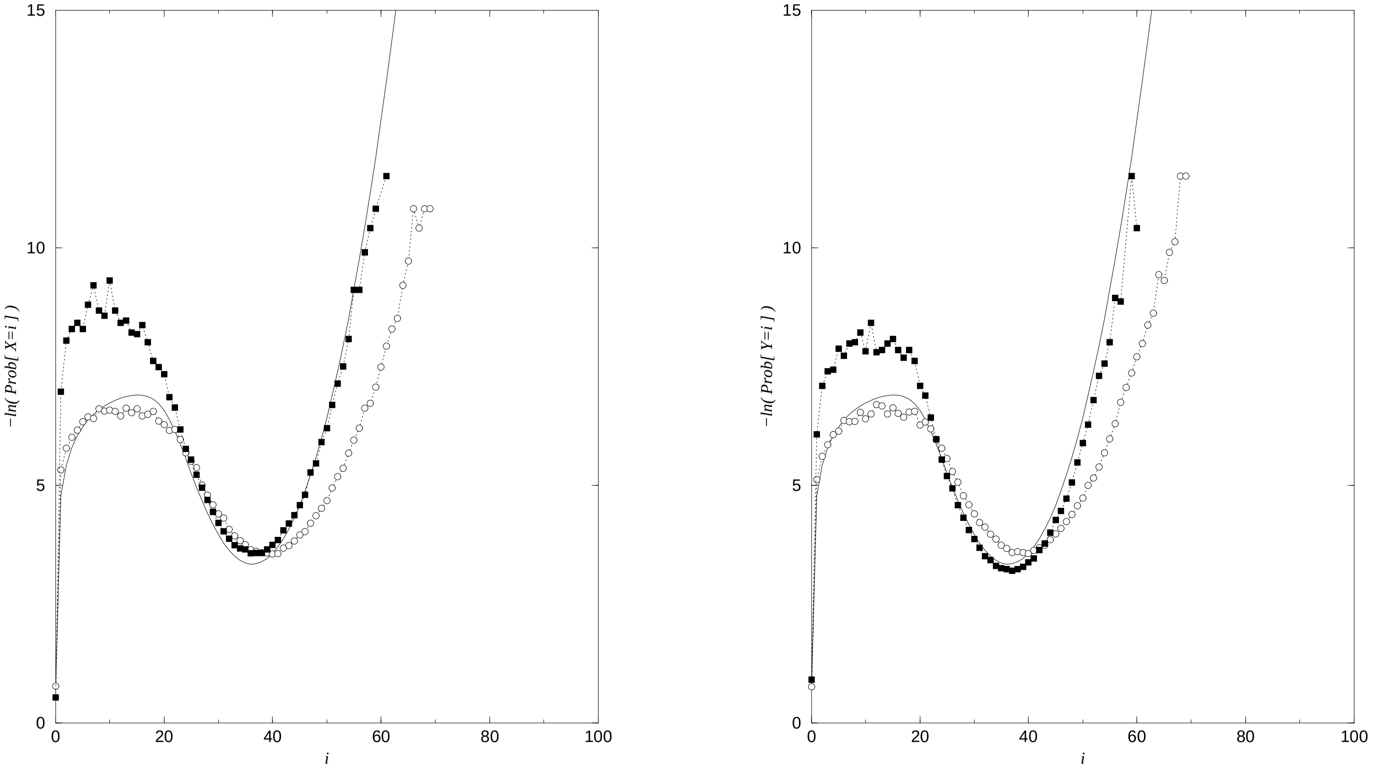


Figure 21: Stationary probability distribution. Two evolutionary experiments based on (i) the flow reactor dynamics (circles) and (ii) the replication-deletion process (squares) are performed. The underlying double shape landscape is induced by structure pair A and randomly constructed neutral networks. For an average population size of $N = 100$ the random variables X and Y count the fraction of masters on the neutral networks corresponding to s and s' respectively. We plot the negative logarithm of the stationary probability distribution for X and Y. The left part of the figure presents the results obtained for master sequences on the network corresponding to s whereas the right part shows the results for s'. The solid line always gives the analytical curve for the actually investigated network and for comparison the dashed line gives that of the non depicted one. (The underlying alphabet is **(G,C)**). The simulations are started with 50 copies of one random sequence on network $\Gamma_n[s]$ and 50 copies of another sequence on $\Gamma_n[s']$. The averaging was done after 10^5 generations.)

B.

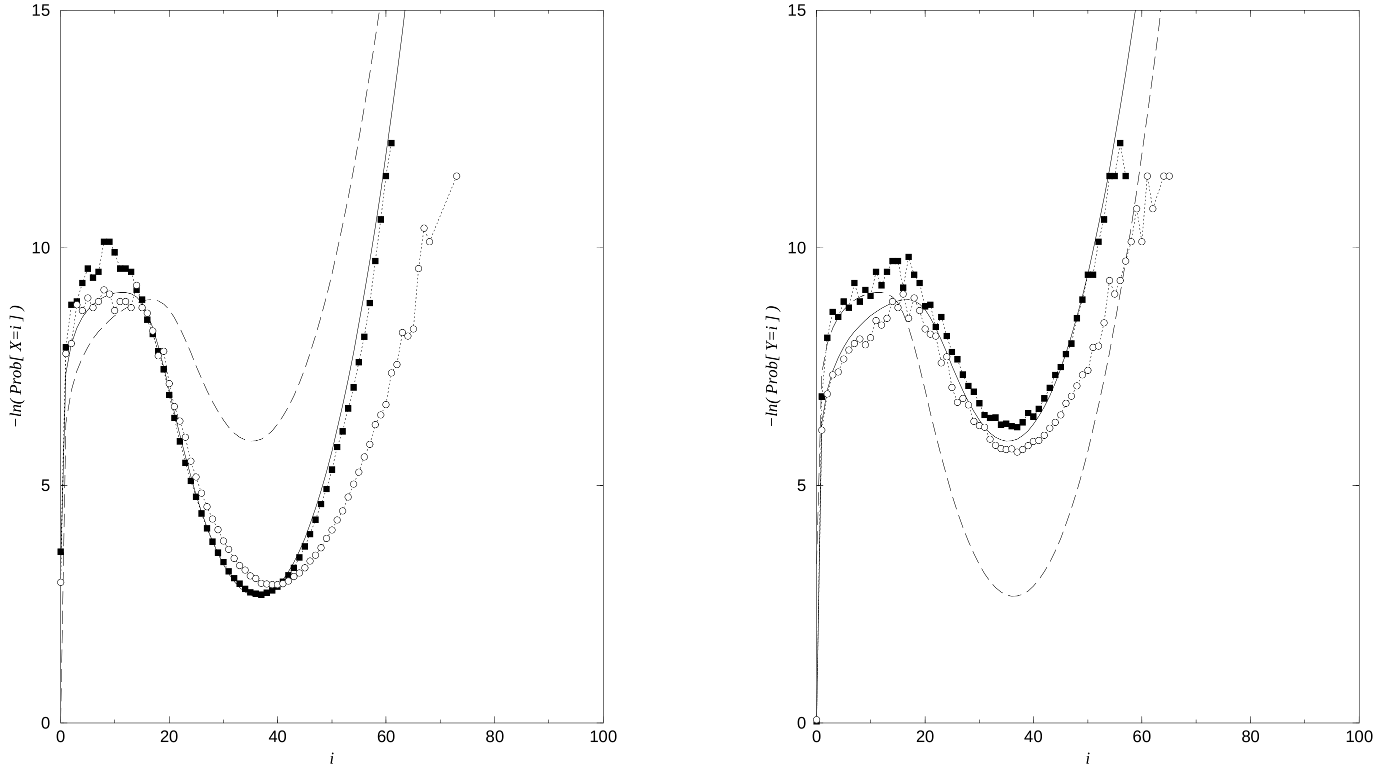


Figure 22: Stationary probability distribution. Two evolutionary experiments based on (i) the flow reactor dynamics (circles) and (ii) the replication-deletion process (squares) are performed. The underlying double shape landscape is induced by structure pair B and randomly constructed neutral networks. For an average population size of $N = 100$ the random variables X and Y count the fraction of masters on the neutral networks corresponding to s and s' respectively. We plot the negative logarithm of the stationary probability distribution for X and Y. The left part of the figure presents the results obtained for master sequences on the network corresponding to s whereas the right part shows the results for s'. The solid line always gives the analytical curve for the actually investigated network and for comparison the dashed line gives that of the non depicted one. (The underlying alphabet is (**G**,**C**). The simulations are started with 50 copies of one random sequence on network $\Gamma_n[s]$ and 50 copies of another sequence on $\Gamma_n[s']$. The averaging was done after 10^5 generations.)

C.

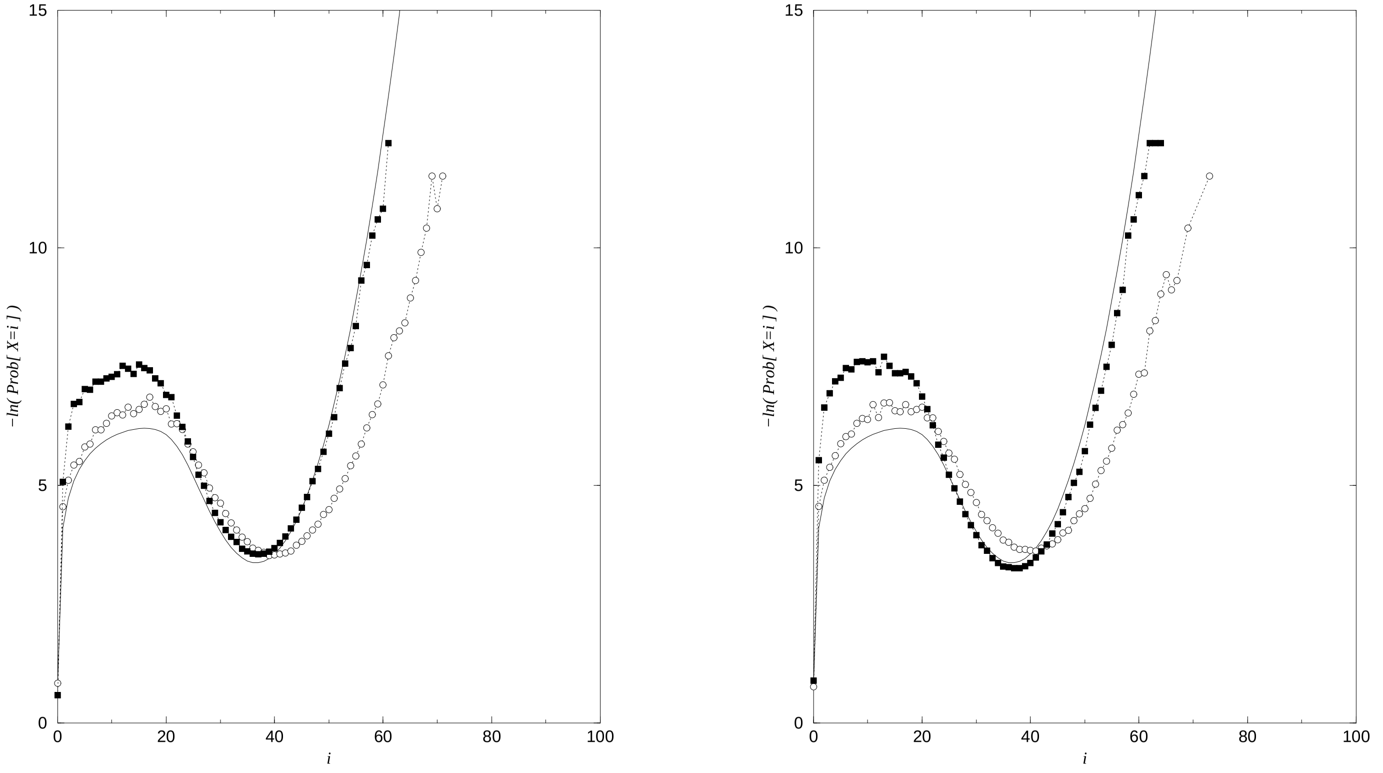


Figure 23: Stationary probability distribution. Two evolutionary experiments based on (i) the flow reactor dynamics (circles) and (ii) the replication-deletion process (squares) are performed. The underlying double shape landscape is induced by structure pair C and randomly constructed neutral networks. For an average population size of $N = 100$ the random variables X and Y count the fraction of masters on the neutral networks corresponding to s and s' respectively. We plot the negative logarithm of the stationary probability distribution for X and Y. The left part of the figure presents the results obtained for master sequences on the network corresponding to s whereas the right part shows the results for s'. The solid line always gives the analytical curve for the actually investigated network and for comparison the dashed line gives that of the non depicted one. (The underlying alphabet is (**G**,**C**). The simulations are started with 50 copies of one random sequence on network $\Gamma_n[s]$ and 50 copies of another sequence on $\Gamma_n[s']$. The averaging was done after 10^5 generations.)

5.3. RNA Folding versus Random Graphs

Again we deal with populations evolving in double shape landscapes. It was shown in the previous section that the stationary distributions of master sequences can be eminently approximated by modeling RNA folding with a random graph ansatz. Now we want to show that evolutionary dynamics exhibit the same properties in case of randomly chosen networks and networks resulting from RNA folding. Therefore we introduce some definitions that allow to classify states of populations evolving in a double shape landscape. We always look at a population $\mathbf{V}(t)$ at time t. Let X_t and Y_t be the variables counting the number of sequences in the population $\mathbf{V}(t)$ located on network $\Gamma_n[s]$ or $\Gamma_n[s']$ respectively. Then we define

(1) The population $\mathbf{V}(t)$ is *fixed* on $\Gamma_n[s]$ at time t iff $X_t > 0$ and $Y_t = 0$,
(2) $\Gamma_n[s]$ is the *dominant* network at time t iff $X_t > 0$ and $Y_t \leq \sigma_{XY}$, and
(3) $\Gamma_n[s]$ and $\Gamma_n[s']$ *coexist* iff $X_t, Y_t > \sigma_{XY}$

where σ_{XY} denotes the stationary standard derivation of the random variable $X_t + Y_t$, i.e. $\sigma_{XY}^2 = \sum_{z=0}^{\infty} p(z) \cdot z^2 - \left(\sum_{z=0}^{\infty} p(z) \cdot z\right)^2$ with $p(z) = \lim_{t \to \infty} Prob\{X_t + Y_t = z\}$. Note that fixation on a network is a special case of dominance.

There are three characteristic features for coevolution of two neutral networks: (i) one dominating network over time, (ii) two coexisting networks, and (iii) networks that mutually come close to dominance or extinction. The characteristic features depend on the size and topology of the networks and of the overlap, the distribution of sequences close to the overlap, the replication accuracy and the population size.

Type (iii) of coevolution is of particular interest because there is no prefered dominating network over time but fast changes between states characterizing the population. We call these switches *transitions*. They are correlated with the number of sequences on or close to the overlap that are present in the population $\mathbf{V}$.

Now we shall perform two evolutionary computer experiments having the following input parameters in common:

- The underlying double shape landscape is induced by structure pair A . The superior fitness of the master sequences is set to 10 as opposed to 1 as minor fitness.
- The underlying sequence space is the generalized hypercube over the alphabet (**G,C,A,U**) with corresponding pairing alphabet (**GC,CG,AU,UA,GU,UG**).
- The initial population size is set to $N = 1000$. We start with 500 copies of one random sequence on $\Gamma_{30}[s]$ and 500 copies of another random sequence on $\Gamma_{30}[s']$.
- The stochastic process is simulated by making use of the flow reactor dynamics.
- The experiments are executed for 10^6 generations.

The simulations differ in taking the neutral networks under consideration (1) from RNA folding under minimum free energy conditions and (2) from random graph construction with $\lambda_u = \lambda_p = 0.5$ for s and s' and $\rho = 0.25$.

Simulation 1

The neutral networks are taken from minimum free energy folding. In figure 24 we present the time development of relative frequencies of master sequences on the elected networks. We find network $\Gamma_{30}[s]$ to be dominant about 90% of time and thereof 93% to be fixed in population. The remaining 10% of time both networks coexist. Despite masters on $\Gamma_{30}[s']$ coming up in this particular realization the corresponding network is never observed to be dominant.

Simulation 2

According to section 5.3 the neutral networks corresponding to the elected pair of secondary structures are constructed dynamically randomly. In figure 25 we present the time development of the relative frequencies of master sequences. We find $\Gamma_{30}[s]$ to be dominant about 73% of time thereof 96% to be fixed in population. $\Gamma_{30}[s']$ is dominant only 12% of time but thereof 94% fixed in population. The remaining 15% are left for coexistence of the two networks.

Simulation 1

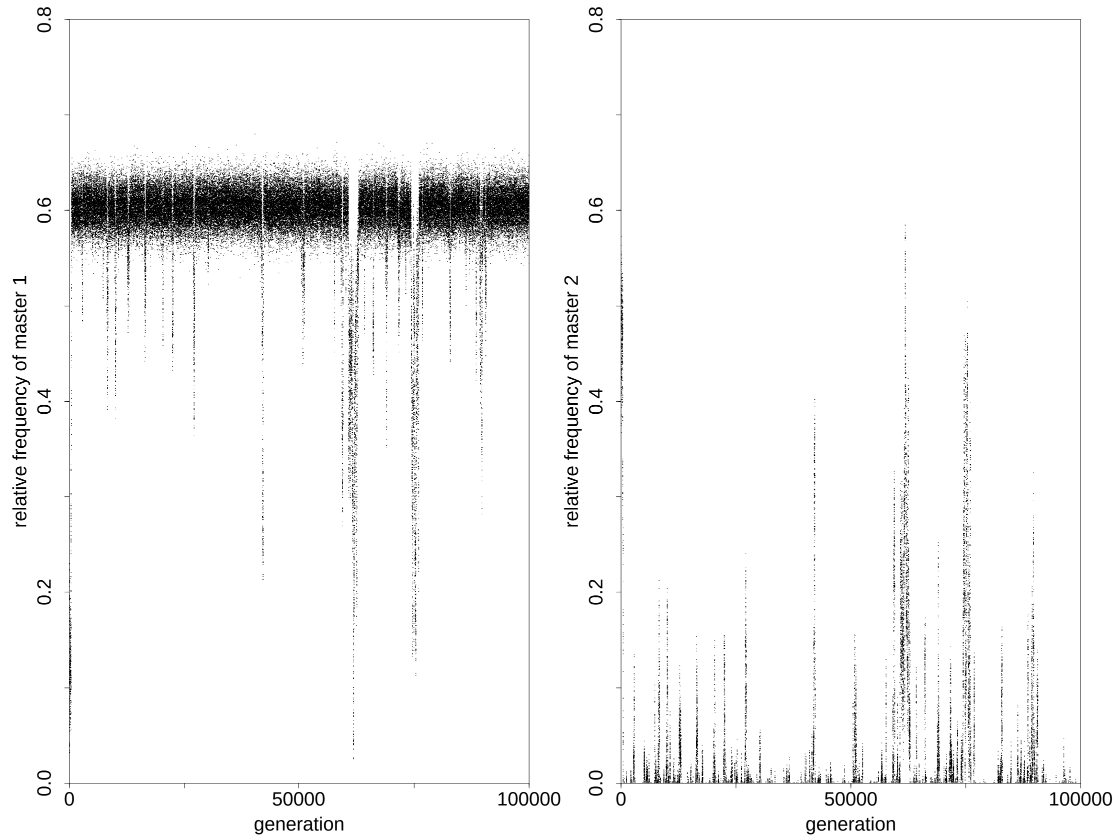


Figure 24: Simulation 1: The relative frequencies of masters on network $\Gamma_n[s]$ (left) and $\Gamma_n[s']$ per generation are presented. The simulation monitors the results for one realization of an evolutionary course on a double shape landscape induced by structure pair A. The underlying sequence space is the generalized hypercube over the physical alphabet (**AUGC**) with pairing alphabet (**GC,CG,AU,UA,GU,UG**). The neutral networks are taken from RNA folding. The average population size is $N = 1000$.

Simulation 2

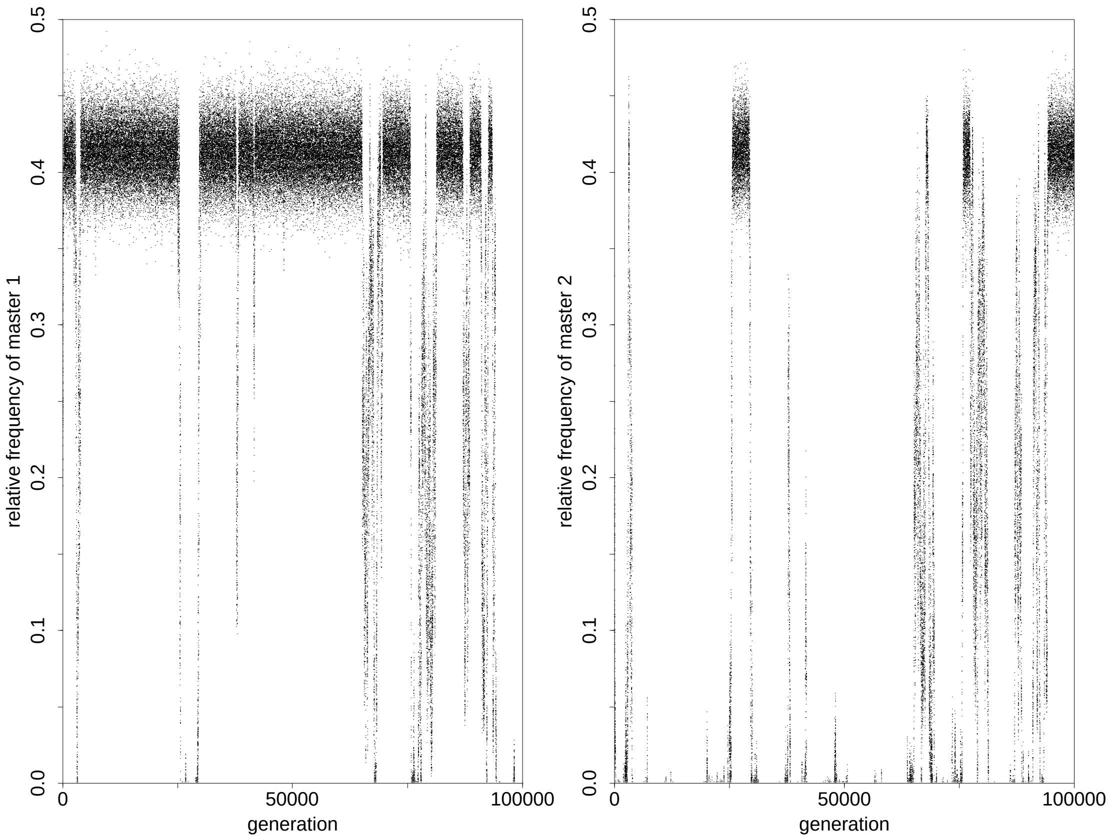


Figure 25: Simulation 2: The relative frequencies of masters on network $\Gamma_n[s]$ (left) and $\Gamma_n[s']$ per generation are presented. The simulation monitors the results for one realization of an evolutionary course on a double shape landscape induced by structure pair A. The underlying sequences space is the generalized hypercube over the the physical alphabet (**AUGC**). The neutral networks are randomly constructed. The average population size is $N = 1000$.

Although no transitions can be observed in simulation 1 we find that the dominance of network $\Gamma_n[s]$ vanishes for about 10 % of time. In figure 26 we show that these coexisting states are directly connected with the occurrence of sequences on the overlap or close to it. It was mentioned earlier (section 4.3) that a population in the context of double shape landscape can be divided into disjoint classes of sequences with respect to the number of incompatible base pairs with s or s'. A corresponding class was denoted by $\mathcal{E}_{ik}$, i.e. $\mathcal{E}_{00}$ is equivalent to the overlap. In the course of the evolutionary simulation sequences are found very seldomly in the overlap. The same effect applies to sequences in distance class $\mathcal{E}_{11}$ that however occur more frequently than those on the overlap. On the other hand we directly see that coexisting states are accompanied from the emergence of sequences in distance classes $\mathcal{E}_{10}$ and $\mathcal{E}_{01}$.

Simulation 2 is an example for an evolutionary trajectory of type (iii) . Note that by random graph construction both networks have the same properties (size, connectivity, density) in average. The alternations from one dominating network to the other occurs via coexisting states. Figure 27 shows that these states are accompanied from the emergence of sequences on the overlap and sequences close to it. It can be observed that sequences in classes $\mathcal{E}_{00}$ and $\mathcal{E}_{11}$ occur less frequently than those in classes $\mathcal{E}_{01}$ and $\mathcal{E}_{10}$ but their emergence is always correlated with coexisting states.

Simulation 1

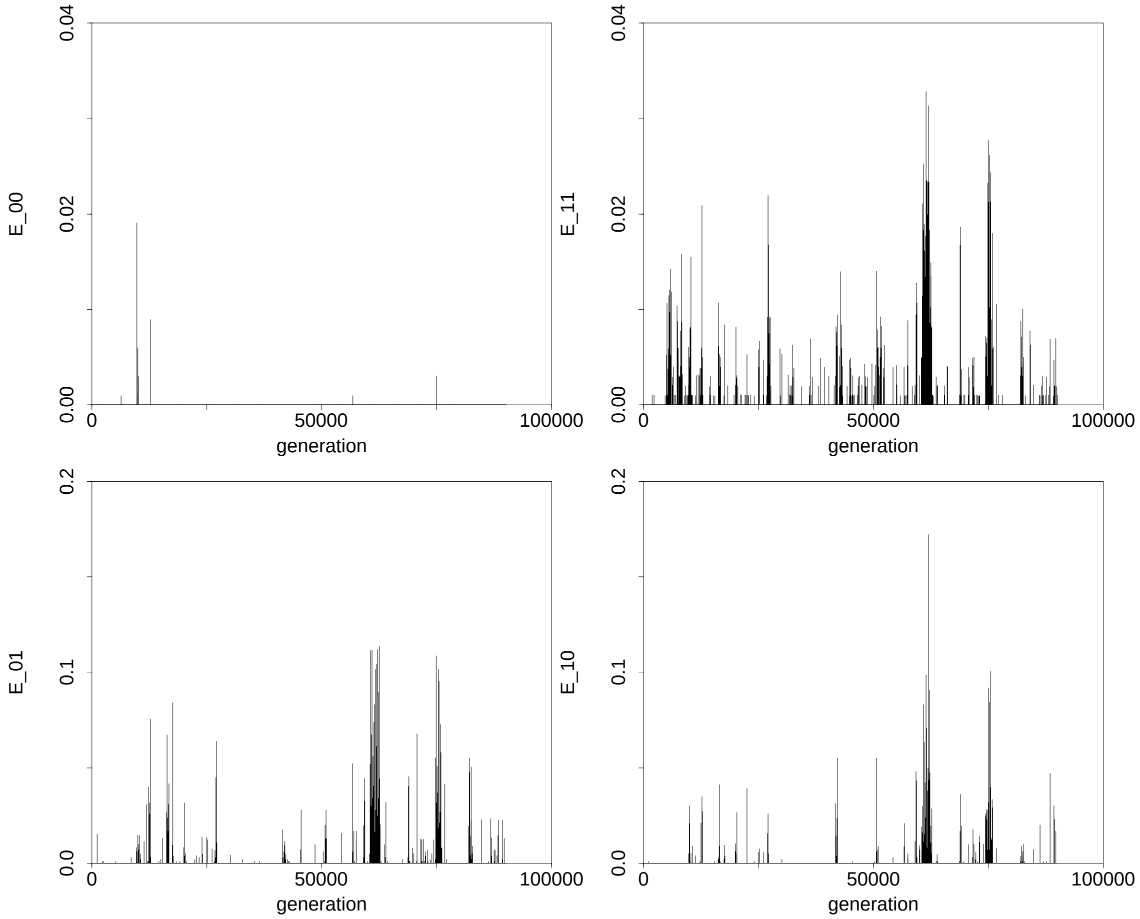


Figure 26: Emergence of sequences in different distance classes. Corresponding to *simulation 1* (figure 24) we document the relative frequencies of sequences on the overlap ($\mathcal{E}_{00}$) and of sequences close to it ($\mathcal{E}_{01}, \mathcal{E}_{10}$ and $\mathcal{E}_{11}$) in steps of 100 generations. Sequences on the overlap and in distance class $\mathcal{E}_{11}$ are found seldomly. However non master sequences having one incompatible base pair with each structure occur nearly all time. The emergence of sequences having only one incompatible base pair with either structure s or structure s' is perfectly correlated with coexisting states of both networks.

Simulation 2

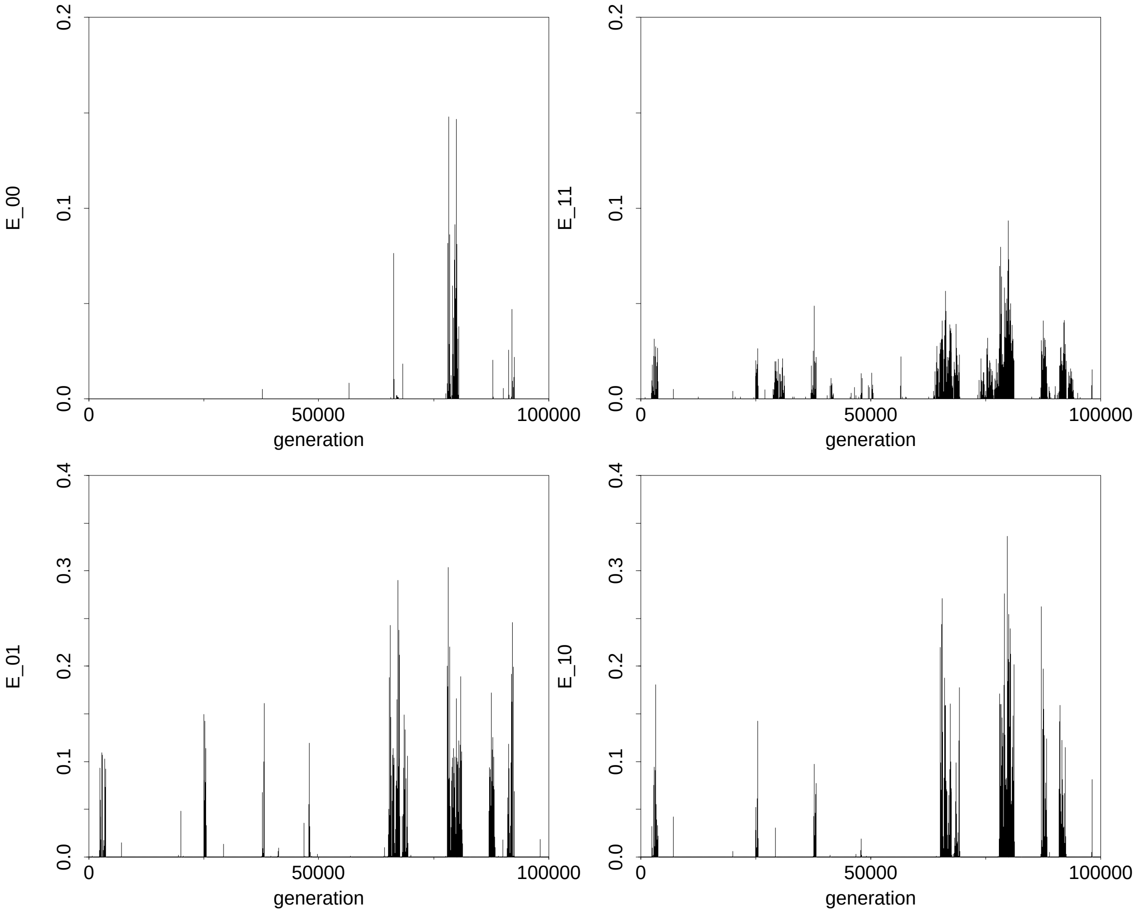


Figure 27: Emergence of sequences in different distance classes. Corresponding to *simulation 2* (figure 25) we document the relative frequencies of sequences on the overlap ($\mathcal{E}_{00}$) and of sequences close to it ($\mathcal{E}_{01}, \mathcal{E}_{10}$ and $\mathcal{E}_{11}$) in steps of 100 generations. In a nice way it is shown that transitions between neutral networks of different kind are supported by sequences on the overlap and nearby, i.e. in Hamming distance up to two. In this particular case sequences on the networks having exactly one incompatible base pair with the other structure are those facilitating transitions.

5.4. A Population in Sequence Space and Time

Except for a small group of theoreticians most people enjoy the graphical illustration of complex phenomena. So in evolutionary optimization it might be interesting to visualize the location of a population in sequence space at a fixed time t. Looking at a population in equidistant time steps may provide an impression how migration through sequence space takes place. The visualization of finite populations in sequence space is tantamount to the projection of a high dimensional space onto three or two dimensions the human being is living in. These projections are always associated with loss of information. However in a natural way a population evolving in a double shape landscape is proper for being pictured in the three dimensional space. According to the number of incompatible base pairs each pair of secondary structures induces a partitioning of sequences present in a population. These *distance classes* have been firstly introduced in section 4.3 .

Before proceeding our considerations we specify the conditions for another evolutionary experiment (simulation 3) whose results later on will be used as reference data.

- The underlying double shape landscape is induced by structure pair C . The superior fitness of the masters is 10 as opposed to 1 as minor fitness of the non-masters.
- The underlying sequence space is the binary hypercube over the alphabet (**G**,**C**) with pairing alphabet (**GC**,**CG**). The neutral networks are formed under minimum free energy conditions (RNA-folding).
- The initial population size is set to $N = 1000$. We start with 500 copies of one random sequence on $\Gamma_{30}[s]$ and another random sequence on $\Gamma_{30}[s']$.
- The stochastic process is simulated by making use of the flow reactor dynamics.
- The experiment is executed $9 \cdot 10^4$ generations.

In figure 28 we present the relative frequencies of masters on network $\Gamma_{30}[s]$ and $\Gamma_{30}[s']$ per generation. This single realization of the evolutionary experiment exhibits an example for the occurrence of sharp and fast transitions which alter the kind of dominating network. Network $\Gamma_{30}[s]$ is found to dominant about 47% of time and thereof it is 64% fixed in population. The masters of $\Gamma_{30}[s']$ are dominant about 42% of the time whereby they are fixed about 77% of this. The remaining 11% of time are left for the coexistence of both networks which indicates that coexistence is a transient state in the evolutionary process.

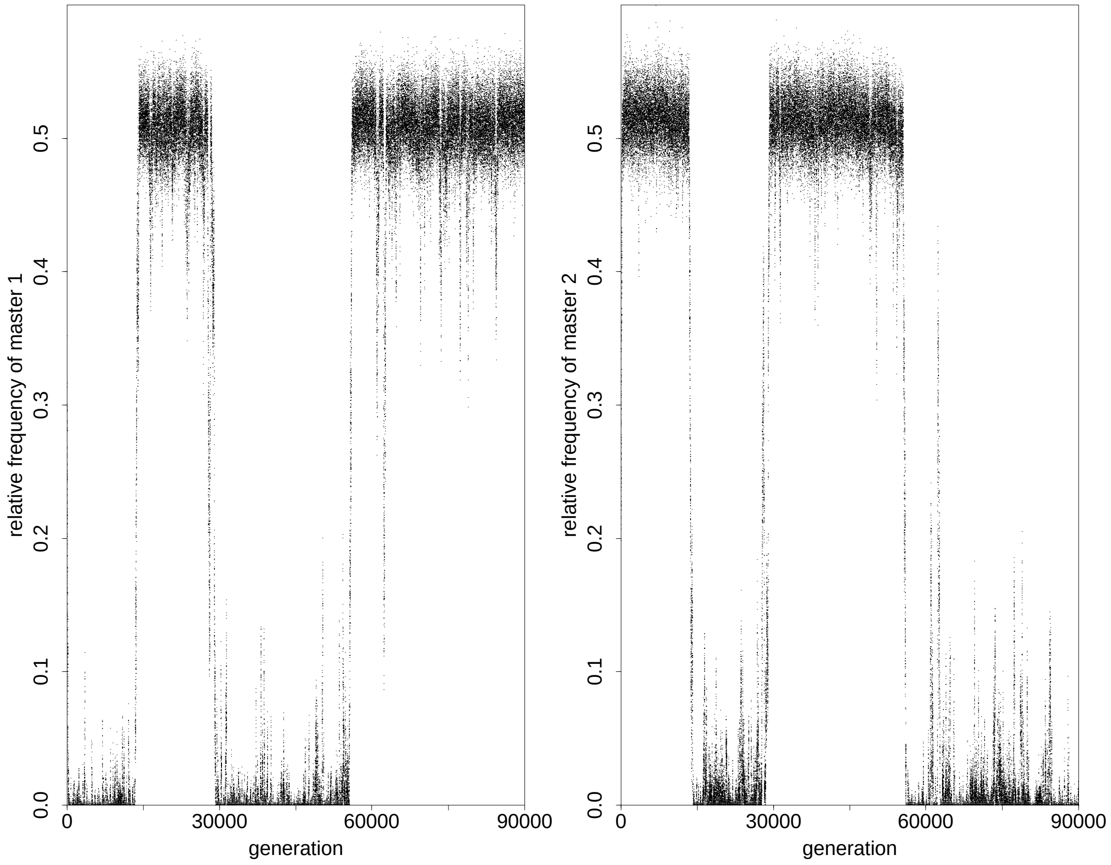


Figure 28: Simulation 3: The relative frequencies of masters on network $\Gamma_n[s]$ (left) and $\Gamma_n[s']$ (right) per generation are presented. The simulation shows the results of one realization of an evolutionary course on a double shape landscape induced by structure pair C . The underlying sequence space is the binary hypercube over the alphabet **(G,C)**. The neutral networks are taken from minimum free energy folding. The average population size is $N = 1000$.

One transition being observed in figure 28 covers the time range from generation 13000 to 14500. For this time interval a series of pictures is presented in figure 29. In steps of 100 generations they report snapshots of the population that is represented according to the partition with respect to the distance classes induced by the structure pair (s, s'), i.e. A reference point (i, k, e_{ik}) in the three dimensional space represents the relative frequency of sequences in the population that are located in distance class $\mathcal{E}_{ik}$. Noticing that structure s and s' belonging to the pair under consideration both have eight and no common base pairs a quadratic grid of length eight is formed.

Up to generation 13100 the population is biased on network $\Gamma_{30}[s']$. A mutant spectrum surrounds the center of the population that is located in distance classes $\mathcal{E}_{4,0}$ to $\mathcal{E}_{5,0}$. This observation corresponds to the statement of lemma 4. In this times the overlap is not occupied. Due to random drift generation 13200 up to 13400 exhibit a broadening of the distribution. The population moves close to the overlap. Then some sequences migrate to the other network as if they tunnel through distance class $\mathcal{E}_{0,0}$. Finally the population is split into two subpopulations each of them occupying one network and being centered around distance classes $\mathcal{E}_{4,0} - \mathcal{E}_{5,0}$ and $\mathcal{E}_{0,4} - \mathcal{E}_{0,5}$ respectively. From generation 13500 to 14000 the subpopulations seem to struggle to overcome this unstable state of coexistence. And finally in generation 14100 network $\Gamma_{30}[s]$ has managed to pull all individuals to its side. Now the whole population becomes centered around distance class $\mathcal{E}_{0,4}$ to $\mathcal{E}_{0,5}$.

This view on the support of a population during time evolution corresponds to the observations made by Fontana *et al.* [23]. The two neutral networks form isolated fitness platforms in sequence space where populations can be trapped on. Evolutionary jumps correspond to transitions between neutral network. In extent to Fontana we are able to specify the type of sequences providing the opportunity to escape from one network to the other. Thus the waiting for a series of any kind of fluctuations is more precisely waiting for sequences on the overlap or 'close' to it.

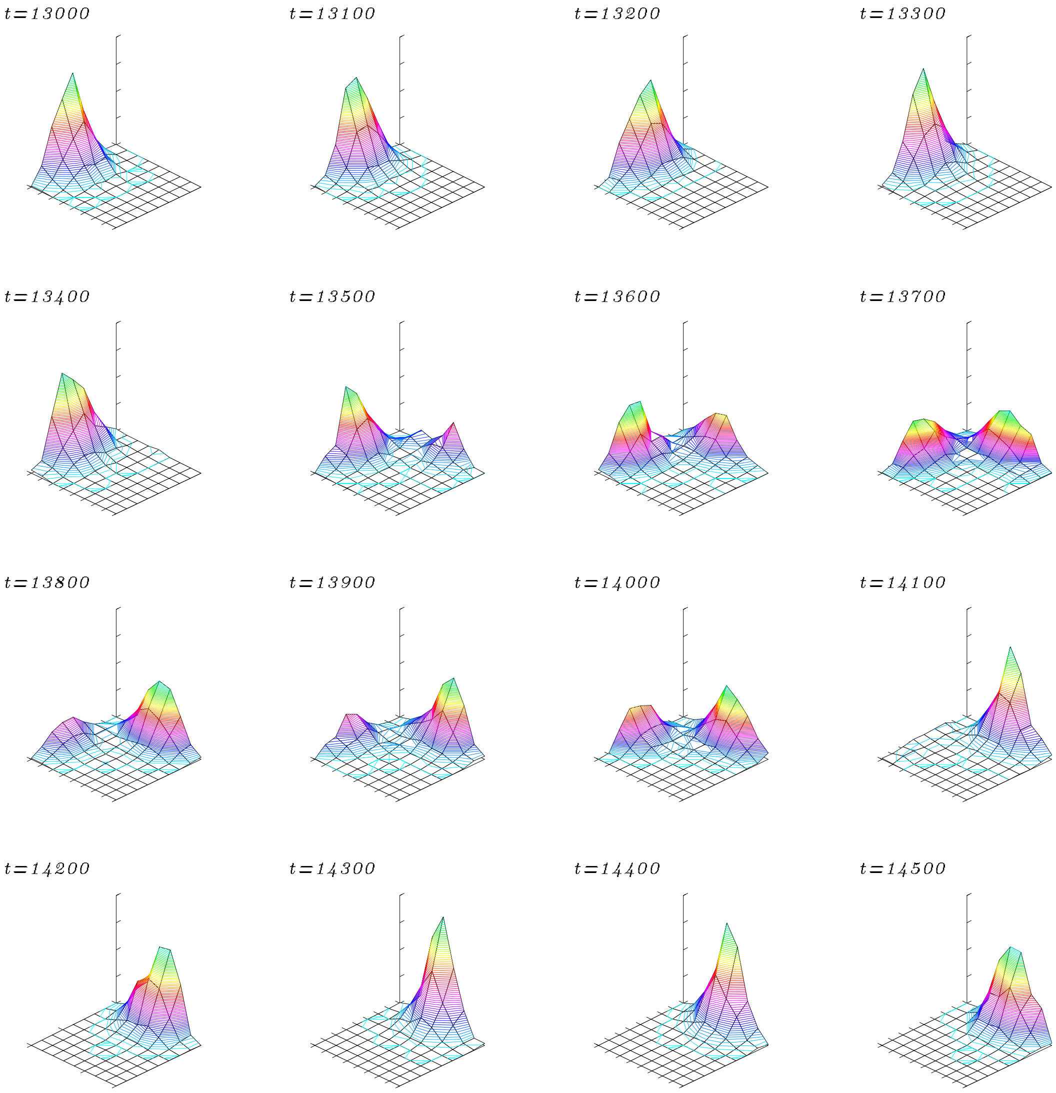
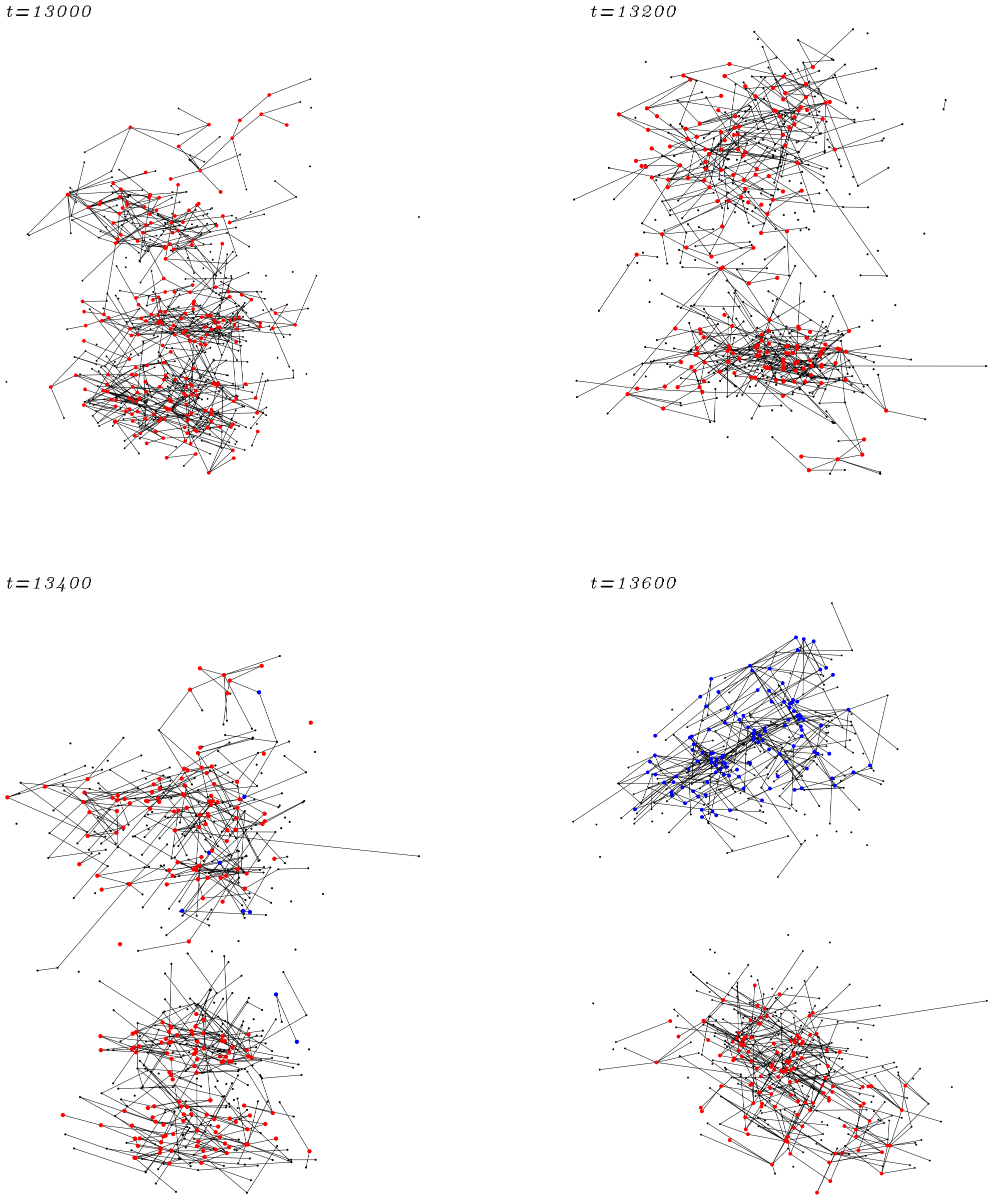


Figure 29: Population in a double shape landscape. In a natural way two structures induce a partitioning of a population, i.e. in disjoint classes of sequences that can be visualized onto three dimensions. Let (i, k, e_{ik}) be an arbitrary coordinate, then i shall indicate the number of incompatible base pairs that a sequence has in common with structure s whereas k complies to the number of incompatible base pairs that this sequence shares with structure s'. Finally e_{ik} denotes the relative frequeny of all sequences present in a population that belong to distance class $\mathcal{E}_{ik}$. According to the simulation results presented in figure 28 we monitor the partitioning of the population starting in generation 13000 up to 14500 in steps of 100 generations.

Derrida and Peliti [10] have predicted by means of analytical calculations that in absence of natural selection due to neutrality a population splits into well-defined clusters in sequence space. This was supported by Huynen *et al.* [40] who derived a diffusion coefficient for the center of mass of a population in sequence space. Moreover by computer simulations he has shown that a population undergoing erroneous replication splits into subpopulations that share the same phenotype and diffuse independently through sequence space. A diffusion coefficient for a population evolving in a single shape landscape was derived by Reidys *et al.* [57]. Accordingly a population on a neutral network behaves like a drop of a fluid. That in fact this behavior applies to populations on double shape landscapes as well could already be guessed from figure 29.

Another kind of visualization that more sophisticated gives insights into the process of clustering is presented in figure 30. Using some theorems from distance geometry [34] the support of a population can be pictured in two dimensions, where each sequence is presented by a point. Again we use the results delivered from the evolutionary course presented in figure 28. We show snapshots from generation 13000 to 13600 in steps of 200. Blue points indicate sequences on network $\Gamma_{30}[s]$ whereas red points those of $\Gamma_{30}[s']$. The minimum spanning tree with respect to Hamming distance is drawn omitting edges of length greater than two. We observe at generation 13000 the population covering network $\Gamma_{30}[s']$. Except of a few non-masters all sequences are connected and thus the population forms a cloud in sequence space. Yet in generation 13200 the population starts splitting into two subpopulations but however it is still connected. Finally in generation 13400 the splitting is completed and two subpopulations are found on $\Gamma_{30}[s']$. One of them is penetrated by sequences that belong to the other network. So this subpopulation completely migrates to network $\Gamma_{30}[s]$. Finally in generation 13600 the population is divided into two clusters each of them occupying one of the master networks. Replication, mutation and selection make this state of coexistence extremely unstable and so after a couple of generations the population will become fixed on either one or the other network.



t=13000 *t=13200* *t=13400* *t=13600*

Figure 30: Population structures in sequence space. All sequences present in at least one copy define the support of a population. Theorems from distance geometry [34] can be applied in order to picture it in two dimensions. A metric Matrix M is computed with entities $m_{ij} = (d_{0i}^2 + d_{0j}^2 - d_{ij}^2)/2$, where d_{ij} is the Hamming distance between sequence i and j and 0 indicates the alternating sequence GCGC.... Sequences are expressed in principal axes coordinates by diagonalizing M. Only the components corresponding to the two largest eigenvalues are kept, yielding a projection onto the plane. According to the simulation presented in figure 28 we exhibit snapshots of the population support in generation 13000 up to 13600 in steps of 200 generations. Blue dots indicate sequences folding into structure s, red dots sequences folding into s' and tiny black dots sequences that neither fold into s nor s'. The dots are connected by edges of the minimum spanning tree. Edges connecting sequences of Hamming distance greater than 3 are omitted.

6. Discussion

The late sixties saw the birth of a new concept in population genetics when Kimura postulated his neutral theory of molecular evolution [43, 44]. The controversy arising over the hypothesis whether or not neutral variants exist in nature and play a role in evolution had an impact on the view of molecular evolution. Due to an immense amount of experimental data from sequence comparison there are no doubts left about the existence of neutrality at the molecular level. Yet one controversial issue has not be clarified at all: What is the relative importance of random drift and positive selection. Even Charles Darwin had admitted the existence and relevance of random drift. Nowadays it has been recognized that its effects cannot be neglected when considering evolutionary dynamics on the molecular level [9, 41, 48]. Kimura emphasized that his theory has two roots. One being the stochastic theory of population genetics and the other being molecular genetics. In our opinion the most important aspect of this unification is the qualitative formulation of a mechanism of heredity. The main goal is to answer the question: How do changes in genotype act on the level of phenotypes? The RNA model considers explicitly the folding of sequences into secondary structures. It makes for the first time investigations of the influence of mutations on the phenotype accessible to mathematical analysis and computer simulations.

Local neutrality being the property of a single sequence to have sequences in its neighborhood that adopt the same structure and the definition of compatible sequences provide the inputs to the description of 'neutral networks' as randomly induced graphs [59, 56]. The direct comparison of features predicted by the random graph model with RNA folding-networks had demonstrated its usefulness [30, 31]. Critical values for the average fraction of neutral neighbors had been derived above which neutral networks are predicted to be dense and connected, with other words particularly well suited for searching in evolution [56, 59]. Those networks are proven to approach each other in sequence space, i.e., for any two networks there exists a set of sequences representing points of contacts between them (sect. 3.3). For the physically meaningful **AUGC** alphabet the cardinality of this set can be obtained as the product of generalized Fibonacci numbers. With the same arguments leading to the size of the overlap it can be shown that it is uniformly distributed in sequence space and has the form of connected islands. These contact sequences are claimed and proven to be of primary importance for evolutionary adaptation (sect. 3.3, 5.3). The random graph model [56] (sect. 3.2) was developed in order to describe and analyze RNA folding landscapes. The existence of sequences that in principle could adopt two structures requires an extension of this model. We have proposed to treat sequences on the overlap separately by introducing a joint neutrality parameter which decides whether the sequence is a master or not. The kind of network the sequence belongs to is taken into account by weighting over the network parameters. The addition of only one degree of freedom to the parameters of the networks turns out to be a useful approximation to RNA-folding (sect. 3.3).

The generation of simple but nevertheless suggestive landscapes is one approach to the question: What are the generic features of evolutionary optimization? That means in order to investigate neutral evolution on RNA folding landscapes it is natural to furnish sequences adopting the same secondary structure with the same fitness value. This approach corresponds to the assumption of a single sequence having superior fitness that leads to the quasi-species [14], or of two sequences having the same superior fitness values yielding a degenerated quasi-species [66]. But instead of asking for the stationary distribution of particular sequences we are interested in the number of sequences being located on specified networks encoding a set of interacting phenotypes. In order to take the effects of genetic drift into account the population size was assumed to be very small compared to the size of the sequence space. This finally demanded a stochastic approach to model the kinetics.

First a basic model was established that assumed a completely flat fitness landscape accommodating two fictitious phenotypes (sect. 4.1). A probability 'w' of a replication to produce an offspring that belongs to the same phenotype as its template was defined. This parameter – the *phenotypic fixation probability* – was shown to have the function of a decoupling constant between the two phenotypes. Depending on the population size the existence of a critical value w^* could be shown. Below this value both phenotypes are likely to coexist over time whereas above it only one phenotype is actually present. The course of evolution is then characterized by fast transitions changing the nature of dominating phenotype. These swaps can be observed for constant as well as for fluctuating population sizes. Distinct fixation probabilities w for each of the two phenotypes change the properties of the model only slightly. Coexistence can be observed if both values are below w^*. If one w-value is smaller and the other one greater than w^*, the phenotype with the greater w-value dominates over time. Transitions between dominating phenotypes can be only expected if both probabilities are above the critical phenotypic fixation probability w^*.

The counter-part of the single-peak landscape [16] is the single-shape landscape on the level of RNA secondary structures [56] (sect. 4.2). It is induced by a fixed neutral network whose sequences are assumed to form a fitter master phenotype than all other sequences (non-masters). This in a natural way defines a partition into two disjoint ‘phenotypes’. A uniform error model [17] and the random graph approach (sect 3.2) provide the inputs for deriving arithmetic expressions for the phenotypic fixation probability ($W_{\mu\mu}(p)$ corresponding to sequences on the network [59] and $W_{\nu\nu}(p)$ for those not being on the network) which depend on the single digit error rate p. In general, the probability to obtain a master individual through replication when the template is a master strongly depends on the master-template itself. The same applies to non-masters as well. So the values for $W_{\mu\mu}$ and $W_{\nu\nu}$ have to be interpreted as average probabilities. They give a good approximation if λ_u and λ_p – the average fractions of neutral neighbors in the unpaired and paired regions of a structure – are interpreted in terms of conditional probability and additionally an average rate of neutrality is introduced. The latter is defined to be equal to $|\Gamma_n[s]|/|\mathbf{C}[s]|$ which in general is much smaller than $\lambda_u \lambda_p$. This assumption takes the local effect called buffering [39] into account.

Likewise as for the basic model, there exist critical values for the probabilities $W_{\mu\mu}$ and $W_{\nu\nu}$ additionally depending on the absolute fitness difference that determine the character of the evolutionary course. For finite (sect. 4.2) as well as for infinite population sizes [57] an error threshold p^* can be given where in the long time limes it becomes most likely to find no sequence on the neutral network, i.e., all sequences that carry the information of this particular phenotype vanish.

The model envisages an increase in complexity and one further step towards an understanding of neutral evolution when the interaction of three phenotypes is investigated. In particular, we are interested in the evolutionary behavior of a population that is characterized by two dominating neutral networks of equal fitness (sect. 4.3). It is known from section 3.3 that for any two fixed neutral networks a number of sequences can be found that come very close to both of them. This gives rise to the assumption that in the course of evolutionary optimization a population mainly living on one network can not only pick up sequences that belong to the other network but eventually support it by producing offspring that adopt the structure or the second network. The parameters of the extended random graph model and the single digit error rate provide the input for the formulation of analytical expressions describing the phenotypic fixation and transition probabilities. The evolutionary dynamics of a finite population in an underlying double-shape landscape turns out to be a combination of the basic model acting on the level of the networks and the dynamics on a single-shape landscape. With respect to the neutral networks three generic features are possible to emerge in the course of evolution

(1) coexistence of genotypes from both networks,

(2) genotypes supporting only one network,

(3) sporadic alternation of extinction and dominance of genotypes belonging to the networks.

These outcomes depend primarily on the topologies of the two networks which are determined by the amount of compatible sequences, the average fractions of neutral neighbors in the unpaired and paired regions, and the average fraction of neutrality and secondly on their overlap as well as on the structural features, single digit error rate, and the population size. In applying our knowledge from the basic model and from the single-shape dynamics it can be recognized that these features are strongly correlated with the fixation rates $W_{\mu_1\mu_1}$ and $W_{\mu_2\mu_2}$ on the two networks. While coexistence of both and extinction of only one network are two outcomes that could be figured out as well by the deterministic approach to the model, mutual alternation of extinction and dominance is caused by the finiteness of the population. The fact that this third feature of evolutionary behavior cannot be detected by the deterministic approach to the model is one more justification for the stochastic view on evolutionary dynamics. A finite population migrates through sequence space and moves on a neutral network by means of diffusion like mechanisms [40, 56]. Depending on the population size it may split into subpopulations that then move autonomously through sequence space. A population being fixed on a predefined network and producing offspring that again are preferentially located on the same network after some time will come close to the other fit network. Then it may 'tunnel' to the other network via sequences on the overlap and become fixed there, or it passes transient states of coexistence and again becomes fixed on the previously dominating network (sect. 5.4). These elaborated interesting stochastic phenomena give rise to new aspects in neutral evolution. Therefore under certain circumstances firstly the replacement of distinct phenotypes of nearly the same fitness takes place only in a few generations (sect. 4.1) and coexistence is only seen to be an intermediate unstable state. This corresponds to the results presented by Fontana *et al.* [23]. But secondly and this is the most astonishing result a phenotype going extinct does not need to be irreversibly lost (sect. 5.4), if it has a similar fitness as the actually dominating phenotype and covers an extended network in sequence space. Then error prone replication causes a diffusion process that allows the population to move close to sequences encoding the lost phenotype and eventually producing offspring that belongs to it.

Appendix A: Parameters

In order to connect the random graph theory on neutral networks of RNA secondary structures with the RNA folding algorithm the average fractions of neutral neighbors (λ_u and λ_p) need to be estimated. In general this can be done by choosing a random sample of sequences on a predefined neutral network and computing the distribution of neutral neighbors separately for unpaired and paired bases. The average fraction of neutrality $\bar{\lambda}$ is derived from a random sample of sequences that are compatible with the given structure. Finally the joint neutrality parameter ρ is computed from a random sample of sequences on the intersection $\mathbf{I}[s, s']$. Due to the data received by exhaustive folding of all **GC** sequences up to chain lengths 30 it was possible to give the exact values for all neutral networks [30, 31]. For convenience once more table 1 recalls the pairs of structures (s, s') underlying the numerical simulations performed in the previous section.

Table 1. pairs of RNA secondary structures forming double shape landscapes

	s	s'
A.	`........(((((((((....)))))))))`	`.........(((((((((...)))))))))`
B.	`........(((((((((....)))))))))`	`((((((((...))))))))((((...))))`
C.	`.......((((((((.......))))))))`	`((((((((.......)))))))).......`

For these three pairs of secondary structures and the sequence space of dimension 30 formed over the alphabet (**G**,**C**) table 2 monitors the parameters describing the corresponding neutral networks.

According to section 4.3 the phenotypic fixation and transition probabilities are calculated and presented in table 3. Here the single digit error rate was chosen to be 0.03. For each pair of structures all sequences adopting structure s are denoted by μ_1 whereas those adopting structure s' are denoted by μ_2. Non-master sequences are indicated by ν.

Table 2. average fractions of neutral neighbors, average fraction of neutrality, joint neutrality with respect to s and s'

	$A.$	$B.$	$C.$
λ_u	0.859944	0.859944	0.666106
λ'_u	0.802359	0.967120	0.665647
λ_p	0.894559	0.894559	0.748079
λ'_p	0.928420	0.965835	0.745359
$\bar{\lambda}$	0.747912	0.747912	0.295754
$\bar{\lambda}'$	0.642873	0.957394	0.293419
ρ	0.3911	0.9160	0.1491

Table 3. phenotypic fixation and transition probabilities for single digit accuracy 0.97

	$A.$	$B.$	$C.$
$W_{\mu_1\mu_1}$	0.557150	0.557358	0.545325
$W_{\mu_2\mu_2}$	0.546931	0.484030	0.545215
$W_{\mu_1\mu_2}$	0.000458	0.000070	0.000444
$W_{\mu_2\mu_1}$	0.000533	0.000534	0.000448
$W_{\nu\mu_1}$	0.000535	0.000535	0.000449
$W_{\nu\mu_2}$	0.000513	0.000082	0.000483

Appendix B: Stochastic Simulation of Finite Populations Dynamics

"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. This deterministic formulation of chemical kinetics leads to N differential equations if there are N chemically active molecular species. An arbitrary equation expresses the time-rate-of-change of the molecular concentration of a single species in functional terms of the concentration of all species as well as reaction constants and stoichiometrics of the reactions it is involved in (mass-action-kinetics, Michealis-Menten-kinetics).

Another approach to chemical kinetics of a spatially homogeneous system is the stochastic formulation. It is somewhat more applicable than the deterministic formulation and as opposed to the mathematically more simple deterministic approach, it takes *fluctuation and correlation* into account. In the stochastic formulation reaction rates are not viewed as deterministic reaction rates but as reaction probabilities per time unit. The temporal behavior of a chemically reacting system takes the form of a Markovian stochastic process 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 probability function over the state space of N species depending on time. From the mathematical point of view the set of deterministic reaction rate equations for a given system is much easier to solve than its corresponding master equation. If neither formulation is tractable by analytical methods computer-oriented methods are required. Gillespie proposed such a method to simulate the Markov process that the master equation describes analytically. In particular he addressed a problem that can be formulated as follows:

- There is given a volume V containing molecules of N chemical active species S_i $(i = 1, \ldots, N)$ and possibly molecules of several inert species.
- Let X_i be the current number of molecules of the species S_i in V with $(i = 1, \ldots, N)$.
- The N species S_i can participate in M chemical reactions R_μ $(\mu = 1 \ldots M)$, each characterized by a numerical *reaction parameter* c_μ.

Remark: A population of haploid replicating molecules evolving in a flow-reactor can be described by two types of reactions $\{R_{\mu_1}\} : S_i \to S_i + S_j$ for replication and $\{R_{\mu_2}\} : S_i \to *$ for the unspecific dilution flux.

The *fundamental hypothesis* of the stochastic formulation of chemical kinetics states that the reaction parameter c_μ can be interpreted as
$c_\mu \delta t$ $\equiv$ average probability, to first order in δt, that a particular combination of R_μ reactant molecules will react accordingly in the next time interval δt.

The principal 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_i molecules of Species S_i for $i = 1, \ldots, N$ 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. In the limit of infinitely many runs the values $X_i(t)$ approach an average value, also the variance of the values $X_i(t)$ is finite.

The usual stochastic approach to the coupled chemical reaction problem focuses upon the probability function $\mathcal{P}(X_1, X_2, \ldots, X_N; t)$. Then the master equation is the time evolution equation for the $\mathcal{P}(X_1, X_2, \ldots, X_N; t)$. Often it turns out to not feasible to solve the master equation both analytically and numerically. That is why the numerical method proposed by Gillespie is not based on $\mathcal{P}$ but one another quantity called the *reaction probability density function*, $P(\tau, \mu)$.

Definition 2 $P(\tau, \mu)\, d\tau \equiv$ *probability at time t that the next reaction in the volume V will occur in the infinitesimal 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 continuous time τ and discrete variables μ.

Gillespie [28, 29] derived 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 h_μ is defined to be the number of distinct molecular reactant combinations for reaction R_μ found to be present in V at time t, whereas $0 \leq \tau < \infty$, $\tau \in \mathbb{R}$, $1 \leq \mu \leq M$, $\mu \in \mathbb{N}$ and $P(\mu, \tau) = 0$ for all other τ, μ.

The computational procedure uses Monte Carlo technique to simulate the stochastic process described by $P(\tau, \mu)$. Then the simulation algorithm can be described 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 $c_1 h_1, \ldots, c_M h_M$. Specify and store a series of "sampling times" $t_1 < t_2 < \ldots$ and a "stopping time" t_{stop}.

Step 1: Generate by a suitable Monte Carlo technique one random pair (τ, μ). (How to do this is shown below.)

Step 2: Using the numbers τ, 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 $c_\nu h_\nu$ quantities for those reactions R_ν whose reactants X_i-values have just been changed.

Step 3 If t hast 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_{stop} only one realization of the stochastic process is obtained. In order to get a statistically complete picture of the temporal evolution of the system, we have to carry out several independent realizations, each starting with the same initial set of molecules and proceeding the same time t_{stop}.

One method to generate the random pair (τ, μ) according to $P(\tau, \mu)$ is called the "direct" method. It is based on fact that $P(\tau, \mu)$ can be written in terms of conditional probabilities. Hence $P(\tau, \mu) = P_1(\tau) \cdot P_2(\mu|\tau)$. Here $P_1(\tau)d\tau$ 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 one gets $P_1(\tau) = \sum_{\mu=1}^{M} P(\tau, \mu)$. Therefore it follows for $P_2(\mu|\tau)$: $P_2(\mu|\tau) = P(\tau, \mu) / \sum\limits_{\nu=1}^{M} P(\tau, \nu)$.

Substituting $P(\tau,\mu)$ finally yields $P_1 = a \cdot \mathrm{e}^{-a\tau}, \quad P_2(\mu|\tau) = a_\mu/a$. where $a_\mu = h_\mu c_\mu$ and $a = \sum\limits_{\mu=1}^{M} a_\mu$. In particular $P_2(\mu|\tau)$ turns out to be independent of τ.

The idea of the 'direct' method is therefore (first) to generate the random value according to $P_1(\tau) = a \cdot \mathrm{e}^{-a\tau}$ and (second) to generate generate a random integer μ according to $P_2(\mu|\tau) = a_\mu/a$.

Using the output of a uniform random number generator, 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 by evaluating a number r_2 from the uniform distribution in the unit interval and taking μ as the integer fulfilling

$$\sum_{\nu=1}^{\mu-1} a_\nu < r_2 a < \sum_{\nu=1}^{\mu} a_\nu .$$

"Replication Deletion Process"

Next we introduce a model that follows the Moran scheme [52]. It has the advantage of allowing explicit expressions for many quantities of evolutionary interest.

Let N be a natural number, with $N \geq 2$ and let $\mathbf{V}$ be a finite family of vertices $\mathbf{V} = (v_i \,|\, i \in \mathbb{N}_N)$ where $\{v_i \,|\, i \in \mathbb{N}_N\} \subset \mathcal{Q}^n_\alpha$. $\mathbf{V}$ is called a *population* in $\mathcal{Q}^n_\alpha$.
We consider an arbitrary partition of the set $\mathrm{v}[\mathcal{Q}^n_\alpha]$ into disjoint sets of vertices G_i, $i = 1,\ldots,m$ with $\bigcup G_i = \mathrm{v}[\mathcal{Q}^n_\alpha]$. A fitness function induced by $\{G_i\}_{i=1,\ldots,m}$ can be given by

$$f(v) := \sigma_i \quad \text{iff } v \in G_i, \quad \sigma_i \in \mathbb{R}^+, \quad i = 1,\ldots,m.$$

A *Replication Deletion Process* with respect to a partition $\{G_i\}_{i=1,\ldots,m}$ is a mapping from a family $\mathbf{V} = (v_i \,|\, i \in \mathbb{N}_N)$ to a family $\mathbf{V}' = (v'_i \,|\, i \in \mathbb{N}_N)$ as follows:

- An ordered pair of vertices (v_ℓ, v_k) is selected from $\mathbf{V} = (v_i \,|\, i \in \mathbb{N}_N)$. For

$$x_j := \mathrm{res}_{G_j}\, \phi(G_j), \quad j = 1,\ldots,m$$

v_ℓ is chosen with probability $\sigma_j x_j / \sum_t \sigma_t x_t$ from $\{v_i \mid i \in \mathbb{N}_N,\}$ with $\mathrm{res}_{G_j}(v_i) > 0$. The second coordinate v_k is selected with uniform probability $1/(N-1)$ from $\{v_i \mid i \neq \ell, i \in \mathbb{N}_N\}$. We assume the time $\hat{T}$ between these mappings to be exponentially distributed (scaled by the mean fitness)

$$Prob\{\hat{T} \leq t\} = \mathrm{e}^{-(\sum_i \sigma_i x_i)t}.$$

- The vertex $v_\ell = (x_1, \ldots, x_n)$ is mapped randomly into the vertex $v^* = (x'_1, \ldots, x'_n)$. This is done by mapping each coordinate x_i to $x'_i \neq x_i$ with probability p where all $x'_i \neq x_i$ are equally distributed and leave the coordinate fixed otherwise. This mapping is called "replication"
- Finally we delete the second coordinate v_k.
- The pair (v_ℓ, v_k) is mapped into the pair (v_ℓ, v^*). Thereby we obtain **V**' by substituting v_k by v^*.

Notation

$\mathcal{A}, \mathcal{B}$	alphabet
$\mathcal{B}(x,\ell)$	$\overset{\text{def}}{=} (\ell-1)!/[x(x+1)\dots(x+\ell-1)], \quad x \in \mathbb{R},\ \ell \in \mathbb{N}$; Betafunction
α, β	size of predefined alphabet
$\mathcal{C}[s]$	graph of compatible sequences with respect to s
$\mathbf{C}[s]$	set of compatible sequences with respect to s
$\mathrm{e}[G]$	edge set of a graph G
$\mathcal{E}_{ik}(s,s')$	$:= \{v \mid d(v,\Gamma_n[s])$ and $d(v,\Gamma_n[s']) = k\}$; distance class with respect to incompatible base pairs in s and s',
f_n	combinatory map
$f_s, f_{s,s'}$	fitness function with respect to neutral networks of s and s'
$\Gamma_n[s]$	neutral network corresponding to secondary structure s of length n
$\mathcal{G}(H)$	set of all induced subgraphs in the finite graph H
$\mathbf{I}[s,s']$	$\mathbf{C}[s] \cap \mathbf{C}[s']$, intersection, overlap
$\mathcal{I}[s,s']$	intersection graph
Λ_{ik}	$\sigma w_{ik} - w_{0k}$
$\bar{\lambda}$	$\lvert\Gamma_n[s]\rvert/\lvert\mathbf{C}[s]\rvert$, fraction of neutrality
λ_u, λ_p	average fraction of neutral neighbors unpaired and paired
N	average or fixed population size
n	parameter of system size, in particular chain length
n_u, n_p	number of unpaired and number of paired bases
P	probability
$P_{x,x'}$	infinitesimal transition probability in Markov chain
Π	pairing rule of an alphabet
$\Pi(s)$	set of contacts of an RNA secondary structure s, pairing rules of s
p	single digit error rate
$1-p$	single digit accuracy
$\pi(s)$	representation of secondary structure s in symmetric group S_n
z	$:= (i_1,\dots,i_k)$, orbit corresponding to a pair of secondary structures s, s', cycle of $\pi(s) \circ \pi(s')$ with respect to the pairing rules $\Pi(s)$ and $\Pi(s')$

$\mathcal{Q}_\alpha^n$	generalized hypercube of dimension n over an alphabet of cardinality α
$\rho(s,s'),\rho$	action probability on the intersection
$\mathcal{S}_n$	set of all secondary structures of sequences of chain length n
s, s'	secondary structure
σ	positive real number greater than 1, superior fitness
$\mathbf{V}$	the population
$\mathbf{V}_\mu, \mathbf{V}_\nu$	number of masters and nonmasters in population
v	vertex of a graph or element of a population $\mathbf{V}$
$\mathrm{v}[G]$	vertex set of a graph G
$W_{\mu\mu}(p)$	average probability by mutating a master with error rate $1-p$ to get a master
$W_{\nu\mu}(p)$	average probability by mutating a nonmaster with error rate $1-p$ to get a master
w_{ij}	short form for $W_{\cdot,\cdot}$; if $i=k$ it is called *phenotypic fixation probability*, otherwise *phenotypic transition probability*
w^*	critical phenotypic fixation probability
X_t, Y_t, Z_t	integer valued random variables

References

[1] C. K. Biebricher. Darwinian selection of self-replicating RNA molecules. *Evol. Biol.*, 16:1–52, 1983.

[2] C. K. Biebricher, M. Eigen, and W. C. Gardiner Jr. Kinetics of RNA replication. *Biochem.*, 22:2544–2559, 1983.

[3] B. Bollobás. *Graph Theory, An Introductory Course.* Springer, New York, 1990.

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

[5] J. Crow and M. Kimura. *An Introduction to Population Genetics Theory.* Alpha Edition, 1970.

[6] K. M. Currey, B. M. Peterlin, and J. V. jr Maizel. Secondary structure of poliovirus RNA: correlation of computer-predicted with electron microscopically observed structure. *Virol.*, 148:33–46, 1986.

[7] C. R. Darwin. The origin of species. In *Everyman's Library*, volume 811, page 81. J. M. Dent & Sons, Aldine House, Bedfort Street, London, UK, 1928.

[8] R. Dawkins. *The selfish gene.* Oxford Univ. Press, 1976.

[9] M. O. Dayhoff and C. M. Park. Cytochrome-c: Building a phylogenetic tree. In M. O. Dayhoff, editor, *Atlas of protein sequence and structure*, pages 7–16. National Biomedical Research Foundation, Silver Springs (Md.), 1969.

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

[11] T. Dobzhansky, F. Ayala, G. Stebbins, and J. Valentine. *Evolution.* W. H. Freeman & Co., San Fransico, CA, 1977.

[12] M. Eigen. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 10:465–523, 1971.

[13] M. Eigen. Macromolecular evolution: Dynamical ordering in sequence space. *Ber. Bunsenges. Phys. Chem.*, 89:658–667, 1985.

[14] M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasispecies. *J. Phys. Chem.*, 92:6881, 1988.

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

[16] M. Eigen and P. Schuster. The Hypercycle A: A principle of natural self-organization: Emergence of the Hypercycle. *Naturwissenschaften*, 64:541–565, 1977.

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

[18] W. J. Ewens. *Mathematical Population Genetics.* Springer, Berlin, 1979.

[19] M. Feinberg. Mathematical aspects of mass action kinetics. In L. Lapidus and N. R. Amundson, editors, *Chemical Reactor Theory. A Review*, pages 1–78. Prentice-Hall, Inc. Englewood, 1977.

[20] M. Fisz. *Wahrscheinlichkeitsrechnung und mathematische Statistik.* VEB Deutscher Verlag der Wissenschaften, Berlin, 1989.

[21] W. Fontana, T. Griesmacher, W. Schnabl, P. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Mh. Chem.*, 122:795–819, 1991.

[22] W. Fontana, D. Konings, P. Stadler, and P. Schuster. Statistics of RNA secondary structures. *SFI preprint 92-02-007, Biopolymers*, 33(9):1389–1404, 1993.

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

[24] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophys. Chem.*, 26:123–147, 1987.

[25] 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.

[26] 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, 1986.

[27] C. W. Gardiner. *Handbook of stochastic Methods.* Springer, 1990.

[28] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[29] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Chem. Phys.*, 81:2340–2361, 1977.

[30] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Mh. Chem.*, 127:355–374, 1996.

[31] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Mh. Chem.*, 127:375–389, 1996.

[32] H. Haken. *Synergetics.* Springer, Berlin, 1983.

[33] R. W. Hamming. *Coding and Information Theory.* Prentice Hall, Englewood Cliffs (N. J.), 1980.

[34] T. F. Havel, I. D. Kuntz, and G. M. Crippen. The theory and practice of distance geometry. *Bull. Math. Biol.*, 45(5):665–720, 1983.

[35] D. D. Hegedus, T. A. Pfeifer, J. M. MacPherson, and G. G. Khachatourians. Cloning and analysis of five mitochondrial tRNA-encoding genes from the fungus beauveria bassiana. *Gene*, 109:149–54, 1991.

[36] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Mh. Chem.*, 125(2):167–188, 1994.

[37] J. Hofbauer and K. Sigmund. *Evolutionstheorie und dynamische Systeme.* Paul Parey, 1984.

[38] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucleic acids research*, 12:67–74, 1984.

[39] M. A. Huynen. Exploring phenotype space through neutral evolution. *subm. to J. Mol. Biol.*, 1995. SFI preprint 95-10-100.

[40] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996. SFI preprint 95-01-006, LAUR-94-3763.

[41] T. H. Jukes. Comparison of the polypeptide chains of globin. *J. Mol. Evol.*, 1:46–62, 1971.

[42] S. Karlin and H. M. Taylor. *A first course in stochastic processes.* Academic Press, second edition, 1975.

[43] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

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

[45] M. Kimura and T. Ohta. *Theoretical Aspects of Population Genetics.* Princeton Univ. Press, 1971.

[46] D. Konings and P. Hogeweg. Pattern analysis of RNA secondary structure, similarity and consensus of minimal-energy folding. *J. Mol. Biol.*, 207:597, 1989.

[47] Y. Ma and M. B. Mathews. Comparative analysis of the structure and function of adenovirus virus-associated RNAs. *J. Virol.*, 67:6605–6617, 1993.

[48] G. A. Mackie. Nucleotide sequence of the gene for ribosomal protein S20 and its flanking regions. *J. Biol. Chem.*, 256:8177–82, 1981.

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

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

[51] J. G. Mendel. Versuche über Pflanzenhybride. *Verh. Naturw. Vereins, Brünn*, 4:3–57, 1865.

[52] P. A. P. Moran. Random processes in genetics. *Proc. Camb. Phil. Soc.*, 54:60–70, 1958.

[53] M. Nowak and P. Schuster. Error thresholds of replication in finite populations, mutation frequencies and the onset of Muller's ratchet. *J. theor. Biol.*, 137:375–395, 1989.

[54] 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.

[55] J. Pütz, J. Puglisi, C. Florentz, and R. Giege. Identity elements for specific aminoacylation of yeast tRNA (asp) by cognate aspartyl-tRNA synthetase. *Science*, 252:1695–1699, 1991.

[56] C. Reidys. *Neutral Networks of RNA Secondary Structures.* Dissertation, Friedrich-Schiller-Universität Jena, 1995.

[57] C. Reidys, C. V. Forst, and P. Schuster. Replication on neutral networks of RNA secondary structures. *subm. to Bull. Math. Biol.*, 1994.

[58] C. Reidys and P. F. Stadler. Bio-molecular shapes and algebraic structures. *Computers Chem.*, 20(1):85–94, 1996.

[59] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps and application on RNA secondary structures. *SFI preprint 95-07-058*, 1995.

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

[61] P. Schuster. Dynamics of molecular evolution. *Physica D*, 22:100 – 119, 1986.

[62] P. Schuster. The role of neutral mutations in the evolution of RNA molecules. In S. Suhai, editor, *Computational Methods In Genome Research.* Plenum Press, New York, 1996.

[63] 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.

[64] P. Schuster and K. Sigmund. Dynamics of evolutionary optimization. *Ber. Bunsenges. Phys. Chem.*, 89:668–682, 1985.

[65] P. Schuster and P. F. Stadler. Landscapes: Complex optimization problems and biomolecular structures. *Computers Chem.*, 18:295 – 324, 1994.

[66] P. Schuster and J. Swetina. Stationary mutant distributions and evolutionary optimization. *Bull. Math. Biol.*, 50:635, 1988.

[67] P. Schuster, J. Weber, W. Grüner, and C. Reidys. Molecular evolutionary biology: From concepts to technology. In H. Flyvbjerg, J. Hertz, M. H. Jensen, O. G. Mouritsen, and K. Sneppen, editors, *Physics of Biological Systems: From Molecules to Species.* Springer, Berlin, 1996.

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

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

[70] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 17:213, 1971.

[71] P. F. Stadler and P. Schuster. Mutation in autocatalytic networks - an analysis based on perturbation theory. *J. Math. Biol.*, 30:597–631, 1992.

[72] M. Tacker. *A model for an RNA-melting kinetics landscape.* Masterthesis, Inst. Theor. Chem., Univ. Vienna, 1992.

[73] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Robust properties of RNA secondary structure folding algorithms. *Eur. Biophys. J.*, in press, 1994.

[74] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Ann. Rev. Biophys. Biophys. Chem.*, 17:167–192, 1988.

[75] R. D. Wachter, M. Chen, and A. Vandenberghe. Conservation of secondary structure in 5s ribosomal RNA: a uniform model for eukaryotic, eubacterial, archaebacterial and organelle sequences is energetically favorable. *Biochimie*, 64:311–329, 1982.

[76] M. S. Waterman. Secondary structure of single - stranded nucleic acids. In *Studies on foundations and combinatorics, Advances in mathematics supplementary studies*, volume 1, pages 167–212. Academic Press New York, 1978.

[77] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids. A structure for Desoxyribose Nucleic Acid. *Nature*, 171:737–738, 1953.

[78] A. Weismann. The continuity of the germ-plasm as the foundation of a theory of heredity. In J. A. Moore, editor, *Readings in Heredity and Development.* Oxfort Univ. Press, 1885.

[79] S. Wright. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In D. F. Jones, editor, *Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

[80] G. K. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, Reading (Mass.), 1949.

[81] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46(4):591–621, 1984.

[82] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl. Acid. Res.*, 9:133–148, 1981.

Table of Contents

1. Introduction 2

2. Evolutionary Dynamics 7

3. Neutral Networks of RNA Secondary Structures 11

  3.1. RNA Secondary Structures and Compatible Sequences 11

  3.2. Neutral Networks as Randomly Induced Subgraphs 15

  3.3. Two Neutral Networks in Sequence Space 21

4. Dynamics on Neutral Networks 29

  4.1. The Basic Model 32
    4.1.1. Constant Population Size 32
    4.1.2. Fluctuating Population Size 37

  4.2. One Neutral Network – Single Shape Landscape 40
    4.2.1. Phenotypic Fixation Probabilities 40
    4.2.2. Stochastic Approach 43

  4.3. Two Neutral Networks – Double Shape Landscape 45
    4.3.1. Phenotypic Fixation and Transition Probabilities 47
    4.3.2. Deterministic Approach 50
    4.3.3. Stochastic Approach 54

5. Neutral Evolution – Computational Results 61

  5.1. Neutral Evolution Based On RNA Folding 63

  5.2. Neutral Evolution Based On Random Graphs 69

  5.3. RNA Folding versus Random Graphs 74

  5.4. A Population in Sequence Space and Time 81

6. Discussion 87

Appendix A: Parameters 92

Appendix B: Stochastic Simulation of Finite Populations Dynamics 94

Notation 99

References 101

List of Publications

[1] C.V. Forst, C. Reidys, and J. Weber. Evolutionary dynamics and optimization: Neutral Networks as model-landscapes for RNA secondary-structure folding-landscapes. In F. Morán, A. Moreno, J. Merelo, and P. Chacón, editors, *Advances in Artificial Life*, volume 929 of *Lecture Notes in Artificial Intelligence*, Berlin, Heidelberg, New York, 995. ECAL '95, Springer. Santa Fe Preprint 95-10-094

[2] P. Schuster, J. Weber, W. Grüner, and C. Reidys. Molecular Evolutionary Biology: From Concepts to Technology. In H. Flyvbjerg, J. Hertz, M. H. Jensen, O. G. Mouritsen, and K. Sneppen, editors, *Physics of Biological Systems: From Molecules to Species*, Springer-Verlag Berlin/Heidelberg, 1996

[3] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler and P. Schuster. Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration. I. Neutral Networks. *Monath. Chem.*, 1996, 127, 355-374

[4] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler and P. Schuster. Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration. II. Neutral Networks. *Monath. Chem.*, 1996, 127, 375-389

Tabellarischer Lebenslauf

Name:	Jacqueline Weber
Anschrift:	Langendembach 70 A, 07381 Langenorla
	Tel.: 03647/418056; e-mail: jac@imb-jena.de
Geburtsdatum:	5. September 1969
Geburtsort:	Bergen, Rügen
Familienstand:	ledig
Staatsangehörigkeit:	deutsch
Schulausbildung:	1976-1986, Polytechnische Oberschule Wolgast
	1986-1988, Erweiterte Oberschule Wolgast
	1988, Abitur
Hochschulstudium:	1988-1993, Ernst-Moritz-Arndt-Universität Greifswald
	Diplomstudiengang Mathematik mit Beifach Physik
Examen:	7/1993, Abschlußnote ”sehr gut”
Arbeitsverhältnisse:	seit 10/1993, wiss. Mitarbeiter am IMB,
	Abteilung Molekulare Evolutionsbiologie
Dissertation:	”Dynamik neutraler Evolution”; 10/1993-9/1996
Besondere Kenntnisse:	Englisch, Französisch, Russisch
Ehrenamtliche Tätigkeit:	Übungsleiter Fach Tauchsport
Hobbys:	Unterwasserrugby, Tauchsport

Ort/Datum

Unterschrift

Selbständigkeitserklärung

Ich erkläre, daß ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel und Literatur angefertigt habe.

Ort/Datum	Unterschrift