

Neutral Networks of Minimum Free Energy

RNA Secondary Structures

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium

Vorgelegt der

Formal- und Naturwissenschaftlichen Fakultät

der Alma Mater Rudolphina zu Wien

von

Ulrike Göbel

im Mai 2000

Zusammenfassung

Die vorliegende Arbeit betrachtet die Sequenz-Sekundärstruktur-Abbildung in RNA von drei verschiedenen Blickwinkeln aus. Die einfachste Herangehensweise, die gleichzeitig die detaillierteste Information liefert, zählt systematisch alle Sequenzen einer gegebenen Länge auf und bestimmt für jede von ihnen die Struktur. Mit einem realistischen Strukturvorhersagealgorithmus (wie minimale-freie-Energie-Faltung) und dem natürlichen Basenalphabet kommt dieser Ansatz aus Zeit- und Platzgründen jedoch nur für sehr kurze Sequenzen in Frage. Als ein Beispiel werden in dieser Arbeit durch vollständige Aufzählung gewonnene Faltungsdaten für den Sequenzraum \mathcal{Q}_{AUGC}^{16} der Sequenzen der Länge 16 über dem natürlichen Alphabet vorgestellt und die Anwendbarkeit eines Zufallsgraphenmodells für die neutralen Netze dieses Raumes wird diskutiert. Die Faltungslandschaft zeigt die bekannten Eigenschaften der Sequenz-Sekundärstruktur-Abbildung von RNA: es gibt wenige häufige und viele seltene Strukturen, Sequenzen, die in häufige Strukturen falten, sind annähernd isotrop über den Sequenzraum verteilt und gleichzeitig räumlich nah zu Sequenzen, die in beliebige andere häufige Strukturen falten, und der Raum besitzt ausgedehnte neutrale Netze. Die bei einer so kurzen Sequenzlänge ausgebildeten Strukturen sind notwendigerweise wenig stabil, wodurch Mutationen häufig kontext- und positionsabhängige Effekte haben. Das ist wahrscheinlich ein Grund dafür, daß die Netze in \mathcal{Q}_{AUGC}^{16} weniger dicht sind als vom Zufallsgraphenmodell vorhergesagt, jedoch stärker verbunden. Außerdem ist die offene Kette die bei weitem häufigste Konformation und daher erscheinen die neutralen Netze der Strukturen in ihrem Netz eingebettet.

Vollständige Aufzählung kommt für Sequenzen der Länge, wie sie natürlicherweise in Zellen gefunden werden, nicht in Frage, so daß die Graphstruktur der neutralen Netze solcher Sequenzen im allgemeinen unbekannt ist. Im zweiten Teil der Arbeit wird eine Methode vorgeschlagen, die es erlaubt, den Grad der Verbundenheit eines Netzwerks aus einer kleinen Stichprobe seiner Sequenzen zu schätzen. Die Methode setzt voraus, daß für zwei Sequenzen, die zur gleichen Komponente des Netzwerks gehören, ein verbindender Pfad existiert, der in der *Hülle* des kürzestmöglichen Pfades enthalten ist (*lokale Verbundenheit*). Diese für allgemeine Graphen nicht voraussetzbare Eigenschaft wird vom Zufallsgraphenmodell für die neutralen Netze in Faltungslandschaften vorhergesagt. In der Arbeit wird gezeigt, daß sie in Netzen des Raumes \mathcal{Q}_{GC}^{30} so weit erfüllt ist, daß der Anteil lokal verbundener Sequenzpaare in einer Stichprobe das Vorhandensein einer oder mehrerer Zusammenhangskomponenten im gesamten Netzwerk zuverlässig schätzt.

Im letzten Teil der Arbeit wird schließlich ein neues Konzept struktureller Neutralität eingeführt, das in mancher Hinsicht der biologischen Realität besser entspricht. Zu einem *teilstruktur-neutralen Netz* sollen alle diejenigen RNA-Sequenzen gehören, in deren Sekundärstruktur eine bestimmte Teilstruktur auftritt. Als Objekt spezifischer Erkennung reichen solche Struktur motive häufig aus, um eine bestimmte Funktion zu vermitteln. Ob Sequenzanteile zusätzlich zur für die Funktion benötigten Teilstruktur den Anteil neutraler Mutationen erhöhen oder nicht, hängt von zwei Variablen ab: von der Anzahl von RNA-Sequenzen gegebener Länge, die mit genau n Kopien der Teilstruktur kompatibel sind (für diese Zahl wird eine geschlossene Formel angegeben), und von der Wahrscheinlichkeit, mit der eine kompatible Teilsequenz bei verschiedenen Gesamtlängen tatsächlich die Teilstruktur annimmt. Längere Gesamtlängen werden begünstigt, wenn die Teilstruktur wenig komplex ist (mit entsprechend hoher Wahrscheinlichkeit der *de novo* Ausbildung kompatibler Teilsequenzen in den zusätzlichen Sequenzanteilen) und gleichzeitig stabil genug, um auch in einem ausgedehnten Sequenzkontext noch zu falten. Als erster Schritt hin zu einem Zufallsgraphenmodell von teilstruktur neutralen Netzen werden Alternativen der Graphenstruktur diskutiert, die man auf den zu einem Strukturmotiv kompatiblen Sequenzen definieren kann. Das Netzwerk würde dann als Teilgraph des Graphen der Kompatiblen modelliert.

Abstract

In this thesis the sequence to secondary structure mapping of natural RNA molecules is approached from three different directions. Exhaustive folding of an entire sequence space yields the most highly resolved picture of the landscape, yet if the model is to be close to biological reality (that is, the natural base alphabet is used and the folding algorithm is the biophysically meaningful minimization of the free energy of folding) then this approach is too costly with respect to computation for all but very small sequence spaces. Exhaustive folding data is presented for the sequence space Q_{AUGC}^{16} and the applicability of a random graph model to the neutral networks of this space is discussed. The folding landscape exhibits the features which are known to be characteristic for the RNA sequence to secondary structure mapping: a distinction of common and rare structures, a near isotropic distribution of the sequences which fold into the former, closeness in space of sequences which fold into different common structures, and huge neutral networks. Because of the necessarily restricted number of base pairs at such a short chain length, the structures are not very stable, leading to frequent context and position specific effects of mutations. As a result, the networks of the space are less dense than predicted by the random graph model, but they are more connected. In addition, the open chain represents by far the most frequent conformation and, hence, the neutral networks of all structures appear to be embedded into its net.

For RNA sequences over the natural alphabet which are of lengths typically found in cells, it is not feasible to exhaustively fold the entire landscape, and therefore the graph structure of a neutral network is generally unknown. In the second part of the thesis, we propose a method which estimates the total degree of connectivity of a neutral network while using only a small sample of its constituent sequences. It assumes that any two sequences in the same component are connected by a path which is at most a single mutation away from the shortest possible connecting path. This property, *local connectivity*, is predicted for neutral networks of folding landscapes by random random graph theory. We show that in Q_{GC}^{30} it holds sufficiently well to distinguish connected networks from disconnected ones based on the the fraction of locally connected pairs in a small sample.

Finally, we introduce a new concept of structural neutrality, which in many cases comes closer to biological reality. A *substructure neutral network* includes all RNA sequences which, in their minimum free energy secondary structure, share a common structural motif. Often such motifs, as the target of specific recognition, are enough to convey a common function. Whether or not an increased total sequence length leads to enhanced neutrality depends on two parameters: the number of RNA sequences of a given length which are compatible with exactly n copies of the substructure (for which we give a closed formula) and the probability of a compatible subsequence to actually adopt the substructure at different total lengths. Longer total lengths are favoured with small, inherently stable substructures. Small size means a high probability of more than a single compatible subsequence per sequence, which however can only be exploited if folding is not too much affected by a larger sequence context. As a first step towards a random graph model of substructure neutral networks, we discuss graph structures which can be imposed on the set of compatible sequences (modelling the neutral network as an induced subgraph in the compatibles).

Danksagung

An dieser Stelle möchte ich allen danken, die am Zustandekommen dieser Arbeit beteiligt waren.

Mein besonderer Dank gilt dabei

Prof. Peter Schuster für seine stete Bereitschaft zur Unterstützung und seine nie endende Geduld,

Dr. Christian Forst und der Arbeitsgruppe Molekulare Evolutionstheorie am Institut für Molekulare Biotechnologie (IMB) in Jena für ein ausgezeichnetes Arbeitsklima, das leider zu früh den Umständen zum Opfer fiel,

Prof. André Rosenthal und der Abteilung Genomanalyse am IMB, die mir vom Januar 1998 bis zum Juli 1999 einen Arbeitsplatz zur Verfügung stellten und damit die Fortführung der Arbeit ermöglichten,

Dr. Klausdieter Weller und *Susanne Fabisch* für die Begleitung bei Reisen nach Wien,

Prof. Thomas Mitchell-Olds und der Bioinformatik-Crew der Abteilung Genetik und Evolution am Max-Planck-Institut für Chemische Ökologie in Jena (*Dr. Thomas Wiehe*, *Steffi Gebauer-Jung* und *Dr. Bernhard Haubold*) für die gastliche Aufnahme und Unterstützung während der heißen Phase der Fertigstellung der Arbeit. Dr. Thomas Wiehe verdanke ich nicht nur zahlreiche Anregungen, sondern er hat mich mit seiner Freude am wissenschaftlichen Dialog auch immer wieder zum Weitermachen ermutigt.

Meinen Eltern danke ich für ihren großen Einsatz nicht nur in den letzten fünf Jahren.

Last not least: diese Arbeit würde nicht existieren ohne *Katrin Vader*, die mir in einer kritischen Zeit beigestanden hat. Ihr ist sie gewidmet.

Table of Contents

1. Introduction	1
1.1. Where are We Now: Views on Life from Antiquity to Molecular Biology	1
1.2. The Molecular Basis of Evolution	6
1.3. Organization of this Work	9
2. Configuration Spaces in Molecular Evolution	12
2.1. Biological Macromolecules	12
2.2. The Main Mathematical Tool: Graphs	16
2.3. Fixed Sequence Length: Hamming Graphs	17
2.4. When the Sequence Length Is not Fixed	19
2.5. Landscapes	21
2.6. Neutral Networks of RNA Secondary Structures	24
3. Sequence Space and Shape Space of Q_{AUGC}^{16}	34
3.1. Introduction	34
3.2. Methods	35
3.3. Which Structures Are Realized in Q_{AUGC}^{16} ?	41
3.4. Global Properties of Networks	43
3.5. Graph Topological Features	54
3.6. Energy Landscapes	81
4. Probing Large Neutral Networks	89
4.1. Introduction	89
4.2. Algorithm	90
4.3. Local Connectivity in Minimum Free Energy Neutral Networks of Q_{GC}^{30}	94
4.4. Outlook: Analyzing Large Sequence Spaces Over the Natural Alphabet	102
5. Neutrality With Respect To A Substructure	105
5.1. Introduction	105
5.2. Definitions and Abbreviations	106
5.3. The Compatible Sequences of a Substructure	107
5.4. Substructure Neutral Networks	121
6. Discussion	131
7. Appendix	137
Bibliography	155
Tabellarischer Lebenslauf	161
List of Publications	162

1 Introduction

1.1. Where are We Now: Views on Life from Antiquity to Molecular Biology

Up to well into the twentieth century it was not really understood what it is that distinguishes organisms from inorganic matter [62, 105, 101]. Living beings obviously contain dead material, but it is put to very special uses when part of them. In particular, many traits and actions of organisms are *purposeful*. According to an old tradition of thought, also commonly found in contemporary primitive cultures, all nature is actually alive, and there exists a continuity between inorganic and organic matter. In the European Middle Ages and Renaissance that was reflected in the believe in spontaneous generation and the idea of the “chain of being”, ascending from inorganic matter to low life forms, to higher organisms, to man. Spontaneous generation of higher organisms was disproven by Francesco Redi in 1668. With the growing acknowledgment of the complexity and diversity of organisms their functioning and their very existance became increasingly miraculous. The attempts to explain these phenomena followed two main traditions, which date back to classical antiquity and which were thought to be incommensurate. The *vitalistic* tradition, started by Aristotle, assumed that purposeful traits and actions must be caused by an analogue of the human free will. Such a *vis vitalis* (force of life) may or may not be supernatural. In the form of presumed physical forces which act solely on living matter it was still held up in the first half of the 20th century [18, 15]. *Mechanicism* stressed linear causality as it is known from the crafts. A mechanical cause predates its effect in time, and typically consists of a force acting on matter. The most radical proponent of mechanism in biology was Descartes [62], according to whom an organisms is nothing but a sophisticated machine.

Mechanicism, although seemingly more modern, does not give an explanation for the observed harmonious design of the “machines”. Jean Baptiste de Lamarck in 1800 [14], by falsely connecting two known facts, arrived at a completely new view at this problem: organisms may *self-design* by way of inheritance of those environment-induced alterations which can indeed be observed to occur (e.g. exercise leads to muscle development). Self-design (*adaptation*) in a changing environment will lead to *evolution* (change of organisms over time), and therefore will not only allow for a self contained mechanistic theory, but might also explain the observed diversity of life.

The structure of Lamarck’s theory is in fact the same as that of the vitalistic theories, with the final goal replaced by a physical “mould”, the environment, a perfect fit with which is actively strived for by the organisms. The rate of change in the environment sets an upper limit to the diversity which can be produced in a given time interval. *Charles Darwin*, whose name is today most tightly associated with the change of paradigm from a static organic world to evolution, in 1859 proposed a different mechanism, which would not only lead to adaptation and evolution, but in addition redefines “goals” in an important way.

Darwin’s ideas on the principles of heredity were as vague and speculative as those of Lamarck, as well as, for that matter, those of the entire scientific community of his time. The theories of the time could neither explain the relative stability of genetic information over the generations nor the occasional occurrence of hereditary variations. His own genetic theory, Pangenesis, was equally unsuccessful in this respect. That Darwin’s theory of evolution of 1859 [13] remained essentially valid up to our days was exactly due to the fact that it did *not* require any assumptions on how nature accomplishes this near-stability.

He observed two things: first, there is hereditary variation among the offspring of organisms, as is demonstrated by the ability of breeders to gradually improve their stocks by selecting the best offspring for further breeding. Second, as was pointed out by Thomas R. Malthus [60], the environment is a necessary factor for regulating the size of (human) populations: it makes sure

that only that fraction of all offspring will reach maturity which the environment can bear. In the presence of variation, Darwin concluded, those that survive would not just be a random sample, but rather those which are best adapted to the environment. If small improvements are hereditary, then complex adaptations can be gradually built up, leading to as perfect a “fit to the mould” as with Lamarck’s mechanism. The difference between Darwin’s and Lamarck’s theory (except for the mechanism, of course) is rooted in a different view on the relation between individual and species: for Lamarck, all individuals of a species are essentially indistinguishable and undergo a concerted adaptation towards a common “goal”, which is set by the environment. In Darwin’s view, it depends on the variations which happen to occur in the offspring of an individual which aspects of the environment are exploited for adaptation. Different individuals may take different routes, eventually leading to speciation. Thus the partial independence of evolutionary dynamics from the environment (and its dependence on the dynamics of occurrence of genetic alterations, mutations) is as inherent to the theory as its stress on diversity and speciation. This has been neglected by many of Darwin’s followers.

The 19th century laid an excellent groundwork of detailed descriptions, from whole organisms to subcellular structures. Morphological descriptions were sometimes supplemented by chemical analyses (discovery of the nucleic acids of the cell nucleus by Friedrich Miescher in 1869). The general trend was towards an interpretation of this body of knowledge in simple laws, in the style of physics. The phenomenon of individual development, as puzzlingly purpose directed as species specific adaptations, gained special interest. In 1883 Augustus Weismann introduced the distinction between a “legislative” part of the living substance and an “executive” one and located the former in the cell nucleus. His theory of inheritance assumes that only germ cells contain the full set of “genetic determinants”, while body cells lose specific subsets during the course of differentiation. All phenomena of transmission genetics are brought about by an unbroken lineage of germ cells, the germ track, and accordingly are not influenced by alterations to parts of the body (soma track):

the cause-effect relation between germ track and soma is strictly linear. This tenet, which greatly facilitates the study of genetic processes, is still part of contemporary molecular biology.

The basic rules of transmission genetics were found by Gregor Mendel in 1866 [63] and rediscovered during the first decade of the 20th century by Hugo de Vries, C. Correns and E. Tschermak. The simplicity of the rules was appealing, yet in the eyes of many biologists they did not capture that kind of continuous variation which were thought to be not only typical for organisms but also a prerequisite for Darwinian evolution. This discomfort, together with the lack of an adequate theory of the developmental effect of the hypothetical genes, led to a broad spectrum of genetic, developmental, and evolutionary theories during the first three decades of the twentieth century – one could call this the pre-paradigm phase of modern biology [56]. The first step towards its resolution was the gradual acceptance of the chromosomal theory of inheritance, which is connected with the name of T.H. Morgan. The Morgan lab demonstrated the existence of spontaneous alterations of Mendelian genes (mutations), and it was later shown that environmental factors can trigger their appearance, but do not determine the direction of change [16]. R.A. Fisher and J.B.S. Haldane demonstrated the ability of alleles with small positive selective values to spread through a population [25, 40]. Thus all assumptions of Darwin had been shown to be consistent with a theory of inheritance which assumed particulate genes, lined up in a linear order on the chromosomes, each specifying a unique trait and possibly being present in more than a single variant (allele) in a population. The paradigm which emerged during the period from approximately 1936 to 1947 is called the Synthetic Theory. It adopts the Darwinian explanation of adaptations by Natural Selection, but is a more comprehensive theory in that it is able to make quantitative predictions on the basis of Mendelian population genetics.

Classical genetics, elaborate as it was in the middle of the 20th century, was still merely a parsimonious description of the passage of traits through the generations. The physical nature of a Mendelian gene and how it causes a trait to appear was unknown. In 1944 Oswald Avery proved

that the transforming principle in aqueous cell-free extracts of pneumococci was desoxyribonucleic acid (DNA) [62]. That finding led to an intense interest in nucleic acids in the ensuing years, culminating in the elucidation of the molecular structure of DNA by James Watson and Francis Crick in 1953. The structure itself – two linear heteropolymeric strands held together by monomer-specific hydrogen bonding – suggested that the molecule might *replicate* by growing a complementary copy on each strand. During the 1950s and 1960s not only the mechanism of replication was proven, but a principle of gene action was discovered which had not been anticipated by any of the earlier hypotheses. Genes consist of segments of nucleic acids which *code* for the sequence of a protein molecule by means of a trinucleotide code. The old question whether biological processes are initiated by an outside agent or keep themselves going by unbroken chains of interactions was answered on an unexpected middle ground: the cell does indeed contain a “body of knowledge” by which it is able to initiate biochemical processes if needed, and to carry them out in ways which may be so far from a direct cause-effect relationship that it looks like design by an intelligent being. That body of knowledge however is a material part of the cell, synthesized by the cellular machinery on replication, and expressing itself via physical interactions. Causation on the subcellular level often is tantamount to the actualization of coded *information*: that is the missing link between vitalism and mechanicism.

Genes were found to express themselves via a short lived complementary copy made of ribonucleic acid (RNA). It is that copy which is translated into a protein chain, by means of a translation apparatus which, too, contains RNA molecules at crucial places. In the early days of molecular biology therefore the so called central dogma was coined: DNA makes RNA makes protein, the flow of information through this chain being strictly one way. It was a revival of August Weismann’s concept of an unbroken lineage of genetic material (DNA), which constitutes a linear cause for the body (proteins), with no back flow of information. During the 1970s it was found that certain viruses are capable of retrotranscribing RNA into DNA, and according to recent discoveries the

coding nucleotide sequence of some messenger RNAs is subjected to directed changes by complexes of proteins and RNA molecules [8]. Yet it still holds today that no mechanism is known which would “retrotranslate” a protein sequence into RNA or DNA.

It becomes however increasingly clear that linear causation does not capture the essentials of life. A network of mutual regulation pervades all biochemical processes, including the genome (the totality of the DNA of a cell or virus). The environment *can* influence the expression of genetic information by inducing the binding of a regulatory protein molecule or a chemical modification of the DNA. Such *epigenetic* changes can even be passed on to daughter cells. What distinguishes these processes from an inheritance of acquired characters is the fact that the environment can only trigger reactions which preexist in the sense that they are coded for in the genetic material.

1.2. The Molecular Basis of Evolution

According to modern thought, true heritable alterations of the genetic material are confined to random changes of coding nucleotide sequences. The number of possible nucleotide sequences of even moderate length is enormous. If there is a single optimal one, how can it be found by selection ? From the 1960s on that problem has been attacked from several directions, which today merge into a unified view on molecular evolution.

The two main arguments in resolving the problem stress a) mechanisms which maintain genetic diversity in natural populations, effectively parallelizing the search for better adapted variants, and b) mechanisms which shield the phenotype (the sum of traits of an organism which are exposed to selection) from alterations of the genotype (the genetic material), thereby allowing more variants to be tried out.

Manfred Eigen, Peter Schuster, and John McCaskill during the 1970s and 1980s developed a theory

of the maintainance and optimization of information in polynucleotide sequences under error prone replication. Manfred Eigen demonstrated in 1971 [19] that the only prerequisite for Darwinian evolution is an open system with replication far from thermodynamical equilibrium and living on limited resources. In such a system every polynucleotide sequence is associated with a spectrum of mutants, the *molecular quasispecies*, which, assuming a fixed mutation rate and sequence specific fitness distribution, is completely defined by the sequence (for a review see Manfred Eigen, John McCaskill and Peter Schuster 1989 [21]). The important point is that selection of a sequence is tantamount to selection of a quasispecies: the Darwinian process itself assures the maintainance of variability. The quasispecies is however only stable up to a defined mutation rate, the *error threshold*, which at a fixed single-nucleotide replication accuracy depends on the sequence length. Manfred Eigen and Peter Schuster in 1977 proposed a mechanism which would allow to store more information in polynucleotides without enlarging the individual sequences: the *hypercycle* [22, 23]. RNA, as a potentially genetically active nucleotide sequence which at the same time folds up into a defined structure (see chapter 2) by which it is able to specifically interact with other molecules and catalyze biochemical reactions [83, 65, 66], gained attention as a model system for molecular evolution during the 1980s and 1990s. Single RNA molecules can undergo Darwinian optimization, as has already been demonstrated by Sol Spiegelman in the early 1970s [86, 55]. A realistic approximation to their secondary structure, which often is the main determinant of tertiary structure and thus function, can be readily computed [45, 46, 26]. Walter Fontana and coworkers during the 1980s focused on Darwinian evolution scenarios on RNA sequences in which selection is for features of the (computed) secondary structure which is adopted by a sequence [30, 29, 27, 28, 32]. In the course of this work it became increasingly apparent that selectively neutral variants play an important role at least in this model landscape.

Ever since during the 1960s new electrophoretic methods revealed an unexpected diversity on the level of proteins, it has been discussed to what extent these variants were visible to selection. In

1969 Motoo Kimura [52] suggested that most of them were products of *neutral alleles*, drifting randomly through the population. This idea, which had already been proposed by Sewall Wright in 1932 [103] as a means to enhance the search capacity of evolving populations, was at first strongly opposed by most biologists. Only during the 1970s and 1980s, when more and more crystal structures of biological macromolecules were solved, it became understandable how very different amino acid (or RNA) sequences could indeed be selectively neutral: these sequences adopt three dimensional structures which are so similar that they are essentially indistinguishable in any relevant molecular interactions [5] (see chapter 2 for a review of the relation between sequences and structures). At the same time it became apparent that the problem of molecular optimization is not so hard as originally thought: very often it is enough to find one out of many equally “optimal” solutions. Neutral variants, by shielding the phenotype from genotypic diversity, themselves constitute a means to maintain this diversity in the population. In contrast to the quasispecies, there is no inherent limit to the amount of sequence diversity which can be stored this way.

During the 1990s the group of Peter Schuster concentrated on the role of neutral variants in molecular evolution, especially as exemplified by the RNA folding landscape. They were able to show analytically how increasing neutrality in the phenotypes leads to an increase of the *phenotypic error threshold* [33, 72]. What matters here is not the absolute number of genotypic realizations of a given phenotype, but the expected number of neutral variants among the outcomes of single mutational events. This leads to the concept of a *neutral network*: this is the set of all sequences which are neutral with respect to some phenotype together with the neighbourhood relation “accessible by a single mutation”, which turns the set into a graph. Such “evolutionary networks” had already been observed for the mapping of protein sequences to their (idealized) lattice structures by D.J. Lipman and W.J. Wilbur in 1991 [59]. For the mapping of RNA sequences over the binary alphabets $\{A, U\}$ and $\{G, C\}$ into their computed secondary structures

they were demonstrated by Walter Grüner *et al.* (1996) [37, 38]. Neutral networks convey an ideal combination of search capacity and robustness to mutations: the genotypes may diffuse over the network by single nucleotide exchanges without losing the currently optimal structure, until a non neutral mutant is encountered which increases fitness. The population will then switch to the network of that structure [48, 49, 96, 91]. In order to be of maximal use, they should be connected graphs and a maximal number of new phenotypes (structures) should be available in the 1-error neighbourhood of the network or even of a single sequence (*shape space covering*). (Walter Fontana and Peter Schuster proposed to base a topology in phenotype space on the accessibility of one network from the other [31].) That the sequence secondary structure mapping in RNA comes close to such an ideal case is suggested by theory [71] and from the available data (complete networks over the binary alphabets and statistical evaluation of networks over the natural alphabet [44] [88, 89]). This suggests that the success of the organic evolution to no small part is due to an excellent solution by Nature of the representation problem of her optimization task [98].

1.3. Organization of this Work

This work is devoted to the study of neutral networks of RNA minimum free energy secondary structures. So far the investigation of neutral networks of RNA secondary structures has followed two lines: a mathematical model of network topology based on random graph theory has been developed by Christian Reidys *et al.* [71]. It assumes a sequence independent probability of folding, which might not hold in natural structures. Walter Grüner *et al.* (1996) [37, 38] have explicitly determined the networks of RNA sequences up to length 30 over the binary alphabets $\{A, U\}$ and $\{G, C\}$. The folding model was that of a thermodynamic (minimum free energy) folding algorithm [46, 45] and thus can be assumed to have captured the essential features of

natural structures. Neutrality was defined with respect to the full length structure (that is, two sequences are members of the same neutral network if their structures coincide at every position in the sequence). The restriction to binary alphabets was due to technical constraints (limited storage).

This thesis extends the previous work in two respects. In the first part we present for the first time data on all neutral networks in a space of RNA sequences over the natural alphabet $\{A, U, G, C\}$. For reasons of limited storage the length of the sequences is restricted to 16. In the second part the concept of structural neutrality is extended to include *neutrality with respect to a substructure*, which comes closer to biological reality than requiring all positions to exactly match some reference secondary structure.

The work is organized as follows:

In chapter 2 we will summarize the biophysical background of the sequence structure mapping in biological macromolecules, as well as the mathematical tools which have been used to describe it, with a special focus is on neutral networks of RNA secondary structures.

Chapter 3 presents the results of the complete mapping of an RNA sequence space of length 16 over the natural alphabet $\{A, U, G, C\}$ to the respective minimum free energy secondary structures [45]. It is shown that the mapping from sequence to shape space for this length and alphabet possesses the properties described above which aid optimization: most neutral networks are *connect-ed graphs* (every pair of sequences on a network can be interconverted by a series of single nucleotide mutants which are on the network) and for a certain subset of structures it holds that nearly all of them can be reached by at least one single nucleotide mutation of a member of the network of any of these structures. For reasons which will be discussed the degree of connectivity is even higher than expected from mathematical theory.

Chapter 4 deals with the case of sequence lengths which no longer permit exhaustive folding. Mathematical theory predicts a property of neutral networks, *local connectivity*, which allows to

test pairs of sequences for the existence of a connecting path on the network in linear time. If this property holds then one can estimate the degree of connectivity of the total network from that of a sample of its constituent sequences. We show that the property holds quite well for minimum free energy secondary structure neutral networks of RNA sequences of length 30 over the binary alphabet $\{G, C\}$ and that it is possible to exploit this fact for a prediction of global connectivity. In chapter 5 the concept of structural neutrality is relaxed to include all sequences which share some *substructure*. It is well known from both the structures of natural macromolecules and the outcomes of artificial selection experiments that a specific function is often conveyed by a defined part of the total structure, other parts being more or less free to vary. In the first part of the chapter we give a closed formula for computing the number of RNA sequences of a given length which are *compatible* with exactly n copies of the substructure, that is, which contain exactly n (possibly overlapping) subsequences which can form all base pairs that are required in the substructure. In the second part properties of the substructure neutral network at different total sequence lengths are discussed for two example substructures of different complexity. It is shown that the population dynamics of a set of replicating sequences on the network is considerably influenced by the degree of complexity of the substructure.

Finally, the results obtained are discussed in chapter 6.

2 Configuration Spaces in Molecular Evolution

2.1. Biological Macromolecules

2.1.1. Sequences and Structures

The chemical constitution of biological macromolecules (DNA, RNA, proteins) is that of a linear heteropolymer. The constituent monomers ((desoxy)ribonucleotides, amino acids) engage in noncovalent interactions with other monomers of the same or different chains, thereby folding the chain up into a three dimensional structure.

Structure is described on several levels. *Primary structure* is the covalently linked sequence of monomers. On the level of *secondary structure*, only patterns of specific interactions are considered. In nucleic acids, secondary structure is tantamount to *base pairing* (Watson-Crick and G-U pairs in RNA). Protein secondary structure in contrast does not include long rang interactions, which is why it is possible to assign a secondary structural state (helix, extended, turn, coil) to a single residue in this case. *Tertiary structure* describes the detailed foldup of the chain in three dimensional space, including distances and relative orientations of interacting monomers. Finally, *quaternary structure* deals with associations of more than one macromolecule into supramolecular complexes. Under suitable conditions most biological macromolecules spontaneously adopt a well defined secondary and tertiary structure. It is the linear covalent configuration of the chain which somehow specifies a cascade of interactions, eventually resulting in a correct foldup. (Exactly how it does so is still unknown. Solving the “folding problem”, especially in proteins, is one of the big challenges of contemporary bioinformatics.) The folding pathway is one example of *coded information* which

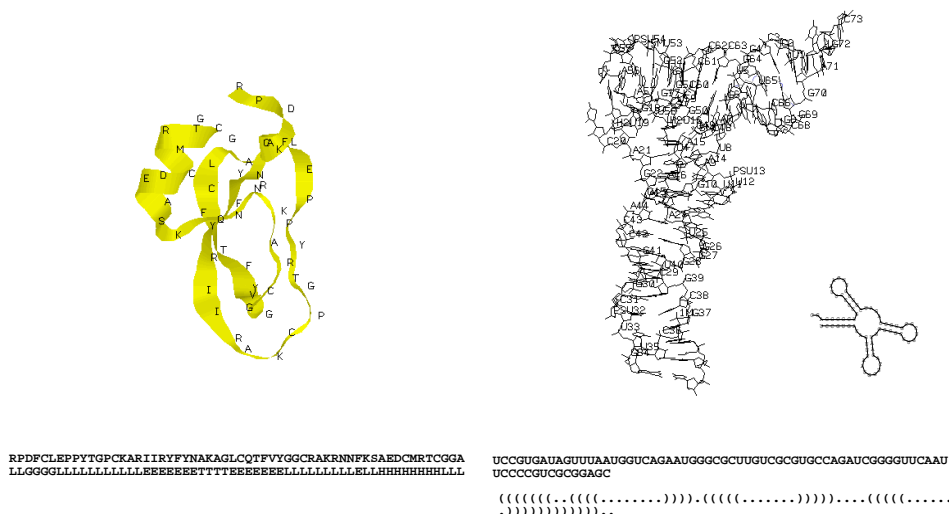


Figure 1: Structure in proteins and nucleic acids. *Left:* The structure of 6pti (pancreatic trypsin inhibitor). The secondary structure of the molecule consists of two helices and a central twisted beta sheet. Protein secondary structures are locally defined: a single residue is assigned to some structural state according to its torsion angles (states in this example are *H* – *alpha helix*, *G* – *3-10 helix*, *E* – *extended (beta strand)*, *L* – *loop*). A secondary structure element consists of several adjacent residues which are in the same state. How a beta *sheet* is composed of its constituent beta *strands* belongs to the level of tertiary structure in proteins. *Right:* The structure of yeast *tRNA^{Asp}*. In nucleic acids, the secondary structure is defined to be the pattern of Watson-Crick and possibly *G – U* base pairs. In a linear representation, the two positions which form a pair are assigned to a left and right matching parenthesis. Alternatively, the pure secondary structure can be depicted as a planar graph (inset). The relative orientation of the paired stacks into an L-shaped three dimensional structure belongs to the level of tertiary structure.

is stored in the primary structure. Coding (in the primary sequence) and serving as molecular machines by means of a sophisticated tertiary structure are the two main aspects of biological macromolecules. In modern organisms different kinds of molecules are specialized in one or the other of these tasks. DNA takes the idea of coding to the extreme by specifying, in its primary sequence, the primary sequence of *other* macromolecules (RNA, proteins). Higher order structure does not play a role in DNA: it is a linear molecule, which by means of base pairing forms a single intermolecular double helix with a complementary strand. Proteins are at the other extreme of the spectrum: they can adopt a variety of complicated three dimensional structures (and therefore

serve as the working horses of the cell), yet all their primary sequence codes for its own folding pathway.

In between there is RNA. It is intermediate in its capabilities as well as a true intermediate in the process by which the genetic information stored in DNA is expressed: the immediate product of a coding segment (gene) in DNA is always a complementary RNA molecule, which then either directly serves a functional role or else in a second step is translated into a protein chain. RNA molecules adopt complicated secondary structures by means of *intramolecular* base pairing, which then further fold up into tertiary structure. Although less structurally versatile than proteins, cells do use them for structural and catalytic tasks [83]. Like the peptidyl transferase activity in ribosomes (which at least in part depends on the RNA moiety [66, 51], these tasks are sometimes very basic to the cellular machinery. RNA molecules in addition can readily evolve new functions under appropriate selection [11, 106] and they are able to autocatalytically rearrange themselves, mimicking recombination [10]. For all these reasons many researchers place RNA very close to the origin of life (assuming an early “RNA world” [35]).

Both primary and secondary structure can conveniently be displayed as a string over some alphabet. For primary structure, the alphabet consists of the set of names of the constituent monomers or their abbreviations (**A**denosine, **U**ridine, **C**ytidine, **G**uanosine in RNA). Because secondary structure in proteins is defined on a per-residue basis, it, too, can be described as a string of abbreviations (of the structural states). Secondary structure in RNA, with which we are mainly concerned in this work, is more complicated because it includes long range interactions. A proper RNA secondary structure however has properties which nevertheless allow for a succinct string representation: a position is not only either unpaired or paired with exactly one partner, but in addition it holds that for two positions p_1, p_2 with $p_1 < p_2$ the pairing partner of p_2 must occur to the left of the partner of p_1 . The pairs thus behave like the parentheses in a correctly bracketed expression. Accordingly, an RNA secondary structure is suitably displayed as a string over the

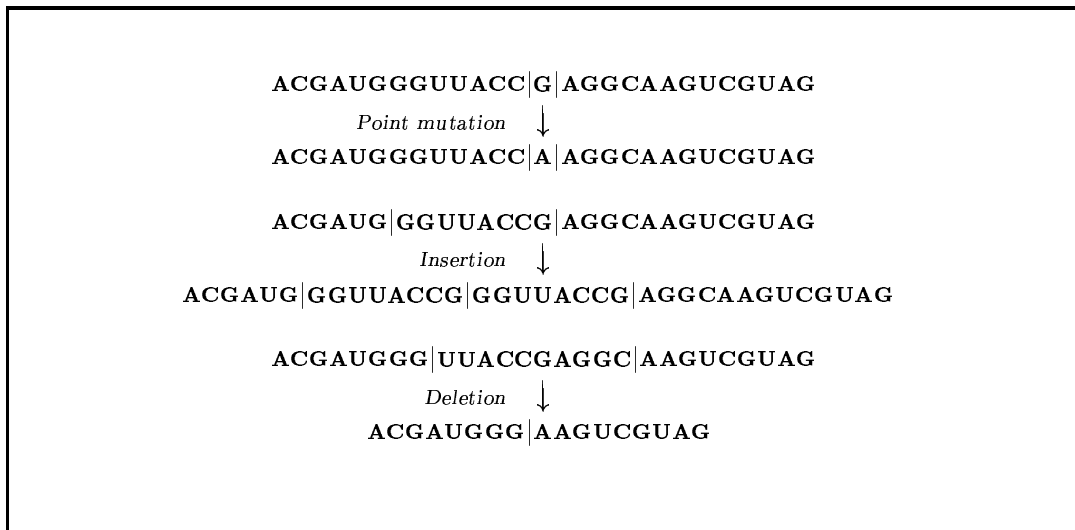


Figure 2: Three classes of mutations. Point mutations are copying errors with single base exchanges; they leave the chain lengths constant. In case of insertions part of the template sequence is duplicated during replication. A deletion leads to an error copy which is shorter than the original.

alphabet $\{(,),.\}$, with a dot indicating that the position is unpaired, and a pair of matching parentheses corresponding to a base pair.

2.1.2. Mutations

A given cell or virus inherits its genetically active nucleic acid molecules from other cells or viruses by way of *replication* (making a complementary copy of the chain). Several types of errors may occur in this process. By far the most common one is substitution or point mutation, which replaces a nucleotide by a different one. Neither the chain length nor the relative order of the elements of the chain is affected by this type of mutation. Mutational events which lead to a reordering of the positions of the chain call for more sophisticated concepts of neighbourhood in an abstract space of sequences. *Insertions* and *deletions* of one or more nucleotides, which may either occur by slippage of the copying enzyme or by imprecise recombinational events, are generally taken into account when comparing biological sequences. Frameshift mutations are less likely to occur in the first place (in Taq polymerase, their rate is 2.4×10^{-5} per site, as compared to 1.1×10^{-4} for base substitutions). Because of their nonlocal effect on a genetic message they will, even if they

occur, hardly be accepted in a coding sequence. This latter constraint does not apply to functional nucleic acids (RNA). Insertion and deletion mutants have been observed in ribosomal RNA[17].

2.2. The Main Mathematical Tool: Graphs

The concept of a graph is the basic mathematical tool to describe pairwise relationships between the elements of a set. Formally it is defined as follows:

Definition 1. (Graph). *A graph G is a pair $(v[G], e[G])$ and a map $o \otimes t: e[G] \rightarrow v[G] \times v[G]$. An element $V \in G$ is called a vertex of G with $v[G]$ as vertex set. Analogously $\epsilon \in G$ is called an edge and $e[G]$ is the edge set. The vertices $o(\epsilon)$ and $t(\epsilon)$ are called the origin and terminus of ϵ respectively. $o(\epsilon)$ and $t(\epsilon)$ are the extremities of an edge ϵ . If from $(v, v') \in e[G]$ it follows that $(v', v) \in e[G]$ then the graph is called undirected, otherwise it is called directed.*

Definition 2. (Regular graph). *A graph is called regular if for each node v the number of nodes v' for which $(v, v') \in e[G]$ is equal to a constant.*

Although any binary relation can be succinctly expressed as a graph, edges often have the meaning of a “reachability” of one or the other sort – be it spatial closeness of bases in a secondary structure graph, or the fact that origin and terminus of the edge can be interconverted by a unit operation of some process. By chaining together more than a single such unit step nodes can be reached from a starting node which are not directly linked to it by an edge: if this is the case, there is a *path* between the nodes.

Definition 3. (Path). *A path in G is a sequence $(v_1, \epsilon_1, v_2, \epsilon_2, \dots, v_n, \epsilon_n, v_{n+1})$, where $v_i \in G, \epsilon_i \in G, o(\epsilon_i) = v_i$ and $t(\epsilon_i) = v_{i+1}$.*

Definition 4. (Connected vertices). *Two vertices $v, v' \in G$ are called connected if there exists a path in G in which both vertices occur.*

The minimal length of a path connecting two vertices is a natural distance measure on the nodes (also called the *canonical metric* on the graph [9]). If there is no path between two nodes, the distance is set to infinity, which preserves the triangle inequality.

Connected nodes are in some sense equivalent (if one “sits” on one of them it is nevertheless clear how the other one can be reached). In a maximal equivalence class every node is connected with every other node inside the class but with no outside node. Such a class is called a *component* of the graph. If there is only one component, the graph is called *connected*.

Definition 5. (Connected graph). *A graph G is connected if any two vertices $V, V' \in v[G]$ are connected.*

Definition 6. (Component). *A component is a maximal connected subgraph of a graph.*

In later sections we will deal with scenarios in which the nodes of a graph are sampled with some probability. One outcome of such an experiment yields some subset of $v[G]$. The structure of the original graph may be transferred to this subset in the following way:

Definition 7. (Induced subgraph). *A subgraph H is called an induced subgraph of a graph G if, for any $v, v' \in H$ being extremities of an edge $\epsilon \in G$, it follows that $\epsilon \in H$.*

That is, two nodes in the subset are joined by an edge if this edge is in $e[G]$.

There is a special term for subgraphs which cover “nearly all” nodes of the underlying graph:

Definition 8. (Boundary). *Let H be a finite graph. The boundary $\partial_H v[G]$ of a subgraph $G < H$ consists of those vertices $v \in H \setminus G$ for which there exists $(v, v') \in e[H]$ such that $v' \in G$.*

Definition 9. (Dense graph). *Let H be a finite graph. A subgraph $G < H$ is called dense in H if and only if the closure $\overline{v[G]}$ of G , $v[G] \cup \partial_H v[G]$ is equal to $v[H]$.*

If an induced subgraph is dense, then there is no node in the underlying graph which not either is part of the subgraph or is joined by an edge to a node in the subgraph.

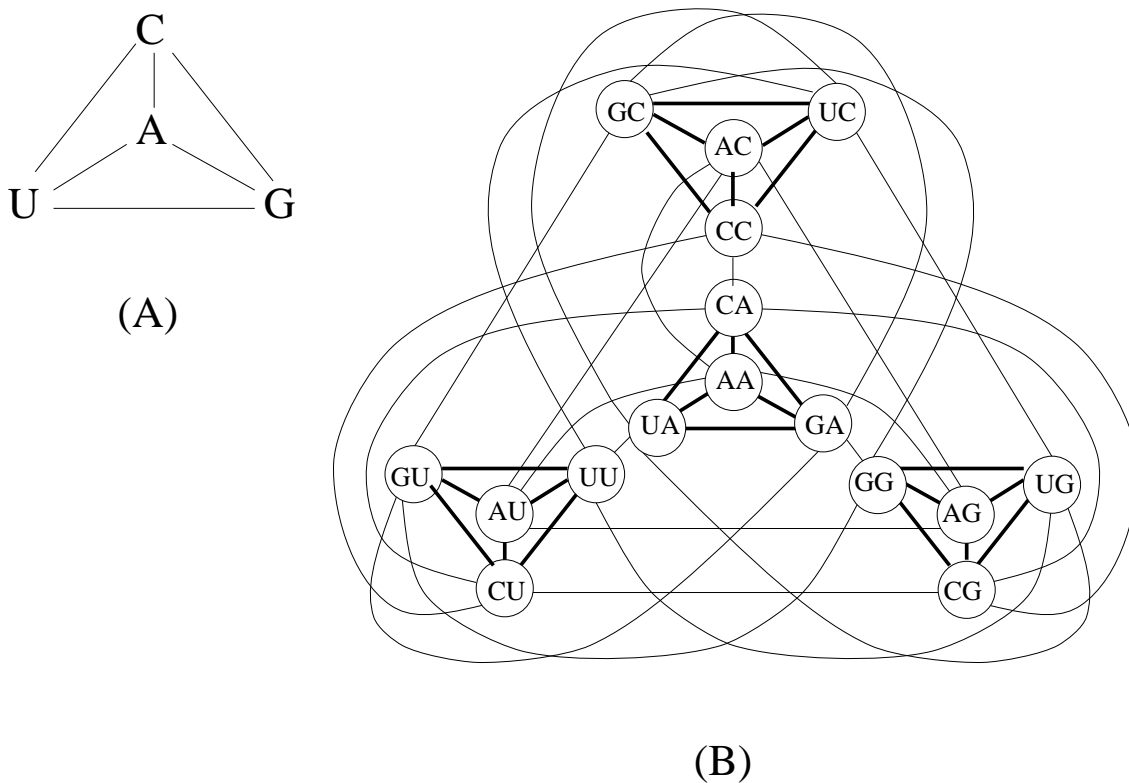


Figure 3: Hamming graphs of dimension 1 (A) and 2 (B). The graph of dimension n consists of $|\Sigma|$ copies of that of dimension $n - 1$ (Σ being the alphabet). The figure illustrates the strong interconnectedness of the graph already at small dimensions.

2.3. Fixed Sequence Length: Hamming Graphs

A Hamming graph [41] (also called generalized hypercube or sequence space [20]) is defined for a particular string length and a particular alphabet Σ . Its node set consists of all $|\Sigma|^n$ possible strings (sequences).

Two sequences are connected by an edge if they differ by a single symbol (base) exchange (see figure 3). Biologically speaking, point mutation is assumed to be the only mutational operator. The number of substitutions by which two sequences differ is a metric distance measure (Hamming distance). It is the canonical metric on the Hamming graph.

A *generalized Hamming graph* [87] is the direct product of several sequence spaces, each of which has its own sequence length and alphabet. By means of this extension the concept of Hamming space can be carried over to the case of strings with position specific symbol sets: each subspace corresponds to those positions which accept symbols from the same set. We will see below that in the case of RNA sequences, it is natural to distinguish between unpaired positions (at which bases can be exchanged independently of the remaining sequence) and paired positions, which are so highly correlated that both partners are effectively substituted in a single step. The distance between two vertices of a generalized Hamming graph is the sum of the Hamming distances within each of the subspaces.

2.4. When the Sequence Length Is not Fixed

2.4.1. Structure of the Underlying Space

For a given base alphabet there are infinitely many possible Hamming spaces, one for each sequence length n , $n = 1, \infty$. These graphs can be combined into a single graph (of infinite size) by introducing insertion and deletion of single bases as additional mutational operators.

Definition 10. (Levenshtein Graph). *The Levenshtein graph is that graph on the infinite set of strings over a given alphabet in which two strings are neighbours if they either differ by a symbol exchange, or by the insertion of a single symbol, or by a single symbol deletion.*

The graph, named after V.I. Levenshtein [58], is made up of “layers”, each corresponding to a sequence space of fixed length. It is obviously not regular: each Q_α^n has its characteristic outdegree, equal to $(\alpha - 1)n$. But there is a second source of heterogeneity, pertaining to the edges which connect layers: the number of deletion neighbours depends on the sequence. For each run

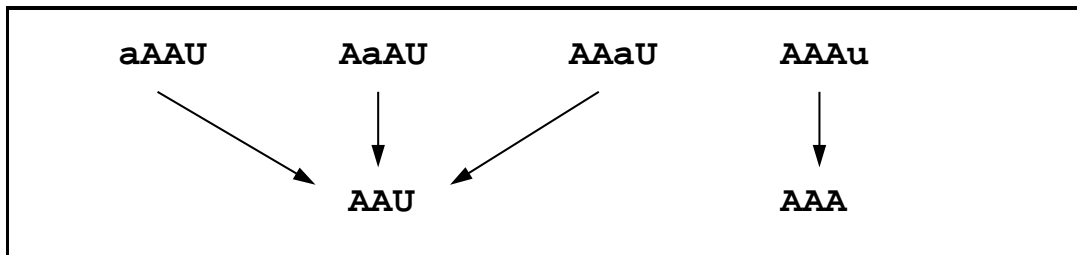


Figure 4: 1-deletion neighbours of the sequence “AAAU”. A lowercase letter indicates that this position is to be deleted. There are as many deletion neighbours as there are *runs*, this is, contiguous stretches of sequence which are made up of one symbol only. The example contains two such runs, namely, “AAA” and “U”, so it has two deletion neighbours. In the general case a sequence has at least one deletion neighbour (if the whole sequence consists of one long run) and at most *length* such neighbours (if any two adjacent symbols are different).

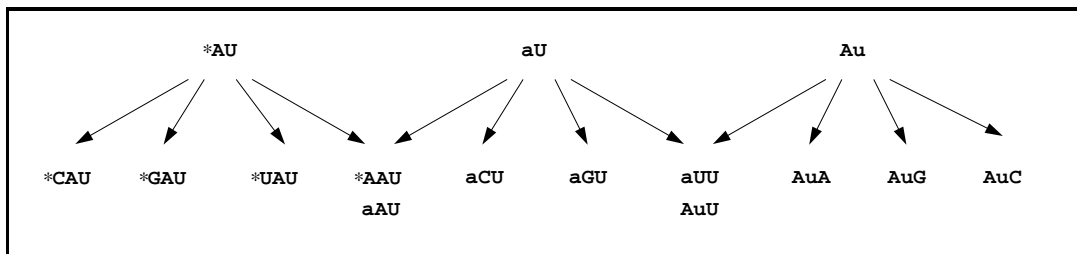


Figure 5: 1-insertion neighbours of the sequence “AU”. There are $length + 1$ places to insert a symbol: after positions $1 \dots length$, and before position 1. Insertion before position 1 can be regarded as insertion *after* a hypothetical position 0, which is marked as an asterisk in the figure. The position after which to insert is shown in lowercase, except if it is position 0. In contrast to the case with deletion neighbours, all sequences of a given length have the same number of 1-insertion neighbours. This number is equal to $nsymbols \times (length + 1) - length$: Any pair of adjacent positions (of which there are *length* pairs, including position zero) generates one redundant neighbour (the redundant neighbour is the one in which the symbol which occupies position $i + 1$ is inserted after position i . This leads to a run of length two. The same run is generated if this symbol is inserted after position $i + 1$).

(a substring which is made up of one type of symbol only) there is but one distinguishable deletion neighbour, irrespective of the length of the run. In contrast, the number of insertion neighbours depends only on n , although on first glance there is also an effect related to indistinguishable runs. Assume position i is occupied by symbol s_i , and position $i + 1$ by symbol s_{i+1} . Inserting symbol s_{i+1} after position i and after position $i + 1$ then results in the same string. For every pair of neighbouring positions, irrespective of the symbol assignments, there is however exactly one such

redundant neighbour, all other insertion neighbours being distinguishable (see Fig. 5). Therefore the number of different insertion neighbours of a string of length n over an alphabet of size α is $\alpha(n + 1) - n$. Insertion of a symbol before position 1 is modelled as insertion after a hypothetical position 0, so that there are $n + 1$ positions in total, including n pairs of neighbouring positions. Then from the $\alpha(n + 1)$ ways to insert a symbol after a position, n are redundant, and the formula follows.

2.4.2. Distance Measures For Variable Sequence Length

The canonical metric on the Levenshtein graph is the so called *modified Levenshtein distance*, which is equal to the minimum number of substitutions, insertions, and deletions needed to transform two strings into each other (plain *Levenshtein distance* amounts to the minimal number of insertions and deletions needed for a transformation). The modified Levenshtein distance is equivalent to the Sellers metric, which in turn, with an appropriate choice of parameters, is equivalent to the negative of the Needleman-Wunsch similarity measure [85]. The latter two measures are routinely applied when comparing biological sequences of unequal length.

2.5. Landscapes

A *landscape* consists of a set of configurations (e.g all strings of a given length over a given alphabet) together with two mappings: first, an equivalence relation indicating which pairs of configurations are neighbours. For biosequences, those strings are neighbours which can be interconverted by one of the mutational operators. Second, a mapping from the domain of configurations into the domain of real numbers, which associates some *fitness* with each configuration.

The term “landscape” goes back to Sewall Wright [103, 104]. The idea is that in searching for

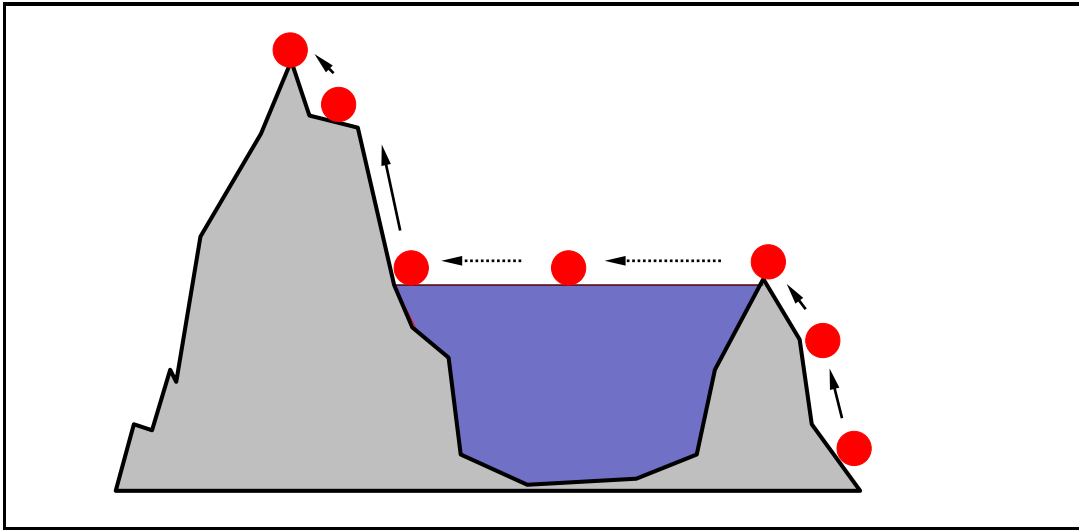


Figure 6: Evolution as hill climbing in an adaptive landscape. The role of neutrality.

the fittest configuration, an evolving system encounters problems which are similar to those of a wanderer seeking the highest point in an unknown natural landscape. Without further information, he will probably try the *steepest ascent* among all available directions. Whether or not this will lead him to his destination depends on features of the landscape as well as on his own ability to perceive and evaluate these features. In a *smooth* landscape, which possesses only one or a few maxima, he will most likely be successful even if he cannot do better than recording the relative heights of the most nearby places and choosing the one which is highest to go next. The more *rugged* the landscape is, this is, the more maxima it contains, the less promising is the “myopic” approach. It will sooner or later lead to a suboptimal maximum (see figure 6). One way to escape from such a trap is by using information from other points besides the immediate surroundings (a human wanderer could for example detect a distant height with a field glass). If this is not an option, it might still be possible to find the global maximum given one prerequisite is fulfilled: there must not be a suboptimal maximum which, in its immediate neighbourhood, does not contain a point which is at the same height (this is, maxima must not be single points but rather *ridges*). The

wanderer would then, if he cannot gain height on any of the currently nearest points, randomly follow a direction which does not descent. Travelling along the ridge, he might encounter ascending directions, which he then will choose. In order for him to finally be successful, the global optimum must either directly abut on the current ridge, or else there must be a series of ridges $r_1, r_2 \dots r_n$, such that r_1 adjoins to the current ridge, r_{i+1} adjoins to r_i , and from r_n the global optimum can be reached. It depends on the landscape whether or not such a series exists.

The human wanderer is a special case of the abstract task to find the best configuration while evaluating only nearest neighbours. “Nearest neighbours” are those configurations which, in the underlying configuration space, share an edge with the current configuration. The best configuration is the one at which the fitness function reaches its maximum value. In both natural and algorithmic applications of gradient searches in a configuration space, the search is normally not performed by a single agent but by a whole population of current configurations. The same configuration may occur several times. The agents do not literally move, but rather undergo Darwinian evolution: they make erroneous copies of themselves which are differentially retained in the population according to the values the fitness function takes on them. (Normally, fitness determines the *probability of survival* rather than survival itself. If this is the case, then the individual agents can accept descending steps, although with low probability. The width of the valley to be crossed, much more than its depth, determines whether or not this can contribute to escapes from local optima [91].) It depends on the mutational operators which mutant configurations can emerge as offspring of a single replication. Over time this process will make the center of gravity of the population move in the direction of fitter configurations. The search task is parallelized in such a scenario and thereby speeded up. Nevertheless the problem of local optima remains, as well as the importance of fitness neutral paths which may connect different optima. In a configuration space, the induced subgraph of those configurations which have the same fitness is called the *neutral network* of this fitness value. The neutral network of the currently highest fitness in a population

is of special importance: except for a possible constraint on the total population size, the whole network may be freely populated. Dependent on the graph structure of the network, this can lead to a greatly enhanced diversity of the current configurations, *which is not counterselected*. More diversity means a higher parallelization of the search for an even fitter neighbouring configuration. The process of unselected spread over a neutral network is called *drift*.

2.6. Neutral Networks of RNA Secondary Structures

2.6.1. Landscapes Based on Sequence Structure Mapping

One way to define a landscape over the elements of a sequence space is to equate some aspect of their structure with fitness. Such aspects may be the exact secondary or tertiary structure, the coarse grained secondary or tertiary structure, or the presence of some secondary or tertiary structural motif somewhere in the sequence.

Although structure is most relevant in proteins, it is not currently feasible to reliably “fold” a protein sequence *in silico*. The minimum free energy secondary (not tertiary) structures of RNA molecules, in contrast, can be quite reliably predicted, at least for short chains [45]. Indeed it is RNA for which a landscape based on the sequence structure mapping makes most sense: RNA can be genetically active, so that the configurations (sequences) directly undergo and inherit mutations, as required in the landscape model. This is not true for proteins. At the same time, the structure of an RNA molecule can be important for its function, and thus has impact on its fitness.

In the following, we will consider the sequence secondary structure mapping in RNA. This is an extremely redundant mapping: there are $n \text{ symbols}^n$ sequences, which are mapped to approximately $1.4848 \times n^{-\frac{3}{2}} (1.8488)^n$ structures [79, 47] For a fitness function which measures similarity to some

reference secondary structure, one therefore expects a high degree of neutrality. This will further increase if only the presence of a certain substructure element is required, rather than an exact match of the secondary structural states of all positions to a target structure.

2.6.2. The Underlying Configuration Space

2.6.2.1. Fixed Sequence Length

The necessary condition for an RNA sequence to adopt some given structure is that it is *compatible* with the structure, that is, that it contains two bases capable of forming a pair at any two positions that are paired in the target structure. With this in mind, there are two different ways to define the configuration space underlying the folding landscape.

The straightforward approach borrows the graph structure from the Hamming graph $\mathcal{Q}_{A,U,G,C}^n$ by defining the space of compatibles to be the respective induced subgraph. This assumes that sequences are interconverted by successive point mutations. Whether or not this graph is connected depends on the alphabet and the pairing rules: assume there are p groups of valid base pairs, such that any two pairs in the same group can be interconverted by point mutations (like $AU \leftrightarrow GU$). The graph of compatibles decomposes into p^{n_p} components, where n_p is equal to the number of base pairs in the reference structure. Because $p > 1$ in all alphabets based on the natural pairing rules (be it $\{G, C\}$, $\{A, U\}$, $\{A, U, G, C\}$, the latter with or without GU pairs permitted), this approach generally leads to a disconnected space of compatibles. If the number of point neighbours is different for different base pairs, it is in addition not regular. Lack of regularity and connectivity is mainly a technical problem: were these conditions fulfilled, random graph theory would permit to estimate global properties of a neutral network (which itself is an induced subgraph in the space of compatibles).

If therefore the focus is on the topology of a single network, rather than on the entire landscape, it

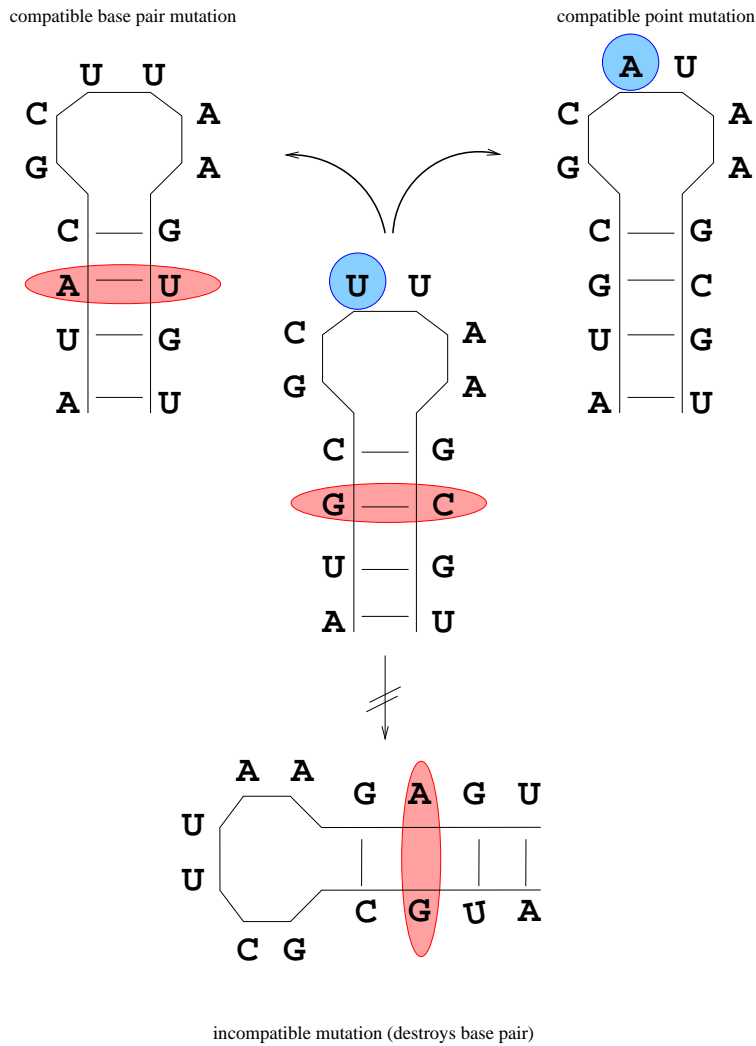


Figure 7: A single mutation in a base pair leads to a sequence which cannot fold into the secondary structure of the original sequence. If the base pair is exchanged as a whole, or else the point mutation hits an unpaired position, then the mutant remains *compatible* with the original structure (all required base pairs can be formed). Depending on the conditions the actual structure of the mutant may nevertheless be different.

is conventional to represent the compatible sequences as a generalized Hamming graph of the form

$\mathcal{Q}_{A,U,G,C}^{n_u} \times \mathcal{Q}_{AU,UA,UG,GC,GU,CG}^{n_p}$, where n_u is the number of unpaired positions in the structure

and n_p is the number of paired ones. The two subspaces are called the *unpaired fiber* and the *paired*

fiber, respectively. By treating a pair of paired positions as one “position”, which is occupied by one of the valid base pairs, it is ensured that all neighbours of a compatible sequence are again compatible. Furthermore, the dependency of the probability to be neutral of a mutation in a base pair upon the pairing partner is removed by this construction. A strong objection which can be raised against this approach is the fact that “base pair mutations” do not exist in nature. Yet, paired positions indeed tend to mutate in a correlated manner. This is due to the fact that any primary mutation is likely to create a mismatch, in which case any compensatory mutation which restores the correct pairing will have a very strong positive effect, and will be selected for [43].

2.6.2.2. Variable Sequence Length

From artificial selection experiments [53] it is known that function is very often due to the presence of some substructure somewhere in the sequence, rather than dependent on the exact structural state of all positions [90, 76, 75, 74, 64]. Thus a constant sequence length is not required.

So far there is no established approach to modelling the space of compatible sequences of a substructure. As with full length structures, one can either ignore the correlations due to the base pairing (the space of compatibles is a subgraph of the Levenshtein graph in this case) or resolve the correlations by mutating pairs of paired positions in a single step. With full length structures, a strong case could be made for the latter alternative, because in contrast to the former, it produced a connected regular graph. The case is less clear with substructures. First, it is at least theoretically possible (in practice depending on the substructure under consideration and the additional sequence length) that the subgraph induced in the Levenshtein graph by the set of compatibles is indeed connected (even if there are base pairs in the substructure !). It is not regular, but neither is a graph which incorporates base pair mutations (see below). It is much less straightforward to include such mutations than it is with full length structures. The fact has to be taken into account that a subsequence which is compatible with the substructure may occur at any

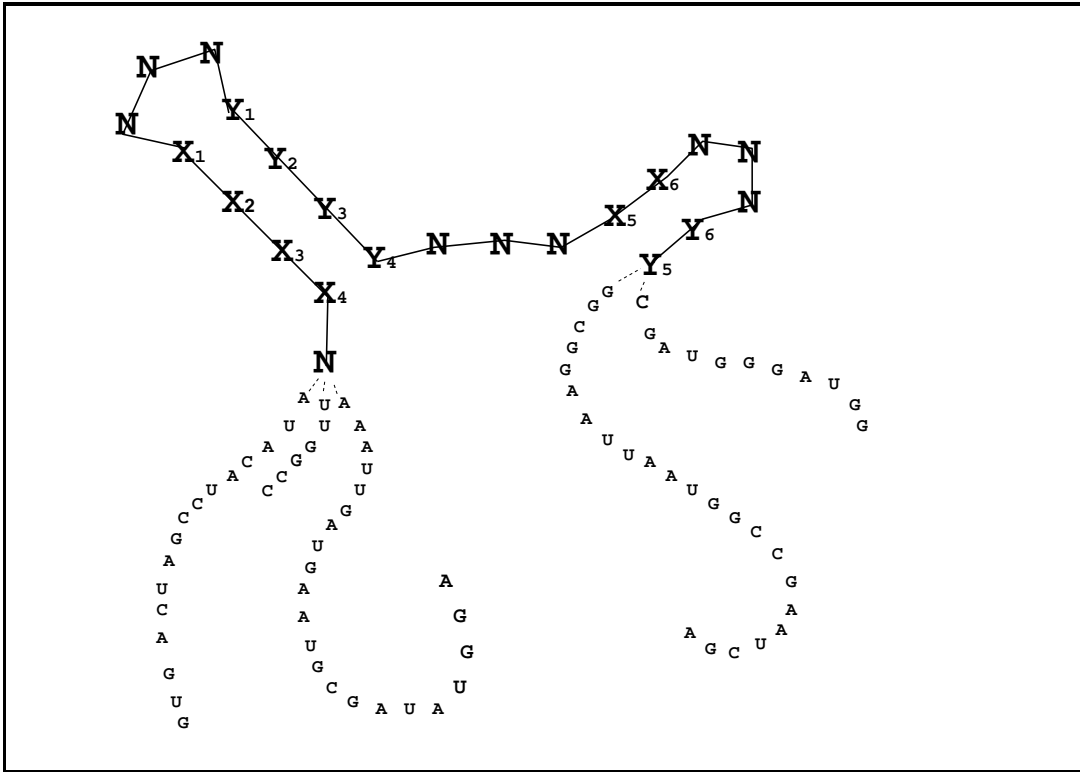


Figure 8: *Neutrality with respect to a substructure.* If fitness hinges on the presence of a certain substructure, a lot more variability is introduced into the neutral set. The amount and nature of additional sequence to the left and to the right of the substructure is completely free. In the region of the substructure itself, permissible sequences are defined by the compatibility rules.

relative offset to position one, so that it is not possible any more to talk about *the* unique pairing partner of a position in an elongated sequence. We show in chapter 5 how this difficulty can be overcome. In this construction pairs of positions are mutated as a whole if both partners belong to a compatible subsequence. If there is more than a single compatible subsequence in a given full length sequence, it may happen that one and the same position is paired with different partners in each of them. Consequently it has as many pair neighbours as there are different partners, and the resulting graph, although connected, is generally not regular.

The whole issue of neutrality with respect to a substructure will be discussed much more in-depth in chapter 5.

2.6.3. Generic Properties of Neutral Networks of RNA Secondary Structures

A given RNA sequence usually is compatible with many different secondary structures. Its minimum free energy structure by definition has to be one of them, the others being the suboptimal structures of the sequence (which sometimes can be triggered by suitable environmental conditions to be adopted instead of the minimum free energy structure). Thus folding into the actual structure may be viewed as *choosing* from a list of candidate structures. It is known that different algorithms for RNA secondary structure do not always yield the same result on a given sequence, yet nevertheless global statistical features of RNA folding landscapes are largely independent of the folding algorithm [88, 89]. This can be understood by the following argument. Varying predictions mean that we do not fully know how the actual structure is chosen from the candidates. The best we can do is assign some *probability* to be chosen to each of the candidates. Simple ways to do this is either assume there is an uniform probability for all possible structures, or else there is a fixed probability λ for a given candidate structure s of interest, all remaining structures being lumped into the event “not s ” of probability $1 - \lambda$. Such a probabilistic view is not only a way to deal with lack of accurate structure predictions. At the same time it provides an algorithm independent *model* of how certain features of RNA folding landscapes come about, and it is these features which indeed are robust across different folding algorithms.

The main generic features are [12]

- (i) There are few frequent and many rare structures. Almost all sequences fold into frequent or “common” structures.
- (ii) Sequences which fold into “common” structures are distributed nearly uniformly in sequence space.
- (iii) Almost all “common” structures can be found close to any point in sequence space. This property is called *shape space covering*.
- (iv) A sequence folding into a “common” structure has a large number of neutral neighbours

(folding into the same structure) and a large number of neighbouring sequences folding into very different structures.

- (v) Neutral paths percolate sequence space along which all sequences fold into the same structure. In fact, there are extended *neutral networks* of sequences folding into the same “common” structure.

2.6.3.1. Zipf’s law(s)

From massive statistical investigations of RNA secondary structures (e.g [89]) as well as from exhaustive folding of sequence spaces over binary alphabets [37, 38] it is known that structures behave like words in a natural language : there are a few common ones and many rare ones. A relationship of this type is known as Zipf’s law after [107]. In its simplest form it states that the frequency of a word times its rank (in an ordering according to decreasing frequency) is equal to a constant, that is, (frequency, rank) pairs will come to lie on a line with slope -1 in a log-log plot.

Zipf1
$$f(r) = \frac{const}{r}$$

There are several generalized versions of the law. One which is highly tunable takes the form

Zipf2
$$f(r) = A \times \left(\frac{B}{B+r}\right)^\gamma.$$

The scaling parameter A is the count of the most frequent word. In a double-log plot, the function is very slowly decreasing up to about rank B , and then gradually changes into a line of slope $-\gamma$. Words with ranks less or equal than B , the abundance of which is more or less independent of their rank, are called common words. Mandelbrot [61] showed that *Zipf2* holds for the word counts in a text which is generated by randomly stringing together the symbols of some alphabet and a space character, followed by splitting into words at the space characters.¹

¹In his original paper George Kingsley Zipf gave a sophisticated explanation for the appearance of the law in natural languages: he argued that a speaker has to invest a minimal effort for conveying a given message if there are a few words which make up the scaffold of *any* message (such as articles and pronouns). Those then can be stored in a “fast access” region of the brain. The specific contents of the message is conveyed by only a small fraction of it, consisting of relatively rare words, which typically are also longer than the high frequency words. This explanation started an argument with Mandelbrot, who advocated the position that the law will necessarily hold for any random text.

Zipf2 is rewritten as

$$\text{Zipf2}' \quad f(r) = A \times \left(1 + \frac{r}{B}\right)^{-\gamma}.$$

In general, the law describes a scenario in which the combinatorial complexity of the most probable events is low (it is likely that a space occurs “relatively soon”, resulting in a short word, yet there are only a few different short words) while the one of the unlikely events is high (there are infinitely many words which are “longer than short words”, and if the alphabet size exceeds 1, there are more different words of a given length n if n grows). In terms of RNA minimum free energy secondary structures, two factors determine whether a structure is “probable”: first, because by definition the free energy is minimized, stable structures are the likely outcome of folding. Second, a structure can only be realized on a sequence which is compatible with it, so that the most “probable” structures are those which both have a large set of compatibles and are comparably stable. With increasing chain length, the ratio of the number of these preferred structures to the number of all structures becomes increasingly small, resulting in a Zipf distribution of the frequencies of realized structures. Another example of a Zipf-like relationship in the context of RNA structure landscapes is the distribution of *intersection* sizes (see section 3.3.3).

2.6.3.2. A Random Graph Model for Neutral Networks

Assume we model folding into a reference structure as choosing the structure from the candidate list of the sequence, with the above mentioned probability λ . Assume further λ is constant for all compatible sequences of the reference structure. (This is certainly never true in detail, be it only because the candidate lists are of different size for different sequences. However the model does not aim at a detailed prediction of whether or not a single sequence folds, but rather at a prediction of global properties of the network. *If* it holds that there is no clustering of sequences with an exceedingly high or low true probability of folding, then the distribution in sequence space and the expected number of selected sequences (although not the exact identity of the sequences)

will be the same if the choosing is done with the average value of the true λ s (which is constant), or with the per-sequence λ s themselves. Thus in this case it is admissible to assume a constant λ .) We have seen above that for fixed length structures the set of compatible sequences is best organized into a (connected and regular) generalized Hamming graph of the form $\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$, n_u and n_p being the number of unpaired and paired bases in the sequences, and α and β the alphabets of bases and valid base pairs, respectively. Those sequences which actually fold induce a subgraph in the generalized Hamming graph – by inheriting the graph structure, the neutral set becomes a neutral “network”.

The compatible sequences which are selected in a single realization of choosing with probability λ induce a subgraph in the generalized Hamming graph, too. For the reasons mentioned above it is not expected to be point by point identical with the “true” neutral network of the reference structure. The true network is however one out of all possible realizations, and if the assumptions about independence of individual folding probabilities are met, it is a typical one. The important point is now that there is a whole body of theory (percolation, random graphs) which, for a random process which selects nodes of a regular graph with some constant probability, predicts *average* global properties of the resultant induced subgraphs. Therefore *if* the model is applicable, it is possible to estimate properties like connectivity or denseness of a neutral network from the average folding probability alone, without a need to fold the individual sequences.

The following model has been used bei Christian Reidys [71]:

By the structure of the generalized Hamming graph, a compatible sequence is split into two objects: an unpaired part v_u which is in $\mathcal{Q}_\alpha^{n_u}$, and a paired part v_p in $\mathcal{Q}_\beta^{n_p}$. There is a distinct choosing probability for each of these parts: it is called λ_u for v_u and λ_p for v_p . One instance of the random process is generated by first selecting sets $\sigma_u \subset v[\mathcal{Q}_\alpha^{n_u}]$ and $\sigma_p \subset v[\mathcal{Q}_\beta^{n_p}]$ using the appropriate choosing probabilities, and then defining the set of selected compatibles to be equal to the cartesian

product $\sigma_u \times \sigma_p$. Christian Reidys showed that there are two threshold probabilities

$$\lambda_u^* = 1 - |\alpha|^{-1} \sqrt{1/|\alpha|}$$

and

$$\lambda_p^* = 1 - |\beta|^{-1} \sqrt{1/|\beta|}$$

such that if both $\lambda_u > \lambda_u^*$ and $\lambda_p > \lambda_p^*$ it holds for the induced subgraph Γ_n in $\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$:

$$\lim_{n \rightarrow \infty} \{\Gamma_n \text{ is dense and connected}\} = 1.$$

Note that for short chain lengths n the relation is not *assured* to hold. Exhaustive enumerations of minimum free energy neutral networks of binary sequences [37, 38] have however shown that the threshold probabilities often do lead to useful predictions for $n \leq 30$. In chapter 2 we will investigate the case of a sequence space of length $n = 16$ over the natural alphabet $\{A, U, G, C\}$.

3 Sequence Space and Shape Space of Q_{AUGC}^{16}

3.1. Introduction

The random graph model introduced in the preceding chapter will, if the appropriate preconditions are met, predict global properties of neutral networks. In cases in which it is impossible to access more detailed information (because the chain length is too long to determine the minimum free energy secondary structure of all sequences in reasonable time, or because neutrality is defined with respect to a quality which cannot be reliably predicted from the sequences) this provides a most valuable hint. One thing which it obviously cannot do is to determine which neutral networks exist in a landscape in the first place. In addition it needs an average choosing probability as input, which means that more than a single member of the network needs to be found beforehand in order to provide a reliable estimate. For RNA secondary structures it has been shown [31] that the networks of rare structures are very hard to find if the landscape is statistically sampled. Thus finding the rare networks is one reason for exhaustively determining the fitness values of an entire landscape, in case this is possible. Another reason is that the details of the graph topological structure of a network which by definition are not predicted by the model can well be of interest in their own right. Last not least, detailed information on a landscape may be used to test the model – this is especially important for short chain lengths for which the mathematical theory itself labels the predictions as unreliable.

In this chapter we provide data on the minimum free energy secondary structures of all RNA sequences of length 16 over the natural alphabet $\{A, U, G, C\}$. These are about 4 billions of sequences, which fold into as few as 274 different secondary structures. The landscape has one

outstanding peculiarity: 63% of all sequences “fold” into the open structure, i.e. they form no base pairs at all. This is because the sequences are too short to permit long stacks to form, and the entropy gain by not folding at all most of the time exceeds the energy gain associated with the formation of a short stack. The frequent instability of base pairs leads to a violation of the assumptions of the random graph model: oftentimes the formation of a pair depends on the sequence context. Because of the abundance of the open structure and the frequent confinement of the remaining networks to only a few dimensions of the entire space (which provide a permissible context) the networks of Q_{AUGC}^{16} are less dense than predicted by the model, but more connected. Obviously a chain length of 16 does not suffice in order to apply the model.

3.2. Methods

3.2.1. Why Length 16 ?

There are two reasons for choosing the rather “unbiologically short” sequence length of 16. First, it is the maximal length that allows a sequence to be mapped onto a unique value of a basic type of the C programming language. By coding a given base by the configuration of two adjacent bits, a sequence of length 16 can be coded for by the value of an unsigned long integer (32 bits). Second, and more importantly, a sequence length of 16 takes us to the limits of affordable computation time and storage resources. The time needed for exhaustive folding of the entire space is dramatically increasing with sequence length : it is 16 hours for Q_{AUGC}^{10} , 8 days for Q_{AUGC}^{12} and about one and a half year for Q_{AUGC}^{16} (on an Indigo2 XL). So for lengths greater than 16 it is expected to reach the order of decades with current computer technology. During the folding of Q_{AUGC}^{16} 10 Gigabytes of data were produced.

One case can be made in favor of a rather short sequence length: The shorter the chain, the less likely it will be kinetically trapped in a structure which is far from its minimum free energy structure. This is, the computed minimum free energy secondary structures are likely not to be too far off from reality.

3.2.2. Algorithm for Determination of the Sequence Of Components

To determine the sequence of components of an undirected graph is a standard problem in graph theory. One common solution is *breadth first traversal* of the nodes of the graph (cf. Fig. 9). The name refers to the fact that the algorithm processes all immediate neighbours of a node before any remaining nodes are processed.

The basic operation during a graph traversal consists of finding the neighbours of the current node. In a general graph setting, this requires adjacency information to be stored in some way. Data structures which do this have a space complexity of $O(|v[G]|^2)$, which is completely infeasible with networks which are of size 5×10^7 . Fortunately, neutral networks of RNA secondary structures have a property by which one can do without this information. Namely, we know that no two sequences can be neighbours in a neutral network unless they are neighbours in the underlying graph of compatibles. Because a sequence completely determines its set of compatible neighbours there is no need to store them: at the price of a little more time they can be constructed from the sequence if needed. In order to determine the actual neighbours in the neutral network, each of the candidate neighbours then has to be looked up in the set of nodes of the network.

For the algorithm in the above form to work, all nodes of the networks have to be accommodated on a random access device, either in working memory or in a database system on disk. Alternatively, one can modify the algorithm so that the network is processed segment by segment, which is what we have done.

The algorithm proceeds in two steps. In the first step, one segment at a time is read into working

list	nodes not yet assigned	comp.no.	member
—	{ 1, 2, 3, 4, 5, 6, 7 }	0	
1 —	{ 2, 3, 4, 5, 6, 7 }	1	
3 —	{ 2, 4, 5, 6, 7 }	1	1
4 — 6 —	{ 2, 5, 7 }	1	3
6 — 5 —	{ 2, 7 }	1	4
5 —	{ 2, 7 }	1	6
—	{ 2, 7 }	1	5
2 —	{ 7 }	2	
7 —	{ }	2	2
—	{ }	2	7

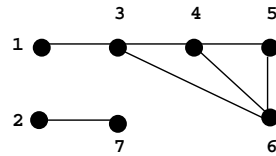


Figure 9: Breadth First Traversal of an undirected graph. At every step, there is a current component number, a set of nodes not yet assigned to any component, and a set of nodes which are known to belong to the current component but are not fully processed yet. In the beginning, the second set, which is represented as a linked list, is initialized to nil, and the set of unassigned nodes is equal to all nodes of the graph. The initial component number is zero. Whenever the list of nodes known to belong to the current component is empty (as it is at the start) one of two actions is taken: if the set of unassigned nodes is empty, too, the algorithm stops. Else an arbitrary node out of the unassigned nodes is chosen, removed from the set, and entered into the list. The component number is incremented by one. Then the following steps are repeated until the list is empty again: the head of the list is removed and reported as the next member of the current component. Those neighbours of this node which are unassigned yet are removed from the set of unassigned nodes and appended to the tail of the list.

memory and a standard component decomposition is done on the subgraph it induces in the graph of compatibles. At the end of the first stage the component structure of all subgraphs induced

by the segments is known. Now we observe that two such components which belong to different segments may be connected in the component structure of the complete network. Assume we can find out which is connected to which. The problem then can again be formulated as a graph, this time on the smaller node set of components of the segments – two components being joined by an edge if they are connected. Let us call the graph Γ . The second stage of the algorithm now consists of a component decomposition of Γ (in practice, it is implemented as a bottom up clustering instead of a graph traversal). Joining the sets of sequences which are in the same component of Γ then yields the component structure of the complete network.

The modified algorithm solves the memory problem. (It is still necessary that two segments at a time fit into working memory. This however can always be achieved by adjusting the segment size). Whether or not it is feasible in terms of time complexity depends on the time it takes to determine whether or not two segment components are connected. Let us call the two sets c_i and c_j , and let us assume $|c_i| \leq |c_j|$. Both sets are read into working memory and are stored in a data structure which permit lookups in $O(\log n)$ time (we use AVL trees [2] as implemented by Kendall Bennett and adapted by Walter Fontana). Then c_i is traversed, and for each element the set of candidate neighbours is constructed. Each candidate is looked up in c_j . As soon as an edge is found, we know that the two sets are connected and the algorithm stops. The worst case occurs if they are not connected, in which case c_i is completely traversed and the time complexity is $O(|c_i| \log(|c_j|))$ (c_i is chosen to be the smaller component in order to keep the number of candidate neighbours which have to be looked up in the worst case small).

Table 1 summarizes the algorithm.

The algorithm takes about 2 weeks on an Indigo2 XL for a connected network of size 31720954 (rank 11). MAXFILL was set to 1000000 in this run. The initial decomposition, taking only about 1 day, resulted in 93 partial components.

```

Read(MAXFILL) ;                               # max. number of sequences in working
                                              # memory
count := Size ;                               # total number of sequences
nsocs := 0 ;

while ( count > 0 )
  n := max(MAXFILL, count) ;
  ReadSeqs(n) ;                               # Store n sequences in an AVL tree

  count := count - n ;
  nsocs := nsocs - 1 ;

  SOC[nsocs] =                               # Find Sequence Of Components for
    BreadthFirstTraversal(n) ;               # sequences just read

end while ;

for i=1 ... nsocs
  for ii=1 ... ncomps[i]
    ReadComp(comp[ii]) ;

    for j=1 ... nsocs
      next if ( i == j ) ;

      for jj=1 ... ncomps[j]
        ReadComp(comp[jj]) ;

        # Is there a connection between
        # subgraphs comp[ii] and comp[jj] ?
        # If so, merge them in output
        # Sequence of Components
        if ( Connected(comp[ii], comp[jj]) )
          Join(comp[ii], comp[jj]);
        end if ;
      end for ;
    end for ;
  end for ;
end for ;

```

Table 1: Algorithm for finding the Sequence Of Components

3.2.3. How to Handle Random Access to Large Data Files

The problem of the entire networks not fitting into working memory recurs at every stage of analysis. Even if only a statistical sample of all sequences of the network is used, one cannot be sure to get a realistic distance distribution when just using a single consecutive part of a listing of the network.

A *B-tree* is a multiway search tree which is normally used to externally implement a dictionary data type (a set with the operations “insert”, “delete”, and “find” defined on it). A multiway search tree is a generalization of the concept of a binary search tree. A node normally holds more than one *key*, this is, data item. To the left and to the right of each data item there are links pointing to sub-multiway trees. Data items are sorted within a given node. A subtree which is pointed to by a link which is bounded by data items i and j is constrained to obey $i < k < j$ for each data item k on any of its nodes. The structure of the tree reduces search time by the reduced height.

In an external implementation of a dictionary, a node is equal to a page in memory.

We have used the Btree methods of Berkeley DB (version 2.4.14) [1] to organize the *Size* unsigned long integers representing the sequences of a network into a database on disk, which can be accessed from within a C program. A very serious drawback is the fact that the database file is about two orders of magnitude larger than the the plain data file (for a network of size 31738681 the database file is about 1.3 Gigabyte in size). For this reason only 220 out of 274 networks have been indexed.

3.3. Which Structures Are Realized in \mathcal{Q}_{AUGC}^{16} ?

The 4^{16} sequences in \mathcal{Q}_{AUGC}^{16} were folded into their minimum free energy secondary structures by the algorithm of Zuker and Stiegler [109, 108] as implemented by Ivo Hofacker et al. in the Vienna RNA Package version 1.03 [46, 45]. During the course of the folding the neutral sets of the landscape, each consisting of all sequences which fold into a particular structure, were built up. In a second step the sequence of components was determined for each of these sets by the algorithm given in table 1. As a first approach to the sequence structure mapping in \mathcal{Q}_{AUGC}^{16} we will in this section entirely ignore the fine grained composition of the neutral sets and just ask the question how many of them occur and what are the distinguishing features of their corresponding structures.

According to [79] the relation between chain length n and the number of different RNA secondary structures takes the form of $s(n) = 1.4848n^{-\frac{3}{2}}1.8488^n$. Thus one would expect $s(16) = 431$ different networks in \mathcal{Q}_{AUGC}^{16} . The number which is actually observed is much lower: it is only equal to 274. The basic element of the structures is a single stack of length 2 to 6. Assuming a minimal loop length of 3, there are $\sum_{l=3}^{16-2n} 16 - (2n + l) + 1$ different structures which consist of a single stack of length n (the variable l is the loop length). For stack lengths of 3 to 6, all of them are realized, which amounts to a total of 70 structures. There is an upper limit of 7 to the loop length in 2-stacks, which is why they contribute only 40 more structures (instead of 55). The remaining 164 structures consist of one of the simple stacks plus some additional base pairs. These pairs may form in the loop and/or in unpaired terminal stretches. Although the two possibilities are combinatorically equivalent, they are not from a thermodynamic point of view: a pair in the loop means less of an additional sterical constraint (and therefore reduction of entropy) than one which involves the terminal regions. That such an effect is in operation is clear from the fact that isolated additional base pairs are quite frequent in long loops, while they never occur outside. The more

stable the underlying simple stack, the less favored are additional pairs in the terminal regions (for example the derivative “(((((((...))))))”) of the stack “..(((((((...))))))...” does not occur). The effect surely contributes to the relatively low number of observed structures. Accordingly this number is likely to change if the energy parameters of the folding algorithm were changed. In fact this is the case with the above example: in version 1.31 of the Vienna RNA package, a sequence which adopts the structure “(((((((...))))))”) is readily found by inverse folding (e.g. GGGCCUGGCAGGCACC). When folded with version 1.03, the terminal base pairs are not formed. Indeed there was a change of energy parameters between the two versions (from the parameter set of Freier *et al.* (1986) [34] to that of Walter *et al.* (1994) [99]).

Although the number of structures in \mathcal{Q}_{AUGC}^{16} is lower than expected, it is high compared to the number of structures in \mathcal{Q}_{AU}^{16} and \mathcal{Q}_{GC}^{16} . There are only 4 structures in \mathcal{Q}_{AU}^{16} (the open structure plus the 3 possible simple stacks of length 6). \mathcal{Q}_{GC}^{16} consists of 195 structures. The 79 structures which are in \mathcal{Q}_{AUGC}^{16} but not in \mathcal{Q}_{GC}^{16} are featured by tetraloop hairpins with a stack of length 1 or 2 and/or isolated base pairs. They contain all but two of the 44 disconnected networks in \mathcal{Q}_{AUGC}^{16} . There are tetraloop structures in \mathcal{Q}_{GC}^{16} , albeit with longer stacks. The natural alphabet permits the formation of certain subsequences which are preferred in tetraloops [57] (these preferences are build into the folding algorithm), leading to loops which are inherently stable. In the binary alphabets, a corresponding stabilization has to be provided by the stack. Isolated base pairs occur more easily in the natural alphabet for combinatorial and energetical reasons: because of the reduced stickiness of the alphabet [28] it is more likely that a lonely pair is the only option for a position (besides not pairing at all). Especially if it is a GC pair, it may actually form, because GC pairs constitute an above average energy gain in this alphabet.

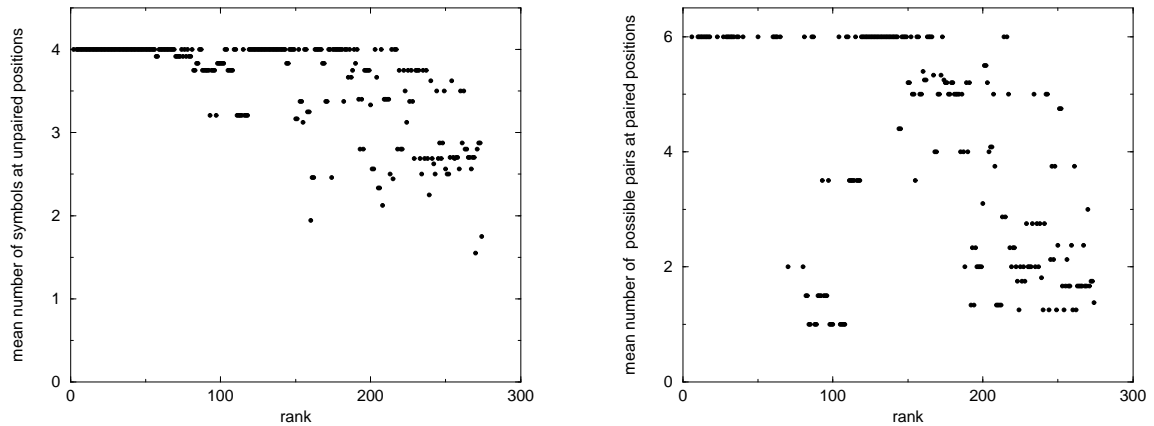


Figure 10: The mean number of bases at unpaired positions (left) and the number of base pairs which can be formed at paired positions, given the mean number of bases at the constituent positions (right). The ordering is according to absolute network size. Small (high ranking) networks tend to be those in which less than the maximal number of bases (pairs) is realized. This means that they are confined to certain dimensions in sequence space.

3.4. Global Properties of Networks

A neutral network, like any graph, is made up of a *set* of nodes together with a binary relation which defines the edges. The simplest characteristics of a network only assume the set property: for example the size of the network, and how this relates to the sizes of certain supersets, subsets, and intersections. Global averages of per-sequence properties are also independent of the graph structure. One such measure which, without making use of any distance information, tells already a lot about the relative location of the network in sequence space, is the profile of per-position base frequencies.

3.4.1. Profiles of Base Frequencies

For all networks except the open structure, the distribution of the bases $s \in \{A, U, G, C\}$, which we will call $count(i, s)$, has been determined for each position i in the reference structure. A simple measure which assigns but one value to a given structure is derived from this distribution as follows.

First, for each position the number of bases which occur at all is determined. These numbers are then averaged either over all positions or over some subset thereof, resulting in a coarse measure of equidistribution of the network (at the respective positions) over all dimensions in sequence space. If the subset is that of the paired positions, it is sensible to give the mean number of occurring pairs (assuming that the network is represented as a generalized Hamming graph). This number has not been determined directly, but it can be approximated by the number of *possible* pairs, which is equal to the number of valid base pairs which can be formed from the symbols which occur at the interacting positions. The numbers of possible pairs are then averaged over all base pairs in the structure. Figure 10 graphs the measure for unpaired and paired positions separately (the ordering of the structures is according to decreasing network size). Larger networks tend to be more equidistributed, with the exception of a group of structures around rank 100. We will see below that these consist of a single stack of length 2. Obviously, the formation of this unstable stack does not only strongly depend on its constituent base pairs, but also, to a lesser extent, on the unpaired context (not all symbols are permitted in the unpaired subsequences, too).

In the appendix, a conservation profile has been derived from these counts for each structure: the conservation at position i , c_i , is defined as

$$c_i = \frac{\max_{s \in \{A,U,G,C\}} \text{count}(i, s) - \min_{s \in \{A,U,G,C\}} \text{count}(i, s)}{\sum_{s \in \{A,U,G,C\}} \text{count}(i, s)}.$$

This measure is zero in the case of a completely uniform distribution and it takes the value of 1.0 if only a single base is permitted at this position. We see that even if the large networks do cover all dimensions of both $\mathcal{Q}_\alpha^{n_u}$ and $\mathcal{Q}_\beta^{n_p}$, the distribution is never completely uniform, at least in the paired fiber. The nonuniformities observed are too regular to be entirely explained by sampling fluctuations of an otherwise uniform random process: stacks are the more conserved the shorter (less stable) they are, unpaired positions are conserved in the vicinity of unstable stacks, and they indeed show a near-uniform distribution if far away from those stacks. Thus the assumptions of the random graph model do *not* hold (exactly) for the networks of \mathcal{Q}_{AUGC}^{16} . Still the model may be

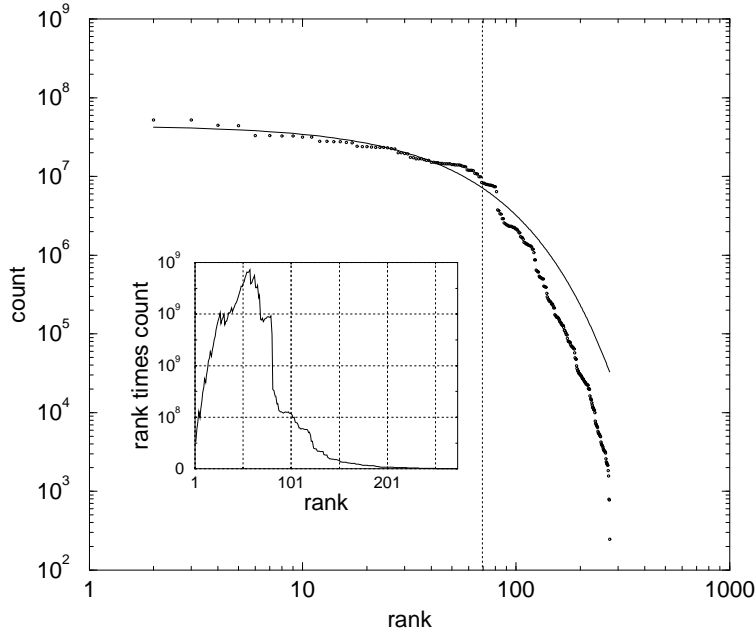


Figure 11: A fit of the rank ordered absolute network sizes to $f = A((1 + r/B)^\gamma)$, $A = 4.46157e + 07$, $B = 127892$, $\gamma = 3369.76$ (the open structure is not displayed). Circles corresponds to observed data points, the continuous line to the fitted function. In the insert the product of rank and count is plotted, which should be constant according to a simple Zipf's law. Obviously this is not the case.

a useful approximation, at least for the large networks.

3.4.2. Absolute Size Distribution

The absolute size of a minimum free energy neutral network is jointly determined by the number of compatible sequences (which sets an upper boundary to the size) and the mean energy gain by preferring a structure over other structures which the sequences may be compatible with. We have argued in section 2.6.3.1 that a Zipf distribution of the sizes will result if the number of energetically favorable structures with large compatible sets is small compared to all structures.

Fig. 11 plots the absolute network sizes versus the rank in the size distribution, on a double logarithmic scale. Note that the figure does not include the open structure. The networks fall into *three* classes: First, there is the open structure which is about 50 times as frequent as the structure which is second in the rank order. With a count of 2709569048 it comprises about 63% of the entire space ($4^{16} = 4294967296$). The counts of the remaining structures are best fit by two lines intersecting at about rank 70. One has a very shallow slope and comprises those structures which can be termed common. The other one, with a steep slope, corresponds to the rare structures. The boundary between the two regimes is more discontinuous than predicted by *Zipf2*. In addition, the counts of the rare structures fall off more rapidly than in the Zipf relationship.

Accordingly there is no very good fit to Zipf's law (formulated as *Zipf2'*). The parameters assigned by a nonlinear fit function are $A = 4.46157e + 07, B = 127892, \gamma = 3369.76$. A is of the same order of magnitude as the size of the biggest network involved (rank 2, with a size of 52505831), but B obviously has no meaning as a rank which separates common and rare structures. Other criteria which have been used to locate the boundary between common and rare structures are the mean redundancy of the sequence to structure mapping [81] or a fixed cutoff of 25% of the rank 1 network [54]. According to the first criterion, the boundary would be located at rank 39, because with a count of 16018755 this network just exceeds $|Q_{AUGC}^{16}|/\#structures = 4^{16}/274 = 15675063$. If the rank 1 network is included when applying the 25% criterion then this network is the only common one in the space: 25% of rank 1 is still about twice the size of the rank 2. Not including it (so that the cutoff is 25% of rank 2) results in a boundary separating rank 58 and 59 (of the full list).

Combining these various definitions, the boundary between common and rare seems to be smeared out over a rank range from about 40 to about 70.

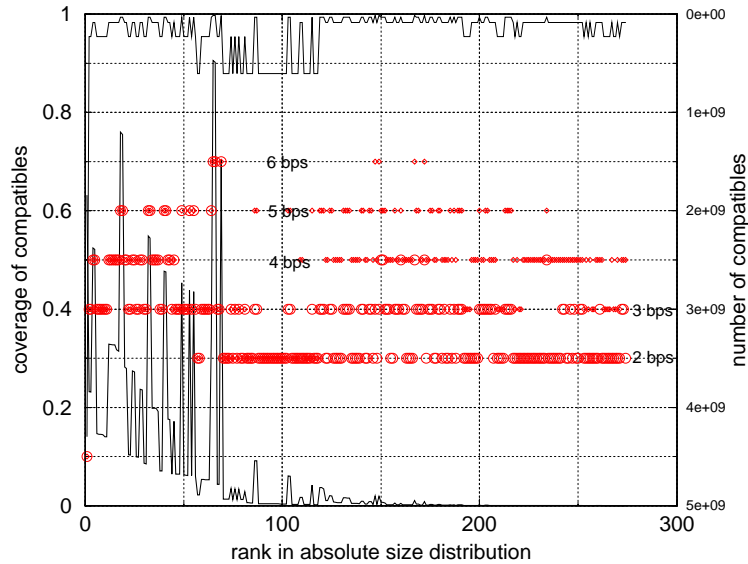


Figure 12: The fraction $n_{compatibles}/Size$ as a function of rank is shown as a bold line using the left y and the lower x axis. The right y axis and upper x axis span an upside-down plot of the number of compatible sequences versus rank. The lengths of the longest stack of the respective structures are indicated by big circles, and the total numbers of base pairs by small diamonds.

3.4.3. Normalized Size Distribution

The size of a network is bounded from above by the number of compatible sequences of the reference structure. This hidden dependency is factored out by dividing the network size by the number of compatibles. The resultant ratio, which we will call the *coverage of compatibles*, is a folding probability per compatible sequence (and can be used as such in the random graph model). Taken times some constant N , it gives the expected network size in case the compatible sets were of a constant size N for all structures. Fig. 12 shows the coverage of compatibles versus the rank in the absolute size distribution, along with the number of compatible sequences and the length of the longest stack per structure. The ambivalent effect of base pairing on the size of a minimum

free energy neutral network is evident: less pair constraints lead to a higher number of compatible sequences, yet the longer a stack, the higher the coverage. The coverage values decay according to (a) power law(s), which may be best fit by different exponents for different stack lengths. The decay within a length class is caused by an increase in loop size, different relative locations of the stack and/or less base pairs in additional structural elements. There are two broad regimes of coverage, which are separated at about rank 70 of the absolute size distribution. The first region, with a mean coverage of 0.25, includes the highest coverages of length classes 3 and 4 as well as all of classes 5 and 6. The end of the region is conspicuously punctuated by the structures ranked 65 and 66, with a coverage of 0.9 each (both have a stack of length 6). Immediately afterwards, coverage drops below 0.1 and stays below this figure for the remaining ranks (mean 0.06). The number of compatible sequences of the structures in this second regime is no different from that of the first one – between ranks 71 and 80 it is even exceptionally high (these structures have only two base pairs). This means that the absolute size of a network is mainly determined by the folding probability. That the boundary between high and low coverage is located at about rank 70 confirms once more that this rank constitutes a sensible point of separation between common and rare structures.

The ordering in figure 9 was according to the rank in the *absolute* size distribution. One can also do a direct rank ordering of the coverage values themselves. This is also a size distribution: namely, the expected sizes of the networks if the number of compatibles were equal to 1 for all structures. (A realistic value for N would not change the shape of the distribution, only its relative location.) Therefore it makes sense to ask for a fit to Zipf's law. The result is shown in figure 13 : the coverage values, which in an ordering according to absolute network size seem to fall onto different analytic curves, do form a common distribution. Note that the open structure *is* included in this case (it occurs at rank 6 in the distribution). The fit to *Zipf2* is clearly better than with the absolute sizes. The assigned parameters are $A = 0.927399$, $B = 89.9916$, $\gamma = 5.30659$. A is a

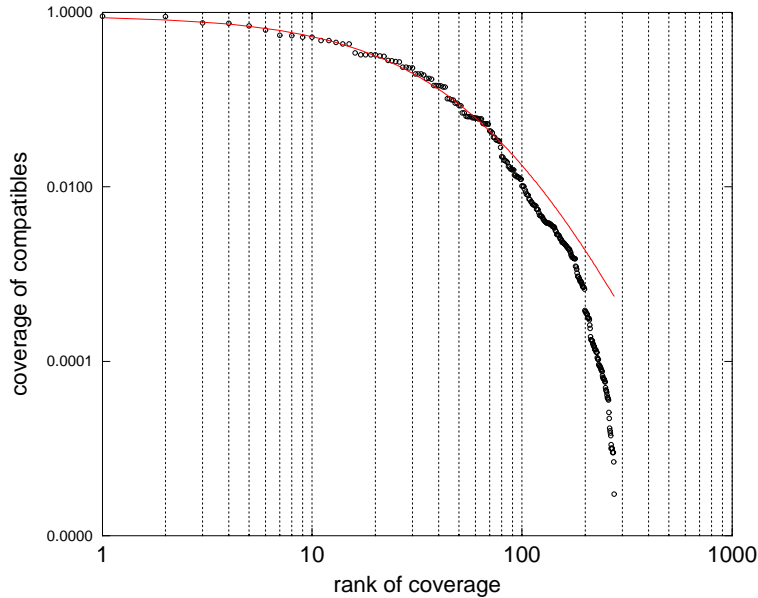


Figure 13: A fit of the rank ordered coverage of compatibles to $f = A((1+r/B)^\gamma)$, $A = 0.927399$, $B = 89.9916$, $\gamma = 5.30659$.

reasonable scaling parameter (the rank 1 coverage (rank 65 of the absolute sizes) is equal to 0.905). B is at least a crude estimate of the rank which separates the shallow from the steep regime of the curve. This boundary is located somewhere between ranks 40 and 70, quite like in the case of the absolute sizes. Yet although many of the structures which are common according to absolute size do show up at ranks less than 70 in the coverage ordered distribution, the two criteria do not pick identical sets of structures. Structures which rank less or equal to 70 according to both criteria are ranks 1 to 56 and 59 to 69 of the absolute size distribution. One could propose to define a common structure by membership in this intersection.

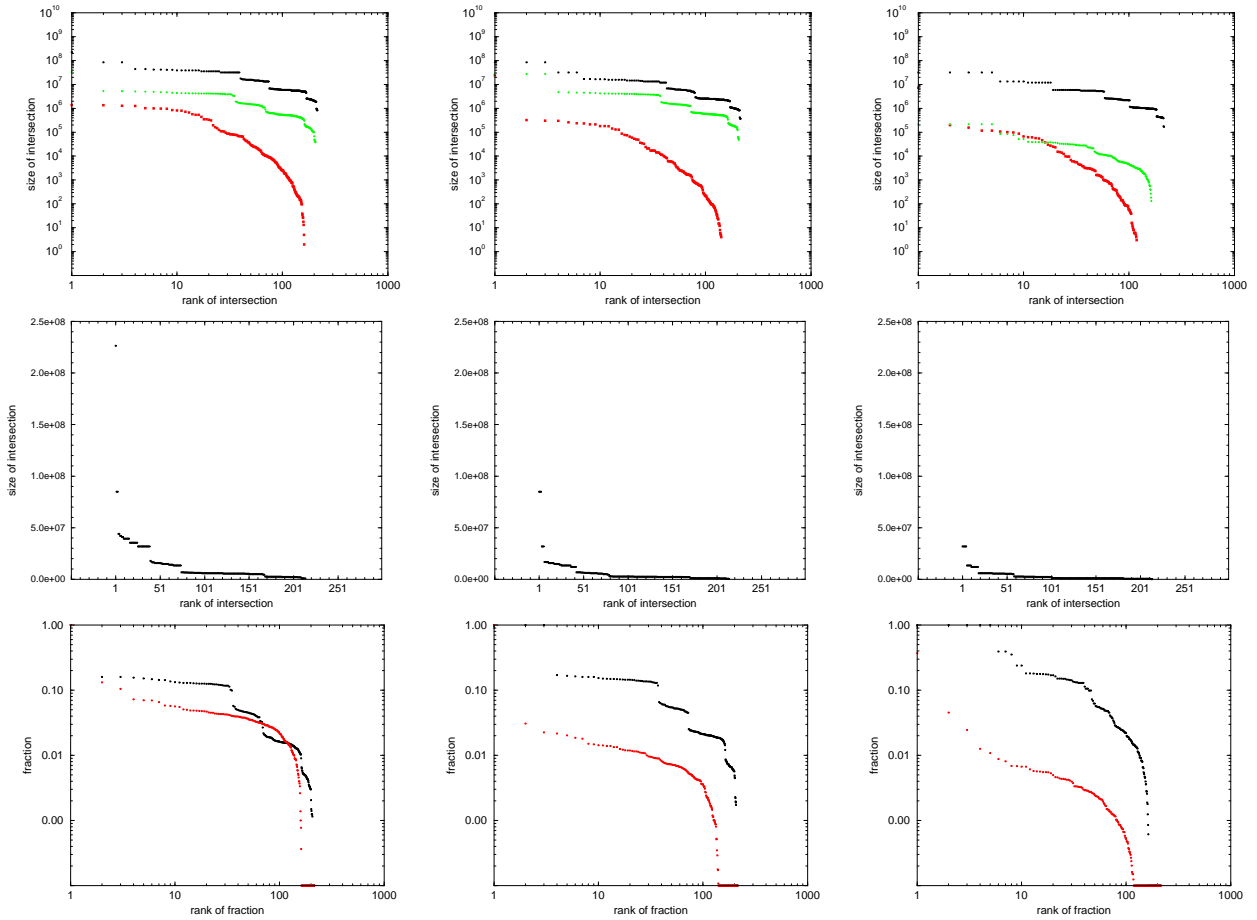


Figure 14: Distributions of intersection sizes and “visibility”. The columns correspond to the 3 reference structures ranked 6, 14, and 150 in the absolute size distribution. The plots in the upper row show count distributions on a double logarithmic scale: $|I[s_1, s_i]|$ (uppermost curve), $A(s_1, s_i)$ (middle curve), and $B(s_1, s_i)$ (lowermost curve). In the second row, $|I[s_1, s_i]|$ is again displayed, this time on a linear scale (which shows more clearly how the distributions are made up of discrete values). In the third row, $A(s_1, s_i)$ and $B(s_1, s_i)$ have been scaled by the respective network sizes to yield “visibilities” $A'(s_1, s_i)$ (dark, upper curve) and $B'(s_1, s_i)$ (light, lower curve). Note that there are several orders of magnitude less on the y scale in the plots of row 3 than in row 1, which is why the tails of $A'(s_1, s_i)$ seems to fall off more rapidly than that of $A(s_1, s_i)$. On equal scales, the shapes of the two distributions are identical.

3.4.4. Intersections

The intersection or overlap $I[s_1, s_2]$ of two RNA secondary structures s_1 and s_2 is defined to consist of those sequences which are simultaneously compatible with both structures [100]. Focusing on one structure, both the neutral network of the structure and the overlap are subsets of the set

of compatibles. The set theoretical relation of the two subsets may be anything between the two extremes of an empty intersection and complete inclusion of the neutral network in the overlap. Being both in $I[s_1, s_2]$ and in one of the neutral networks means that the sequence may directly switch structure under suitable conditions. In a static landscape like the one discussed here this is ignored, yet the number of sequences with this property still is a measure of accessibility of one structure from the other, in the following sense. As discussed in chapter 2, the 1-neighbourhood of an RNA sequence is usually defined with respect to its structure, exchanging paired positions in a single step. *Under this assumption* it follows that *any* neighbour of a sequence on the neutral network of s_1 is compatible with s_1 , whether or not it is a neutral neighbour. Any sequence on the network of s_2 which can be reached by a 1-move from the network of s_1 thus is in $I[s_1, s_2]$. Therefore the cardinality of the intersection of $I[s_1, s_2]$ and the network of s_2 is an upper limit to the number of sequences which fold into s_2 and are in the *boundary* of the network of s_1 . If it is zero, it is clear that the boundary does not contain s_2 , without further inspection.

In the following a reference structure s_1 is fixed and for all structures $s_i \neq s_1$ we determine three basic distributions: a) $|I[s_1, s_i]|$ (by means of the algorithm given by [100]), b) the number of sequences which fold into structure s_1 and at the same time are compatible with structure s_i (in the following referred to as distribution A), and c) the number of sequences which fold into s_i , being compatible with s_1 (distribution B). Together with two derived distributions which will be discussed below, they are displayed in figure 14.

A single count in the distribution of $|I[s_1, s_i]|$ means that s_i appears in the list of candidate structures of a compatible sequence of s_1 . Structures with a high count are those which are in many lists. A structure is the more likely to be in an arbitrary list the less constraints are put on a sequence by the requirement to be compatible with the reference structure, too. The best case, in which there are no additional constraints, occurs when the set of base pairs of s_i is a subset of those of s_1 . The number of structures which fulfill this criterion (or the weaker criterion that there

are no long *orbits* with the reference structure (cf. chapter 5) is small compared to the number of structures which do not fulfill it. By the arguments given in chapter 2, this points to a Zipf like distribution.

The relationship is certainly exponential (cf. figure 14). The linear plots show how the curve is put together from building blocks on two scales: first the intersection sizes come in certain discrete values each of which is the result of some combination of orbit lengths. When plotted in rank order, some of the discrete sizes are so similar that they join into near-constant steps, which are separated from other such steps by large jumps. These higher-order steps are the most conspicuous feature of the distributions. The overall exponential distribution comes about by an approximate bisection of the sizes in subsequent steps. (This indicates that one constraint of some sort is added per step. Roughly but not exactly it is the number of base pairs in s_i which determines the step to which s_i belongs (data not shown)). Averaging over the steps, the distribution looks very much like a *Zipf2* curve. Note that the rank order according to the intersection size with some reference structure is mostly very different from the rank order according to absolute network size (the correlation between the ranks according to the two criteria is (0.12, 0.215, 0.13) for the three reference structures).

More relevant for the topology of the landscape than the intersection sizes themselves is the distribution of sequences in the intersection which actually fold into one of the shapes. It constitutes a limit distribution for the *boundaries* of the landscape's neutral networks (see section 3.5.3.4). In contrast to the boundaries themselves, it can be determined without additional folding once the networks are known (by traversing the reference network and testing the sequences for compatibility with the structure s_i of interest, a test which needs linear time in the sequence length).

In distribution A, only one neutral network is involved (that of the reference structure). The set of sequences which are compatible with some s_i is a subset of the reference network, and it is sampled from $I[s_1, s_i]$, which is a subset of the compatibles of the reference structure. Thus in the most ideal

case, distribution A is derived from the distribution of intersection sizes by a multiplication by the average folding probability of the reference structure. In practice, $I[s_1, s_i]$ is not a uniform subset of the compatibles, which is why the folding probability on this set may differ from the average. (Imagine $s_1 = \dots(((((((\dots))))))\dots)$, $s_i = \dots(((((((\dots))))))\dots)$. The folding probability for s_1 of sequences in the intersection is certainly below average, because it is energetically more favourable to choose s_i instead of s_1 .) Nevertheless, distribution A is essentially a shifted copy of the distribution of intersection sizes in all three cases.

In distribution B, the sampling from $I[s_1, s_i]$ is with respect to a different neutral network for each s_i , so that the folding probability is not even in theory assumed to be constant. We have seen above that the average folding probability of a compatible sequence, which is equal to the coverage of compatibles, is distributed over the structures in a “Zipfian” manner and correlates with the absolute size of the network. Distribution B shares both features. The steps in the underlying distribution have become completely blurred and the curves are more markedly bending downwards with increasing rank in a log-log plot. The correlation of rank order in distribution B and rank with respect to absolute network size is (0.82, 0.73, 0.70) for reference structures (6, 14, 150), while it is only (0.07, 0.10, -0.01) in case of distribution A.

If $B(s_1, s_i)$ is divided by the network size of s_i , a measure results which can be termed a visibility of structure s_i from network s_1 : a value of 1.0 means that all sequences in the network of s_i can potentially be in the boundary of s_1 . If the value is small, then it depends very much on the relative location on network s_i whether or not structure s_1 is accessible (and therefore it may not be very meaningful to talk about “the” accessibility of one structure from the other). The rank orders of $B(s_1, s_i)$ and visibility need of course not coincide (and they do not), but rank ordered visibility, too, exhibits a Zipf-like distribution, with a tail which falls off even more rapidly than in distribution B. On the high frequency end, the curves are bend upwards, resulting in an overall sigmoidal shape. These “overly visible” structures either exactly contain the pairing pattern of s_1

(so that the neutral network is a subset of $I[s_1, s_i]$ and the visibility is 1.0) or at least have only few additional constraints.

The visibility of structure s_1 from network s_i is identical with distribution A, except for a scaling by the size of network s_1 .

3.5. Graph Topological Features

3.5.1. Distance Distributions

Size, coverage of compatibles, and cardinality of the intersection are single numbers associated with one or two networks. The next step towards a more refined view is to investigate the distribution of (Hamming) distances between pairs of sequences of the network. This gives an idea of how the network is located in sequence space, while still ignoring the graph structure.

The observed distance distribution may be tested against several hypotheses. In the order of increasing constraints on the sequences, these are :

- (1) *Is the network a random sample from \mathcal{Q}_{AUGC}^{16} ?*
- (2) *If not so, is the network a random sample from its compatibles ?*
- (3) *If not so, is this caused by position specific probabilities only or are there additional correlations in the sequences ?*

In a “random” sample from a sequence space, symbols occur with uniform probability at each position, and the symbol assignments at any two positions are independent. In such a setting, pairwise distances are distributed as $B(n_{symbols}, \frac{(n_{symbols}-1)}{n_{symbols}})$ ($B(n, p)$ from here on will denote the binomial distribution with n trials and probability of success p) : for every pair of sequences, the Hamming distance is equal to the number of mismatches observed in *length* trials. It is clear that

no network (with the exception of the open structure) can ever be exactly distributed like this, because the base pairs introduce correlations into the sequences. Nevertheless it is interesting to ask how close the distance distribution comes to $B(16, \frac{3}{4})$, because this is the reference distribution for “no clumping in sequence space”. The range of sequence variation which can be covered by a population of sequences while retaining their structure is largest if the network of the structure is approximately “randomly” distributed.

In contrast to condition (1), condition (2) can in principle be met by the members of a network. To test it, one deletes the positions which correspond to the upstream partners of the base pairs, resulting in a sample of sequences of length $length - n_p$, and compares this to $B(16 - n_p, \frac{3}{4})$. Stable stacks result in a good to perfect fit to this distribution (see below). There are however many networks which do not fit it, hence are not uniformly distributed in their compatibles.

In many networks the reason for a lack of fit to $B(16 - n_p, \frac{3}{4})$ is immediately evident from looking at the profile of per-position base frequencies (see section 3.4.1): often, these frequencies are very different for different symbols to the extent that one or more symbols are never found at a given position (a very common case of this sort are A-U or G-U pairs in unstable stacks).

The expected distance distribution under the hypothesis of complete independence of the positions (but taking account of position specific base frequencies) is constructed from the frequencies as follows:

The probability of a *mismatch at position i* is

$$p_{mi}(i) = \sum_{j_1=1}^{nsymbols[i]} f_{j_1}[i] \sum_{j_2=1, j_2 \neq j_1}^{nsymbols[i]} f_{j_2}[i],$$

where $nsymbols[i]$ is the number of bases which are permitted at position i and $f_j[i]$ is the frequency of the j th base (at position i).

Accordingly, a match has probability

$$p_{ma}(i) = 1.0 - p_{mi}(i).$$

Next, all possible subsets of the set of positions are enumerated by successively incrementing a counter N from 0 to $2^{n_{pos}} - 1$, at each step converting N to the vector of zeros and ones which corresponds to its binary representation. Let d be the sum of entries of the vector. If a one is thought of as representing a mismatch in a pair of sequences, and a zero as representing a match, then d is equal to the Hamming distance of two sequences. This particular constellation of matches and mismatches adds

$$\Delta p(d) = \prod_i (v(i) \times p_{mi}(i) + ((1 + v(i)) \bmod 2) \times p_{ma}(i))$$

to the total probability $p(d)$ of distance d which at the beginning had been initialized to zero for all d . Here, v is the above mentioned binary vector. Dependent on $v(i)$, either the mismatch probability or the match probability is selected as the i th factor.

Intermediate between a plain $B(16 - np, \frac{3}{4})$ and the detailed distribution above is the idea to break a sequence into subsequences according to the number of different symbols which can occur (so that this number is constant for each subsequence). Let n_{s_i} be the number of symbols which can occur in subsequence s_i . Then the number of mismatches at positions which belong to s_i is assumed to be binomially distributed with $p_{s_i} = \frac{(n_{s_i}-1)}{n_{s_i}}$ and $n = length(s_i)$. Let d_{s_i} be the distance of the positions of set s_i in a pair of sequences. The total distance d of the pair obviously is equal to $\sum_i d_i$. Then a given constellation of (d_{s_i}) contributes

$$\Delta p(d) = \prod_i p_B(length(s_i), p_{s_i}, d_i)$$

to the total probability of d (where $p_B(n, p, k)$ is the probability of k successes given that successes are distributed as $B(n, p)$).

A common method for testing whether observed per-bin counts match the number of counts which are expected from some distribution is the χ -square test. The χ -square statistics is defined as

$$(1) \quad \chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i}$$

where N_i is the number of events in bin i which are observed, and n_i is the expected number.

$Q(\chi^2|\nu)$, the probability that a correct model will produce a distribution with $\chi_{distr}^2 > \chi^2$, can be calculated in closed form as an incomplete gamma function [69] (ν is the number of degrees of freedom in the model).

From (1) it is apparent that the absolute values of χ^2 are dependent on the order of magnitude of the observed and expected counts (the squared numerator is increasing faster than the linear denominator). In the following comparison of observed distance distributions versus the various expected distributions which have been discussed above, observed and expected counts are divided by the maximum value of the observed counts. The resulting probabilities (which are quite different from the ones obtained without normalizing!) have been found to reflect the “intuitive” similarities of the distributions well: a value of 0.9 means that the distributions nearly coincide, while values near zero means that we are dealing with two different distributions, normally with a considerable relative shift. Yet one has to keep in mind that this “probability” is a heuristic variable, *not* $Q(\chi^2|\nu)$ as it would have resulted from the non normalized counts.

Testing the full length distance distributions versus a $B(16, \frac{3}{4})$, the probability is zero to 6 decimal places for all networks of Q_{AUGC}^{16} – full length sequences as expected can never be completely “dispersed” over the space. The absolute χ^2 values however span a range of several orders of magnitude and in overall tendency increase exponentially with the rank of the network (cf. Fig. 15). Networks up to about rank 60 can be thought of as approximately uniformly distributed. Note that the threshold separating common from rare structures had above been located somewhere between rank 40 and 70.

Regarding the degree of fit of the distance distributions of the shortened sequences to the simple binomial, piecewise binomial, and resolved expected distributions, the structures of Q_{AUGC}^{16} can be partitioned into 3 classes. First, there are the common structures, which exclusively (up to rank 57) consist of single hairpins made of three or more base pairs. Among these structures are

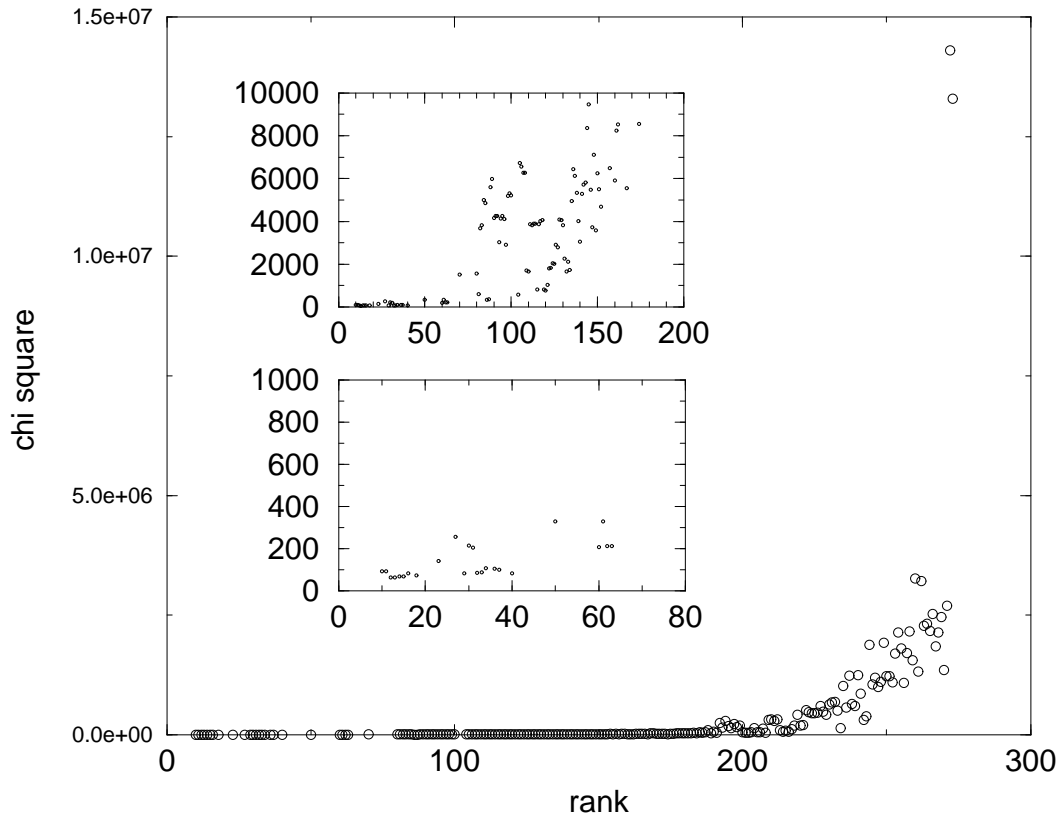


Figure 15: χ_2 distance of the observed distance distributions to $B(16, \frac{3}{4})$. Both observed and expected values are divided by their respective maximum before comparing the distributions. Although there is not a single very good fit, it is possible to distinguish 3 classes of networks. The distributions of the common structures, up to rank 70, are closest to $B(16, \frac{3}{4})$, despite the presence of base pairs. Intermediate χ_2 values characterize structures between ranks 70 and 200, while the distance distributions of structures beyond rank 200 have nothing whatsoever to do any more with $B(16, \frac{3}{4})$.

the only networks which show a near-perfect fit to the simple binomial distribution, that is, which are approximately uniformly distributed over the set of compatibles. Rank 10 is an example.

Fig. 16 shows the distance distribution for full length sequences (solid line) and for shortened sequences (dashed line) compared to $B(16, \frac{3}{4})$ (solid fat line line) and $B(13, \frac{3}{4})$ (dashed fat line). (The rank 10 reference structure contains 3 base pairs). In the full length sequences, there is a considerable underrepresentation of intermediate distances and an overrepresentation of short ones. In this case this is entirely due to the nonindependence of paired positions: the network *is*

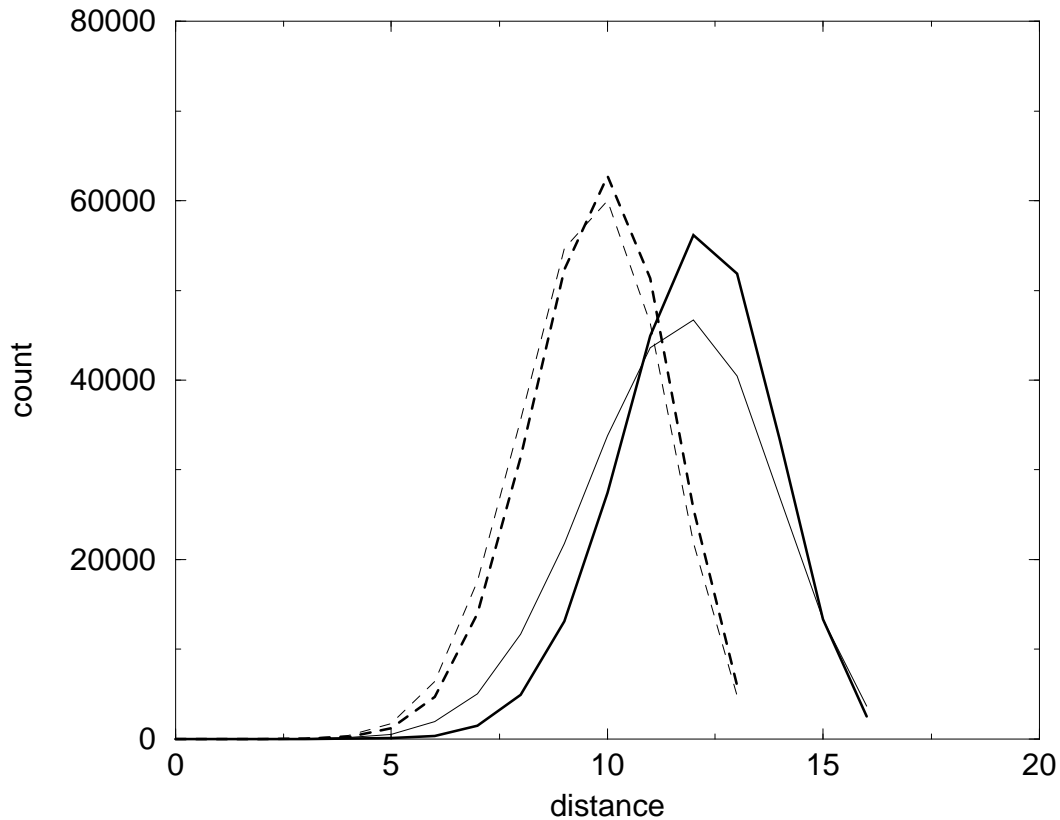


Figure 16: Distance distribution of 500 full length sequences from the rank 10 network (continues light line) compared to $B(16, \frac{3}{4})$ (continues fat line), the distance distribution of the same sequences with the upstream partner positions of the base pairs deleted (light dashed line), and the expected distribution for the shortened sequences, $B(13, \frac{3}{4})$ (fat dashed line). In the full length sequences, there is considerable buffering (overrepresentation of small distances) and an underrepresentation of intermediate distances, compared to $B(16, \frac{3}{4})$. But these effects can be nearly completely be accounted for by the coupling of paired positions: the distance distribution of the shortened sequences nearly coincides with its expectation.

uniformly distributed over its compatibles.

There is another class of structures besides the common ones the distance distributions of which are easy to explain. Those are the structures with two base pairs only. In the rank list, their “realm” is located between the common structures and those rare structures which contain subelements, approximately from range 70 to range 120. The two base pairs are always adjacent, so these structures contain a single structural element, a 2-stack. The distances do not fit a simple binomial

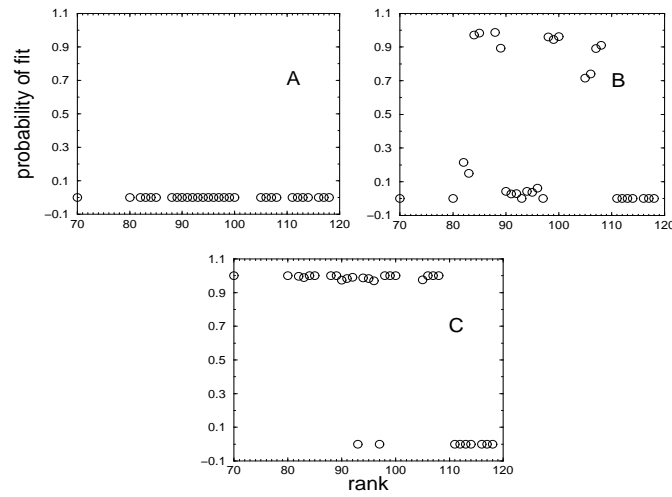
distribution, because there are sequence constraints on the stack (normally $G - C$ only, sometimes restricted to a particular orientation). In some structures, the stack is completely conserved as $5' GC \dots GC 3'$, this being the only sequence constraint. Accordingly the distances of the shortened sequences fit the piecewise binomial distribution well to perfectly (dependent on the degree of uniformity of the base composition of the remaining positions). In some other networks, there are more pronounced position specific base inhomogenities. But in these cases, too, the positions seem to be independent, as judged from the fact that there is a perfect fit of their distance distributions to the resolved distributions. The only $n_p = 2$ networks the distance distribution of which cannot be reproduced in this way correspond to 9 structures which contain a tetraloop. All of these 9 networks decompose into two components each, each with a characteristic pattern of nucleotides in the region of the tetraloop (which in part can be considered an artifact of `RNAfold`, which gives a built-in bonus to certain preferred patterns).

What remains is the class of non-hairpin structures. They commonly show major deviations from a uniform base composition at many positions, which is why the fit to the piecewise binomial distribution is generally bad (as well as the fit to $B(\text{length} - n_p, \frac{3}{4})$). Networks of structures in the rank range from about 100 to about 200 mostly fit the resolved distribution very well, while most of those beyond rank 200 do not. In the latter ones, seemingly not all positions are independent. This explains why these structures are at the very end of the rank list: like in a base pair, mutating one out of a set of non-independent positions will most likely not result in a neutral neighbour, resulting in low neutrality and density of the structure.

3.5.2. Overall Connectivity

The Sequence Of Components decomposition described in section 3.2.2 was carried out for all networks in the landscape except for that of the open structure (the latter was omitted not only because it would have taken too much time to process this huge network, but also because for

structures with 2 base pairs



structures with 3 base pairs

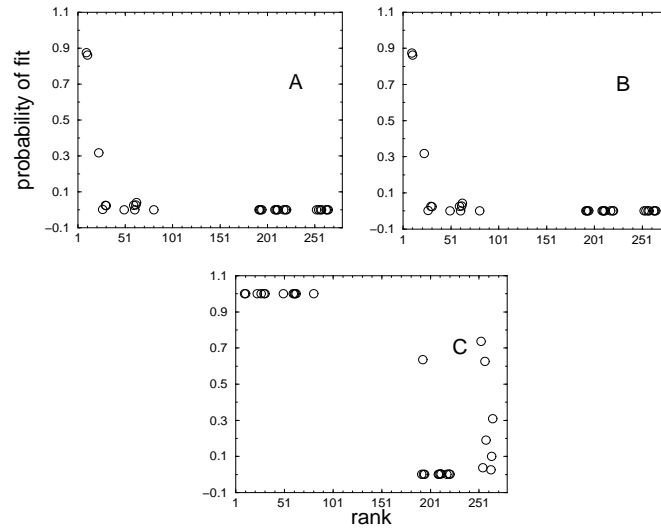
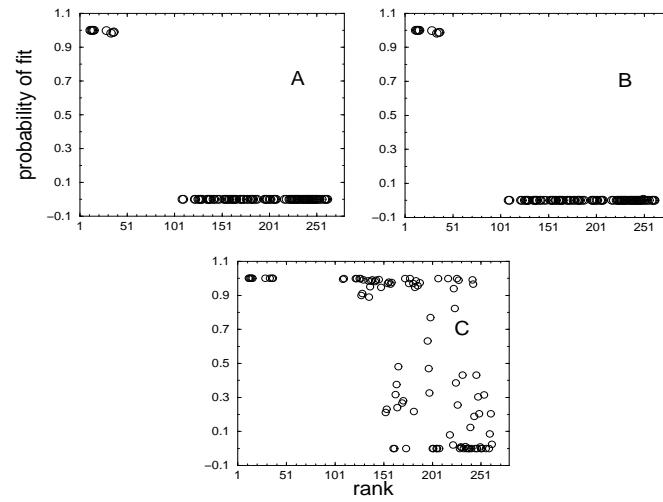


Figure 17: Degree of fit of the observed distance distribution to a) $B(16 - n_p, \frac{3}{4})$ (upper left), b) a piecewise binomial distribution based on the number of symbols which can occur at the different positions (upper right), and c) expected distribution with position specific mismatch probabilities (computed from the observed per-position frequencies of the symbols), but assuming that positions are independent (lower), for $n_p = 2$ and $n_p = 3$.

structures with 4 base pairs



structures with 5 base pairs

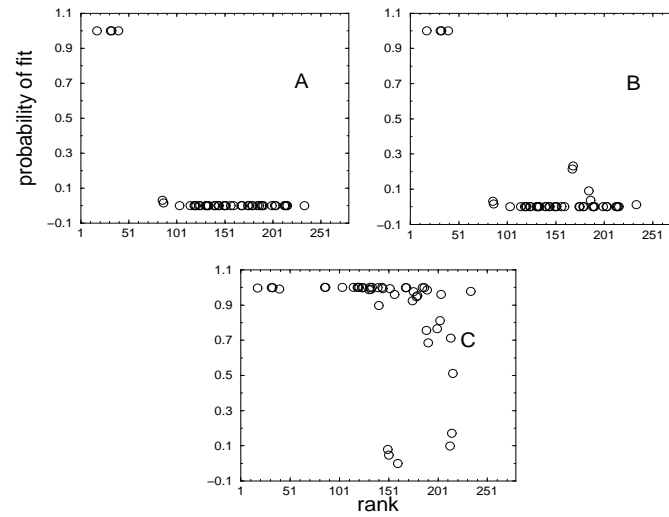


Figure 18: Same as Fig. 7, for $n_p = 4$ and $n_p = 5$.

reasons discussed below there are already strong hints that it is dense and connected).

Of the 273 networks analyzed, 229 are connected graphs. The first disconnected network occurs at

structures with 6 base pairs

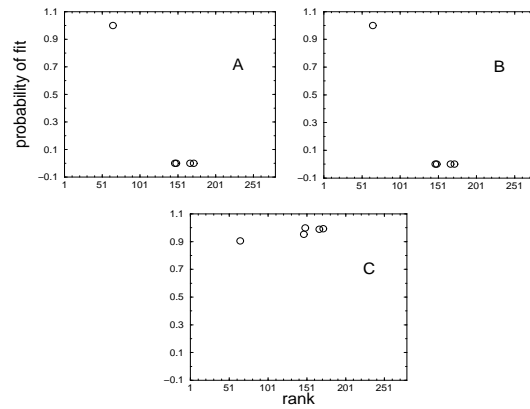


Figure 19: Same as Fig. 7, for $n_p = 6$.

rank 93. Of the 44 disconnected networks, 42 share a common feature: the underlying structure contains a tetraloop with an unstable stack (size of stack 1 or 2) and each component shows a characteristic pattern of nucleotides in the region of the tetraloop. The two remaining disconnected networks correspond to structures which contain 5-loops with a stack size of two. Having an unstable 4- or 5-loop is *not* a sufficient condition for a network to be disconnected. Ranks 163 to 166, for example, correspond to structures which contain tetraloops, yet they are connected. The probable cause is the presence of additional (stabilizing) base pairs outside the 2-stack.

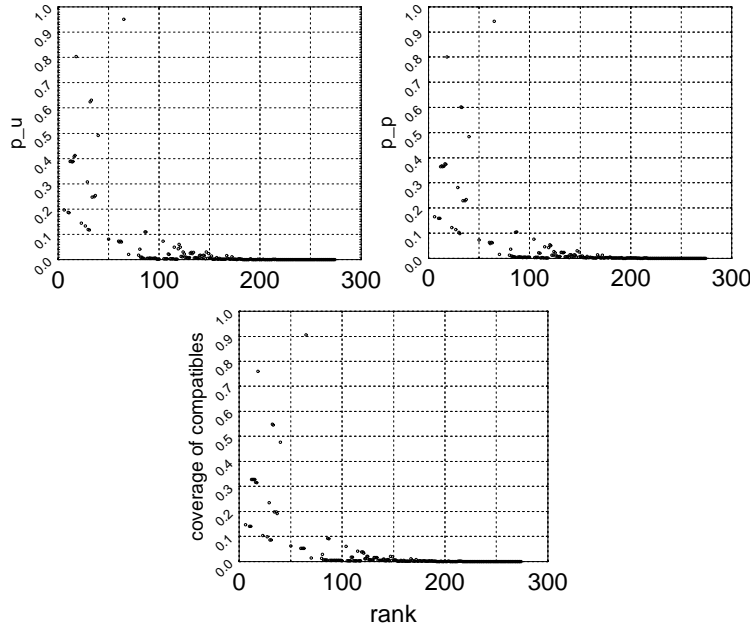


Figure 20: Upper two plots: mean fractions of neutral neighbours of a random compatible point (for 220 network of Q_{AUGC}^{16} , 1000 points per structure). Lower plot: coverage of compatibles of the same networks. p_u and p_p are quite similar to each other as well as to the coverage values.

3.5.3. Distribution of Shapes in the Vicinity of a Sequence

3.5.3.1. Lambda Values

An induced subgraph generated by sampling vertices from a generalized Hamming space $Q_\alpha^{n_u} \times Q_\beta^{n_p}$ is expected to be dense and connected if the choosing probabilities for the paired and unpaired fiber fulfill the relations $p_u > 1 - \alpha^{-1}\sqrt{\alpha-1}$ and $p_p > 1 - \beta^{-1}\sqrt{\beta-1}$ (cf. section 2.6.3.2). In the case of Q_{AUGC}^{16} , $\alpha = 4$ and $\beta = 6$, leading to the values of $p_u^* = 0.37$ and $p_p^* = 0.30$ for the threshold probabilities.

It is known from previous studies [37, 38] that the folding probabilities in minimum free energy neutral networks often depend on whether the average is taken over all compatibles or else only those compatibles are considered which are 1-neighbours of a folding sequence. The reason for this is the fact that the sets of candidate structures of two sequences which differ by a single mutation

(point or pair exchange) are expected to be more similar than the sets of two random compatible sequences. If the reference structure is the energetically most preferred one in one of the lists, then it is more likely that this is the case in its immediate neighbour, too, than in an unrelated compatible sequence. In order to catch this effect, the choosing probability is usually estimated not by direct testing for membership in the network of a sample of compatibles, but by recording the fraction of *neighbours* of a compatible sequence which are on the network. This is actually the reason why neighbourhood information is required for determining this probability. In the following we will call the estimated probability p if the set of reference compatible sequences is uniformly sampled from all compatibles. If the sampling is from folding sequences only, it is called λ , while δ refers to a nonfolding set of reference compatibles. For all three probabilities, there are two separate values which pertain to unpaired and paired neighbours of the reference sequence, *rsp.* They are distinguished by the subscripts u and p .

Fig. 20 shows the distribution of the estimated p_i values, $i \in u, p$, for the 220 networks which have been indexed. The two distributions are very similar to each other and, consequently, to the coverage of compatibles ($p_u + p_p - p_u p_p$ estimates the coverage). p_i values above the respective connectivity thresholds are confined to ranks less or equal to 50 (with the exception of rank 65, which has an unusually high coverage). From this one would conclude that most of the networks are neither dense nor connected. As far as density is concerned, this is true (see below). Yet most of the networks are connected graphs.

This can be partly understood by observing that λ_i , $i \in u, p$, is much higher than p_i in this landscape. Accordingly, δ_i , $i \in u, p$, is lower (the three measures are related as $p_i = \lambda_i + \delta_i$).

Fig. 21 shows the distributions of λ and δ . λ_u , which is about linearly decreasing with rank, drops below the connectivity threshold only from about rank 250 on (with a few outliers at higher ranks). The values of λ_p are generally smaller than the corresponding λ_u values, with a similar negative correlation with rank. (There are some outliers around rank 100 with $\lambda_p = 0$. These

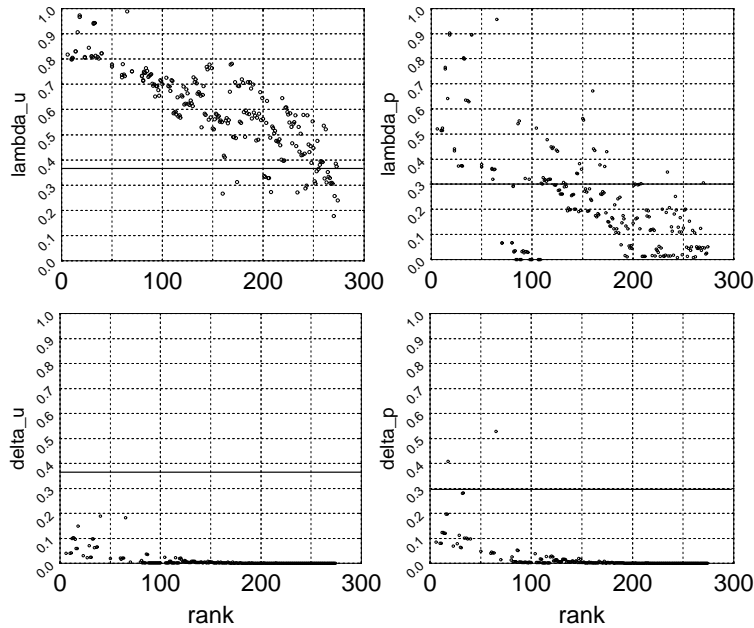


Figure 21: Mean fractions of neutral neighbours for 214 network of Q_{AUGC}^{16} . The upper two plots show the fraction of neutral point (pair, resp.) neighbours for sequences which are on the network (sample size per network 5000, except for ranks 10, 83-100, 104, 106-114, with a sample size of 500). The lower two plots depict the same fractions for sequences which are *not* on the network. The horizontal lines mark the threshold values above which the respective fiber is expected to be dense and connected, if it were constructed by the random graph model of [71]. It is evident that it makes a big difference whether or not the reference sequence is on the network, a fact which is not in agreement with the random graph model.

are 2-stacks in which $5'GC \dots GC3'$ is the only acceptable occupancy of the stack.) A regression line would cross the connectivity threshold at about rank 120. The λ_p s of the common structures all exceed the threshold. The distribution of δ for the unpaired and paired fiber resembles the distribution of coverage of compatibles, with a general shift towards smaller values.

For most other networks than the common ones, the condition $\lambda_u > 1 - \alpha^{-\sqrt{\alpha-1}}$ and $\lambda_p > 1 - \beta^{-\sqrt{\beta-1}}$ still does not hold. Two possible reasons come to mind why many of them nevertheless are connected. Both explanations constitute deviations from the random graph model. First, we have seen that in many structures with unstable elements there are sequence constraints. This means that some dimensions of sequence space are not populated by the network, so that the observed network is but one component of a graph which would have arisen as the result of a

random process sampling all dimensions. Secondly we observe that very small values of λ_p often are associated with a small number of base pairs. The 2-stacks close to rank 100 are an example. The smaller the number of base pairs, the larger the unpaired fiber (obviously). In the random graph model introduced in chapter 2, the number of possible contexts does *not* matter (the unpaired and paired subsequences v_u, v_p are chosen independently with the probabilities λ_u, λ_p and only afterwards combined). It is however known that the formation of poorly stabilized base pairs in minimum free energy structures is certainly context dependent [28, 97] (we have seen this above in section 3.4.1). Choosing (folding) in such a situation would be more appropriately described by a distinct probability for each context, finally accepting a paired subsequence if it has been selected in at least one context. The number of contexts then would play an important role (as well as the distribution of probabilities among them). The more contexts, the more likely it would be that for each pair mutation there is at least one which accepts it with a high probability. Then the network is connected if the unpaired fiber itself is connected, which probably is the case given the mostly high λ_u values.

3.5.3.2. Variance of Outdegrees

If the folding probability is constant, then the number of neighbours on the network per compatible sequence will be binomially distributed. A non constant probability distribution can lead to either overdispersion or underdispersion relative to the binomial expectation [67, 102, 68]. Overdispersion results if each experiment (consisting of n trials of which the successes are counted) has its own parameter p_i , which however is fixed across the n replications. The equivalent in the case of neutral networks is a constant folding probability for all 1-neighbours of a given compatible sequence, taking a different value for each neighbourhood. Of course the latter two conditions can never fully hold simultaneously (if all neighbours take exactly the same probability, then it is constant on all compatibles), but the fact that $\lambda \gg \delta$ points to a considerable correlation between the

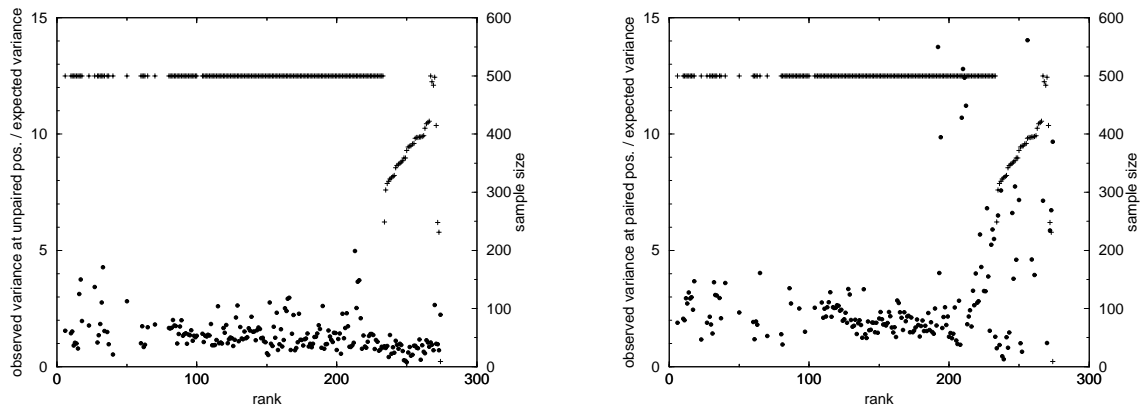


Figure 22: The ratio of the observed variance of the outdegree of network members to the expected variance according to a binomial distribution (filled circles). The number of replications n in the assumed binomial distribution has been set to the mean number of symbols (bases or base pairs) which are actually observed in the respective networks, rather than to the maximum number according to the alphabet. The number of neighbourhoods on which the observed distribution is based is marked by a plus sign for each network.

folding probabilities of neighbouring compatibles. Among-neighbourhood variation of the folding probability amounts to context dependent folding. Underdispersion on the other hand is caused by a variation of the parameter *within* the n replicates of a given experiment [67]. The acceptance of mutations in this case depends not on the context but on the mutated position itself.

Rather than in the distribution of outdegrees over all compatibles, we are interested in that over the sequences on the neutral network. How different the two are depends on how different the distribution of folding probabilities on the folding sequences is from the full distribution. Naturally high probabilities will be overrepresented in the former. If the latter is bimodal, it can happen that the low probability mode is not at all represented in the former, resulting in a much more uniform distribution on the network. If the probabilities correlate with sequence features, then such a situation will manifest itself as exclusion of certain dimensions of sequence space in the members of the network, as has indeed been observed in some of the profiles. Yet the distribution of outdegrees may be close to its binomial expectation (taking into account the reduced number of trials n (dimensions sampled)). A more smooth full distribution on the other hand will be

more uniformly sampled by the network, and it will then again depend on the degree of within-neighbourhood correlation of the folding probabilities whether or not this will cause overdispersion or underdispersion.

For each of the 220 networks indexed, we have sampled 500 reference sequences on the network and have determined their respective number of neutral neighbours. In figure 22, the variance of the resulting (observed) distributions is compared to the variance of $B(n_i, \lambda_i), i \in \{u, p\}$. The λ_i are estimated in the same run from 5000 sequences different from the ones which make up the distributions. n_i is the mean number of bases at unpaired positions in the network, respectively the mean number of possible base pairs, as introduced in section 3.4.1. Note that the maximum sample sizes are not reached any more from rank 234 on, as indicated in the figure by + signs. The networks in this rank range are so small that it has not been possible to find 5500 different sequences by random selection within reasonable time.

The ratio of observed and expected variance at unpaired positions has a mean of 1.35 over all networks. Surprisingly, it is the small networks for which the observed variance is less than the expected one. These are structures in which there are often constraints on individual unpaired positions (see the conservation profiles in the appendix), so that it may actually be the case that the position specific within-neighbourhood variation of folding probability is more influential than the among-neighbourhood variation. Underdispersion would then be expected. Between ranks 70 and 220, most ratios cluster near 1.0, outlier at ratios > 1.0 being more frequent than those below that figure. The distinction of the common structures (rank less or equal than 70) as a separate regime is partly an optical illusion due to the fact that on many of these networks there is no data (they have not been indexed because they are too big). Yet it is true that some of these networks have very high ratios. The only distinguishing features of these structures is that they contain tetraloops (ranks 16, 17, 32, 33, 50). Unlike the tetraloop structures at higher ranks, these networks are connected, so there must be neutral paths from one of the preferred tetraloop

constellations to the other. It might be the case that there are long range variations in folding probability according to the distance from a stable tetraloop configuration.

For the paired fiber, the ratio of the variances exceeds 1.0 for all but 9 structures (mean 2.86). Again there are some high ratios among the common structures. These are actually the most stable ones among the structures analyzed (the ones with the longest stacks). Because of the cooperative stabilization, the distributions of numbers of neutral neighbours peak at the maximum number, while the expected peak is at a smaller number. The high variance is caused by the fact that a broad range of numbers below the maximum is populated, too. In contrast to the unpaired fiber, the highest ratios of the paired fiber occur in the smallest networks. The mostly small average number of possible pairs in these structures (cf. Fig. 8) leads to very small expected variances, so that the absolute values of the observed variances need not be very high in order to produce these ratios. From the enhanced variance we conclude that even the formation of the few permitted pairs is context dependent, which is supported by the fact that there are sequence constraints on the unpaired contexts (leading to reduced observed variances in the unpaired fibers of these structures).

3.5.3.3. Density

In chapter 2 an induced subgraph has been defined to be dense if there is no node in the underlying graph which not either is part of the subgraph or is joined by an edge to a node in the subgraph. According to this definition, subgraphs with any number of isolated nonselected nodes are equally termed not dense. In practice the number of isolated nodes may however make a difference. In the case of a neutral network (viewed as a subgraph of a generalized Hamming graph) nonfolding nodes which have a folding neighbour can play a role in the dynamics of a population on the network (they become populated by off mutations, and, if they have more than a single folding neighbour, they might contribute a new sequence on the network by an on mutation). It is the number of

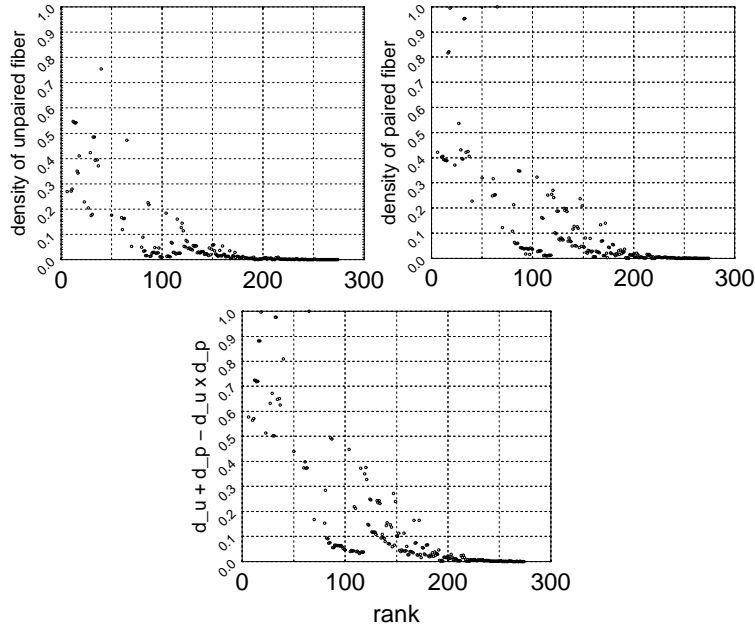


Figure 23: Statistical density values of 220 networks of Q_{AUGC}^{16} . The statistical density is the fraction of compatible sequences *not* on the network which have at least one neighbour which is on the network, out of 2500 compatible sequences not on the network per structure. In the lower plot the densities of the paired and unpaired fiber are combined into a joint density (the probability of a sequence not on the network to be either in the boundary of the paired fiber or in the boundary of the unpaired fiber). For all tested networks from rank 40 on, less than half of the non folding compatible sequences are in the boundary.

such nodes which, together with the population parameters, determines the strength of this effect. Therefore it makes sense to view density as a continuous measure on the interval $[0,1]$, rather than as the boolean property which it is according to the graph theoretical definition. In the following, *sample density* will refer to the fraction of nodes in a sample of nonfolding sequences which have a neighbour on the network (in the hypothetical case in which there is no nonfolding sequence, it is set to 1.0).

Sample density depends on whether or not the event “no neighbour is on the network” is likely for a nonfolding sequence, given the probability δ of a neighbour to fold. For most of the common networks (up to rank 70) the expected number of neighbours exceeds 1 for both the unpaired and the paired fiber. Accordingly, sample density takes the highest values in this regime (see figure 23).

Not distinguishing between unpaired and paired neighbours (lower panel of the figure), the first structure for which less than 50% of all tested nonfolding sequences are in the boundary occurs at rank 50 (the individual sample densities of the unpaired and paired fiber are lower). It is the common structures for which the condition $\lambda_u > 1 - \alpha^{-1}\sqrt{\alpha^{-1}}$ and $\lambda_p > 1 - \beta^{-1}\sqrt{\beta^{-1}}$ is true. The random graph model predicts them to be dense in the graph theoretical sense, which equals a sample density of 1.0. Obviously this does not hold, despite of the relatively high sample densities. The finite chain length is one possible reason. It can also be argued that no network of \mathcal{Q}_{AUGC}^{16} can be truly dense, because the overwhelming count of the open structure has to be accommodated somewhere. In fact the open structure is represented in nearly all tested 1-neighbourhoods of sequences of most structures (see below). Of course the prediction of the model strictly applies only for infinite sequence length. Between ranks 70 and 200, the sample densities decay to values less or equal than 0.02. The structures fall into five groups according to their number of base pairs, each of which shows a characteristic course of the decay (densities of structures with more base pairs being shifted to higher values). Beyond rank 200, densities have come very close to zero. These networks constitute small, isolated islands in sequence space.

3.5.3.4. Boundaries

In the presence of neutrality, the sequence structure mapping is from an “extended point” in sequence space (the neutral network) to a single point in shape space. The benefits of neutrality are to be found both in the increased volume of the “extended point” (increased stability of phenotypes against mutations in the genotypes) and in its increased surface (compared to a single point). The latter is called the *boundary* of the neutral network and consists of those structures which are realized by nonfolding 1-neighbours of the network’s members. If there is selection on a change of structure, it is the boundary which can be directly reached (larger boundaries therefore mean more evolutionary plasticity). Walter Fontana and Peter Schuster [31] have proposed to base

a topology on shape space on this relation of accessibility in sequence space: structure A is considered to be the more close to structure B the more frequently A is found in the 1-neighbourhoods of sequences on network B.

If all structures in the landscape were contained in the boundary of any structure, the discrete topology would result (all shapes are neighbours to each other). This constitutes an extreme of the generic property “shape space covering” of folding landscapes (see section 2.6.3). The proposition that it is true has been called the *adjacency conjecture* by the above authors. It means that on a path from any random genotype to any desired structure only a single change of structure is required (of course the path may include drift on the network of the initial phenotype and therefore may be longer than a single step). The authors find that in the space \mathcal{Q}_{GC}^{25} the conjecture holds for coarse grained structures but not for fine grained ones.

We have estimated the one point mutation boundaries of 199 networks of \mathcal{Q}_{AUGC}^{16} as well as the boundaries with respect to a single compatible move (point or pair mutation) for 190 networks of said landscape, by explicitly folding the neighbourhoods of 1000 sequences on each of the networks. In both cases two measures of closeness of a reference structure α to a structure β in the boundary have been computed in accordance with [31]: The number of neighbourhoods in which β is represented at least once, $N(\beta, \alpha)$, as well as the total number of occurrences of β across all neighbourhoods, $N_t(\beta, \alpha)$.

Fig. 24 presents the distributions $N_t(\beta, \alpha)$ (upper panel) and $N(\beta, \alpha)$ (lower panel) for the three reference shapes ranked 6, 18, and 220 according to network size. Each distribution is given for the point mutation boundary (filled circles) as well as for the single compatible move boundary (open circles). Two points can be made regarding these examples. First, the sigmoidal shape of the distributions on a double logarithmic scale resembles that of the “visibility” $B(s_1, s_i)$ defined in section 3.4.4. Thus the latter is not only a limit distribution for the boundary, but it also predicts some of its global features. Second, from the examples one would conclude that there are at least

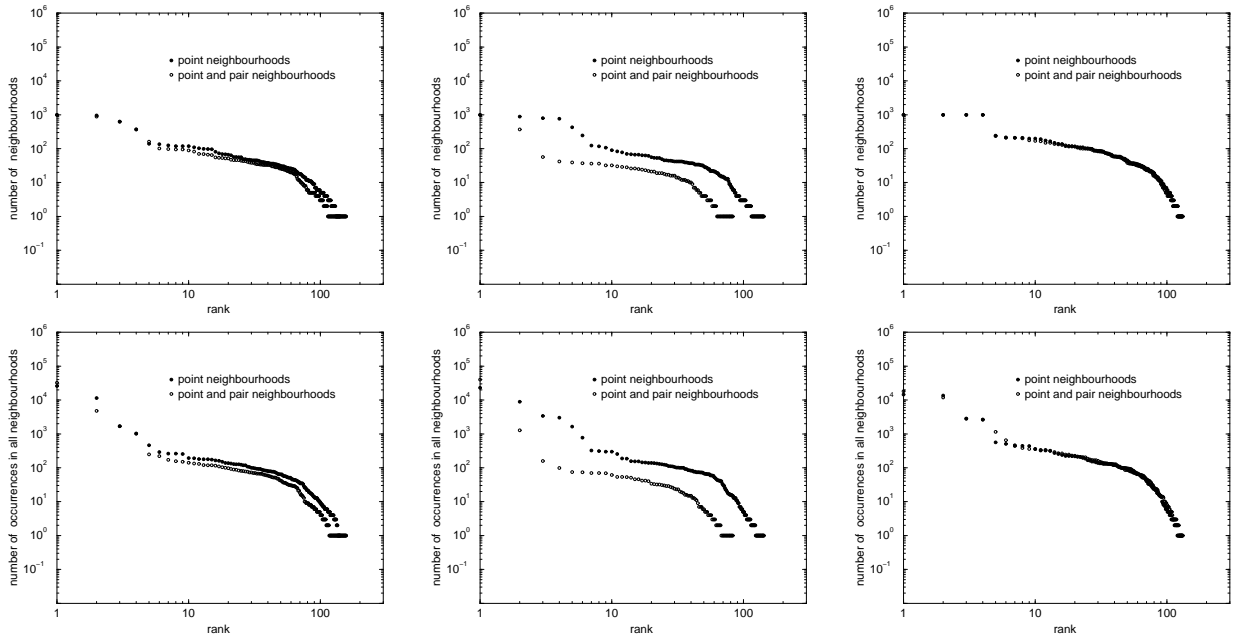


Figure 24: Columns Reference shapes. From left to right: $..(((...))....$ (rank 6 according to network size), $...((((...))))$ (rank 18), $...((.(....).))$ (rank 220). **First row** $N_t(\beta, \alpha)$. Filled circles: point neighbours. Open circles: compatible move neighbours. **Second row** $N(\beta, \alpha)$. The two definitions of neighbourhood are marked as in the upper panel. $N_t(\beta, \alpha)$ is very similar to $N(\beta, \alpha)$ in these examples. Both contain some overrepresented neighbours as well as a tail of rare neighbours. Structure #18, with 5 base pairs, shows a noticeable offset of the point and compatible move distributions.

as many structures in the point boundaries as there are in the compatible move boundaries. This is consistent with the fact that a given sequence has more single point neighbours than single compatible move neighbours, the difference being the more pronounced the more base pairs there are in the structure. Indeed the offset is largest for structure # 18, which contains 5 base pairs.

Tables 2 to 4 compile information on the cardinality of the point mutation boundaries (simply called “boundaries” from here on) and the size of the intersection of pairs of boundaries, for 30 selected networks (lack of space precludes a listing of all pairwise intersections). Of the sample networks, 15 belong to the common structures, while the remaining 15 fall into the medium to small size range. Tables 2 and 3 are within group comparisons for the common and rare sets of structures. There are three different kinds of information in these tables. The diagonal consists of

pairs of numbers $x : y$, where y is the cardinality of the boundary and x is the number of common structures contained in it (a common structure being defined as one whose rank according to network size is less or equal than 70). The upper half matrix lists the intersection sizes for the full boundaries of two shapes i, j . In the lower half matrix, the number of shared common structures is given. Table 4 presents a cross-comparison of the two groups. The matrix is not symmetric in this case, so that both the number of common structures in the intersection and the cardinality of the full intersection need to be given for each entry. The two measures are again separated by a colon. The entries $m_{i,j}$ of all tables have to be put into the perspective of the diagonal elements $m_{i,i}$ and $m_{j,j}$, which they cannot exceed. The relative intersection sizes $m_{i,j}/m_{i,i}$, $i \neq j$ are summarized at the bottom of each table by their range and mean.

On average there are 154 structures in in the boundaries of the sample common structures. The boundary sizes do not correlate with the rank according to network size. We will see below that it is not so much the size but the composition which is different in rare structures. These show a markedly reduced cardinality of the boundary only from rank 200 on (mean 109 for the 10 structures beyond this rank). The estimated boundary sizes of all sample structures are much smaller than the number of structures in the space (274). If we had indeed enumerated all structures in the boundaries (rather than testing a small sample of neighbourhoods, as we did) this result would already imply that the adjacency conjecture in its most comprehensive form is false in this landscape. The counts of rare structures in the boundaries are most sensitive to sampling fluctuations. Thus one can expect to get a more robust result if they are neglected (lower half matrices of tables 2 and 3). Common and rare networks behave differently in this respect: in the boundaries of common networks, nearly all common shapes are found (between 65 and 70 of 70, with a mean of 67.9) and nearly all of them are shared by all pairs of boundaries (mean percentage shared 97%). Thus the adjacency conjecture is *true* for common shapes. Rare networks are still adjacent to 54.7 common shapes on average, the mean percentage shared between boundaries being

equal to 82%. Although these are quite high figures, they are low enough to suggest that some of the common structures are genuinely inaccessible from networks of rare structures (rather than just having been missed in the samples). The picture is different if all neighbouring shapes are considered (with the reservations mentioned above). The mean percentage shared is now 82% for common networks and only 63% for rare ones. Thus the differences between boundaries are due to “endemic” rare neighbouring shapes, and this is more pronouncedly true in rare networks. The total boundary size of such a network need not be reduced: with a count of 153 it is even unusually high in structure #190, but nevertheless the average number of common structures in an intersection with any other boundary in this group is only 47 (73 if all structures are counted), very different from the situation with common reference shapes.

Table 4 lists the number of structures shared by the intersections of the boundary of a common network and that of a rare one. They are dominated by properties of the rare partner: it determines the number of common shapes in the intersection (the common partner mostly possesses all of them anyhow) and, depending on how special its rare neighbouring shapes are, also the total size of the intersection. The result is a wide range of percentages of shared structures: network i shares close to 100% of its common (and often also rare) neighbours with network j if i is the rare partner, but only about half of them if it is the common one.

Concluding this section on boundaries, we have a look at the open structure, which is special because of the huge size of its network. The rank of this structure in both $N(\beta, \alpha)$ and $N_t(\beta, \alpha)$ is sharply bimodal: for most networks it is present in nearly all neighbourhoods examined and reaches a total count close to or even exceeding the count of the reference structure itself, while in 16 networks it is never encountered in the boundary samples. The picture is the same whether point or compatible move boundaries are considered. Surprising as this bimodality is at first glance, it makes sense if the identity of the 16 structures is taken into account: these are exactly those structures which contain more than a single structural element (of the ones on which there are

boundary data). Any mutation will affect only one of the elements. Retaining the other (which is located towards the end of the chain and on unfolding leaves an relatively long contiguous unpaired fragment, favorable in terms of entropy) seems to be preferred over total unfolding in all cases (this can be seen from the fact that the neighbouring shapes which rank highest according to $N(\beta, \alpha)$ and $N_t(\beta, \alpha)$ are those in which one of the structural elements is deleted, the other one being either exactly retained or slightly modified – data not shown). In the networks in which the open structure is represented in the neighbourhoods, it ranks slightly higher according to $N_t(\beta, \alpha)$ than according to $N(\beta, \alpha)$, which means that if it occurs in a neighbourhood then there is more than a single instance of it. For both measures it ranks higher in point boundaries than in compatible move boundaries, this difference being more pronounced in the case of $N_t(\beta, \alpha)$. Recall that the result of a compatible mutation by definition is able to adopt the reference structure, while a point mutation in a base pair can produce a sequence which for combinatorial reasons is unable to form a stable structure. Therefore complete unfolding is expected to result more often from mutations of the latter type.

	<i>6</i>	<i>11</i>	<i>14</i>	<i>16</i>	<i>18</i>	<i>23</i>	<i>27</i>	<i>30</i>	<i>34</i>	<i>37</i>	<i>40</i>	<i>50</i>	<i>60</i>	<i>63</i>	<i>70</i>
<i>6</i>	67:157	110	128	132	125	140	137	142	121	125	117	133	132	134	127
<i>11</i>	62	65:133	115	116	113	118	121	119	109	108	116	119	116	116	111
<i>14</i>	65	63	68:150	135	129	132	132	133	114	124	116	130	130	128	124
<i>16</i>	67	65	68	70:161	130	140	138	142	115	125	122	138	136	134	130
<i>18</i>	66	64	67	69	69:143	129	131	129	113	118	115	130	127	123	121
<i>23</i>	66	64	67	69	68	69:164	146	144	123	126	123	144	134	136	135
<i>27</i>	67	65	68	70	69	69	70:164	141	125	130	119	140	135	140	135
<i>30</i>	66	64	67	69	68	68	69	69:171	124	128	126	138	137	139	136
<i>34</i>	64	62	65	67	66	66	67	66	67:140	111	108	122	118	119	115
<i>37</i>	62	60	63	65	64	64	65	64	63	65:147	115	122	124	125	119
<i>40</i>	64	65	65	67	66	66	67	66	64	62	67:136	120	118	119	114
<i>50</i>	66	64	67	69	68	68	69	68	66	64	66	69:164	136	136	129
<i>60</i>	64	62	65	67	66	66	67	66	64	62	64	66	67:163	132	129
<i>63</i>	64	62	65	67	66	66	67	66	64	62	64	66	64	67:164	136
<i>70</i>	66	64	67	69	68	68	69	68	66	64	66	68	66	66	69:159

	minimum (at)	mean	maximum (at)
all	0.69 (<i>11,30</i>)	0.82	0.93 (<i>30,40</i>)
common	0.92 (<i>11,37</i>)	0.97	1.00 (<i>6,16</i>)

Table 2: Intersections of point mutation boundaries. Common vs. common structures. Italic labels on rows and columns denote the ranks according to network size of the structures which are compared. Upper half matrix: total size of the intersection. Lower half matrix: number of structures in the rank range [1..70] which are in the intersection. Diagonal: in a pair of numbers $x : y$, x denotes the number of common structures which are in the boundary of this rank, while y is the full cardinality of the boundary. The small table below gives the range and mean of the relative sizes of the intersections, $m_{i,j}/m_{i,i}$, $i \neq j$.

	150	160	170	180	190	200	210	220	230	240	250	252	253	254	255
150	63:142	101	116	86	98	88	76	93	73	76	82	105	73	58	77
160	57	64:149	101	95	108	96	84	96	78	74	89	90	82	71	72
170	59	55	61:144	86	98	89	76	94	77	76	84	100	77	60	83
180	55	56	53	62:128	93	75	78	88	58	68	79	76	73	56	65
190	52	57	52	51	58:153	92	78	97	66	83	85	83	86	69	80
200	52	57	52	52	53	58:115	71	81	63	69	73	77	66	60	66
210	46	47	47	53	44	46	53:116	91	59	51	78	69	75	57	73
220	51	53	51	51	50	49	48	57:132	67	63	84	82	82	64	89
230	49	51	49	44	47	45	36	43	52: 89	42	57	64	53	47	57
240	38	38	39	35	37	37	30	34	33	39: 97	61	64	55	50	52
250	52	53	52	56	49	49	50	52	42	33	59:110	70	80	59	70
252	56	51	55	48	47	46	41	46	45	37	46	56:122	62	51	64
253	42	46	42	45	41	43	42	43	36	28	46	36	49:111	53	80
254	37	41	38	38	38	39	35	39	34	28	39	35	31	41: 84	52
255	45	45	45	43	42	43	42	48	37	28	44	40	41	32	49:110

	minimum (at)	mean	maximum (at)
all	0.41(150,254)	0.63	0.88(160,230)
common	0.56(240,250)	0.82	1.00(150,252)

Table 3: Intersections of point mutation boundaries. Rare vs. rare structures. The structure of the matrix and the summary table are the same as in table 2.

	150	160	170	180	190	200	210	220	230	240	250	252	253	254	255
6	60:112	61:121	58:115	59:104	55:114	55: 91	50: 92	54:108	52: 83	37: 76	56: 93	53: 96	48: 88	38: 68	46: 81
11	58:103	59:104	56: 99	62:106	53: 93	55: 86	53: 86	52: 96	47: 69	37: 67	56: 87	51: 85	47: 81	39: 60	44: 77
14	61:109	62:118	59:111	60:104	56:106	56: 90	51: 91	55:105	50: 76	39: 76	57: 98	54: 94	47: 90	40: 67	47: 80
16	63:116	64:123	61:118	62:108	58:116	58: 95	53: 97	57:108	52: 80	39: 76	59: 98	56: 98	49: 91	41: 71	49: 85
18	62:110	63:114	60:114	61:107	57:103	57: 92	52: 87	56:105	51: 75	38: 76	58: 94	55: 97	49: 88	40: 65	49: 83
23	62:120	63:124	60:116	61:112	57:114	57: 94	52: 98	56:112	51: 80	38: 76	58: 96	55:103	48: 93	40: 71	48: 85
27	63:123	64:125	61:120	62:107	58:116	58: 96	53: 98	57:114	52: 82	39: 78	59: 99	56:101	49: 96	41: 70	49: 91
30	62:117	63:123	60:116	61:111	57:122	57: 96	52: 93	56:113	51: 77	38: 79	58: 97	55: 99	48: 91	40: 73	49: 88
34	60:106	63:114	58:108	59: 94	56: 93	56: 85	50: 81	54: 96	51: 81	37: 61	56: 83	54: 92	46: 79	41: 60	46: 75
37	58:105	64:125	56:106	57: 96	58:113	58:100	48: 83	54:103	51: 78	39: 74	54: 92	51: 89	47: 91	41: 72	46: 80
40	60:102	61:111	58: 99	62:107	55:102	56: 91	53: 85	54:100	49: 72	39: 71	57: 89	53: 84	48: 81	40: 64	46: 74
50	62:118	63:121	60:115	61:113	57:109	57: 95	52: 92	56:109	51: 79	38: 77	58: 95	55:103	48: 90	40: 68	48: 81
60	60:114	61:123	58:120	59:109	56:116	55: 98	50: 92	54:106	50: 75	39: 81	56: 93	54:104	46: 93	39: 65	46: 82
63	60:114	61:126	58:111	60:106	56:111	55: 96	52:103	57:115	49: 78	37: 75	57: 96	53: 98	47: 90	41: 78	48: 84
70	62:114	63:118	60:111	61:104	57:110	57: 87	52: 94	56:107	51: 78	39: 75	58: 89	55: 96	48: 89	40: 68	48: 86

	minimum (at)	mean	maximum (at)
all	0.40(60,254)	0.71	0.93(6,230)
common	0.55(18,240)	0.88	1.00(6,230)

Table 4: Intersections of point mutation boundaries. Common vs. rare structures.

3.6. Energy Landscapes

In a static folding landscape, each sequence is uniquely assigned to a shape. Once this assignment is done, it is the shapes (phenotypes) which are evaluated by the fitness function. Sequences which fold into the same shape have the same fitness by definition.

Minimum free energy folding selects the most stable one among the possible structures of a sequence. The exact energy of this “best” structure may well be very different for different sequences. There is no a priori reason to expect that members of the same neutral network are more similar in this respect than random sequences.

In an *energy landscape* over a sequence space, fitness is directly equated with free energy, without explicit reference to shapes [27, 6]. The focus in studies on energy landscapes is less on neutrality but on the degree of ruggedness of the landscape, which determines whether or not very stably folding sequences, given they exist, can be found in reasonable time from a random starting point. One way to characterize ruggedness is by way of the autocorrelation of the energies found on a random walk. In complete sequence spaces, autocorrelation $\rho(d)$ has been found to decay exponentially with distance d [80]. The distance l at which it holds $\rho(l) = 1/e$ is called the *characteristic length* of the landscape.

In the following we combine the two approaches by asking for the *distribution of minimum free energies within a neutral network defined by a shape*. The degree of ruggedness of such a per-network landscape determines how easy it is to optimize the stability of a structure without having to pass through intermediates with a different structure. We have probed the landscapes of the 220 indexed networks by 5 self avoiding random walks of 2000 attempted steps each. Each walk was characterized by its mean energy and variance, as well as by the autocorrelation at offsets 1 to 6, and the distribution of *runs*. A run consists of a series of contiguous steps during which the binned energy values are constant (that is, the raw values are confined to a window of the bin size).

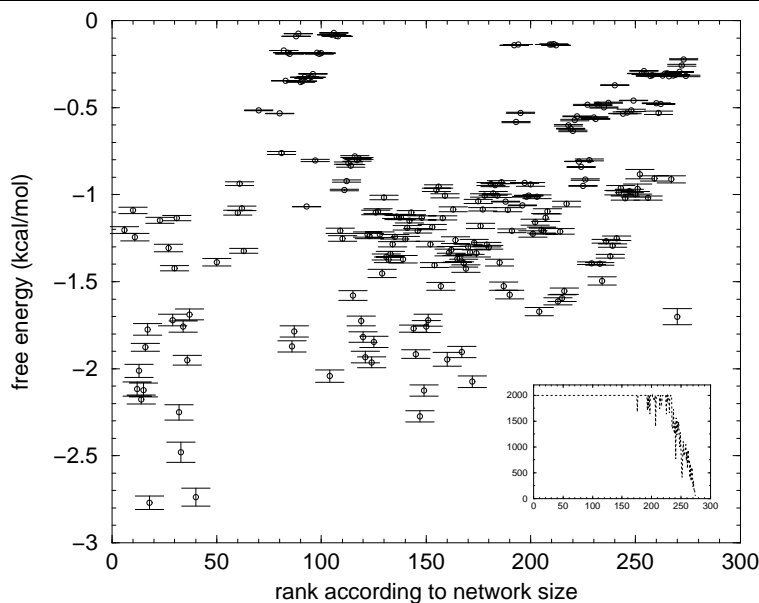


Figure 25: The means (data points) and standard errors (error bars) of the free energy on self avoiding random walks. Data on 5 random walks per network are averaged. The average lengths of the 5 walks are plotted in the insert: up to rank 235, the maximum length of 2000 steps is nearly always reached. The lengths decrease rapidly (and linearly) for smaller networks.

3.6.1. Mean and Variance of Energies

Fig. 25 presents the means and standard errors of free energies encountered on a random walk of maximal length 2000. The data are averaged from 5 random walks per network. Not all walks reach the maximum number of steps: the mean number of actual steps in the 5 instances is displayed in the insert of the figure. Overall there is a tendency for big networks to have the larger absolute values of free energy, but this is far from being a lawlike relation. The standard errors of the energies do not correlate at all with size, but rather with the free energy itself, such that more *stable* structures show the *larger* variation. This is counterintuitive at first, but it can be understood by the following line of reasoning: if the typical free energy is already very small (in absolute terms), then outliers in the direction of even smaller energies are very unlikely (they will probably switch structure instead). From a baseline at a large (absolute) value of free energies outliers may occur in both directions, doubling the possible variance.

Why is there such a poor correlation between free energy and size of the network ? We have argued above that it is the folding probability rather than the number of compatible sequences which dominates the observed size in the networks of Q_{AUGC}^{16} . The folding probability of a sequence for a structure of course depends on how much energy is gained by “choosing” the structure. But, and this is the reason why a correlation of the mean free energy and the folding probability is not a priori to be expected, the crucial measure is the energy gained *relative to other structures with which the sequence is compatible*. In order to fold with a high probability, a sequence must not be compatible with other structures which are even more stable. The very highest ranking structures in Q_{AUGC}^{16} combine three advantageous properties: they are not too unstable (no very small absolute free energies), nevertheless they do not have too many base pairs (which would restrict the number of compatibles), and, for reasons of the relative location in the structure, their stacks cannot be extended (reducing the competition by more stable structures).

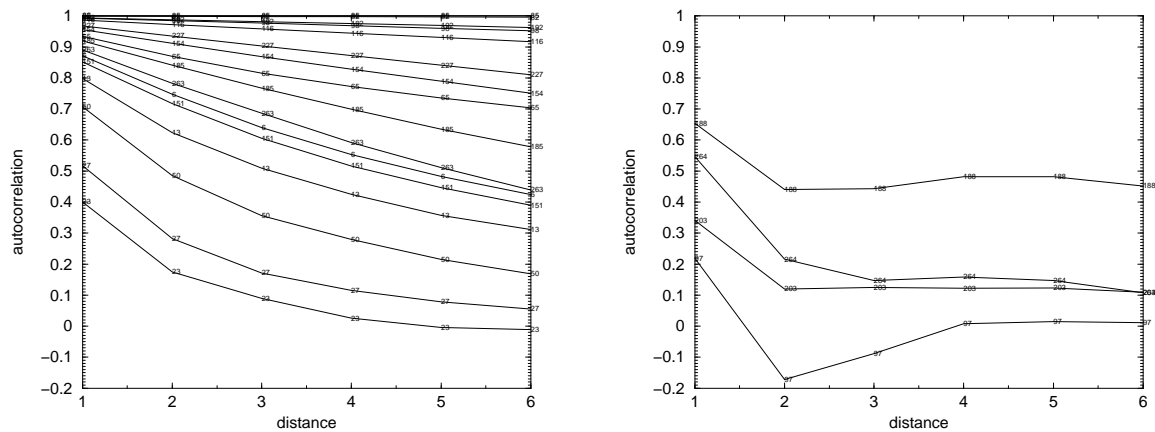


Figure 26: Correlation structures of sample per-network energy landscapes. Neither the type (exponential or other) nor the steepness of the descent with distance correlate with the rank of the network according to size (see labels on the curves).

3.6.2. Autocorrelation

The autocorrelation of a walk at offset d , $\rho(d)$, has been measured as Pearson’s correlation coefficient of the series $(w_i, w_{i+d}), i = 1 \dots n - d$, where n is the number of steps. It has been determined for offsets 1 to 6.

The networks Q_{AUGC}^{16} differ very much with respect to autocorrelation and the derived measures characteristic length (that l with $\rho(l) = 1/e$). Many of them show the expected exponential decay of the autocorrelation, albeit with very different slopes (see figure 26). In others the autocorrelation goes on a plateau already at a distance of less or equal than 4. The background autocorrelation need not be equal to zero: it will be the higher the less different energy values there are. Fig. 26 gives examples of landscapes of both types. The curves are labelled by the rank of the respective network according to size: again this parameter does not correlate with the slopes. Thus, among both large and small networks there are those in which optimization of stability is easy and those in which it is not.

Fig. 27 displays the distribution of the characteristic lengths for all 220 networks. The majority of

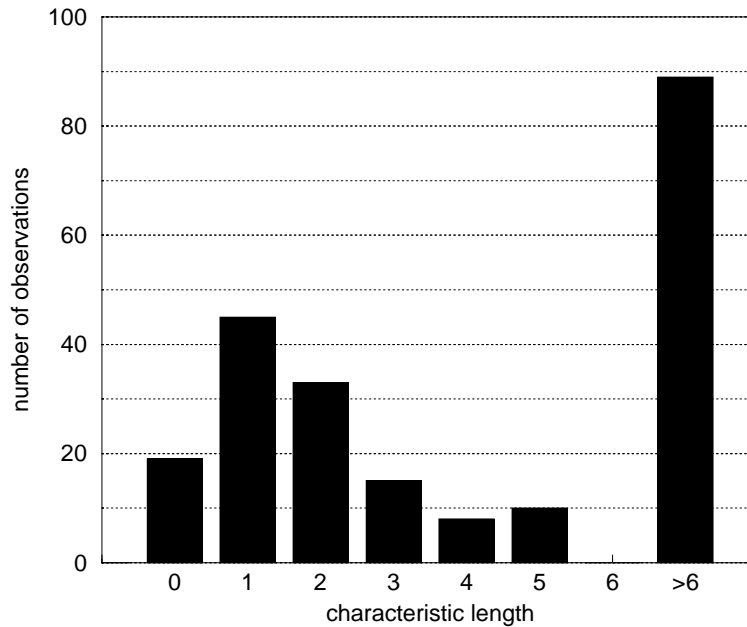


Figure 27: Distribution of characteristic lengths of the energy landscapes of 220 networks. The majority of the characteristic lengths exceed 6. It cannot be decided whether this peak is due to a second mode at longer lengths or whether the individually small frequencies in the right tail add up to it. If the latter is true, then the most frequent individual length is actually short (2).

the networks show quite smooth energy landscapes (characteristic length greater than 6), yet there are also very rugged ones. Thus there is indeed information gained by examining the per-network landscapes instead of averaging over the entire sequence space.

3.6.3. Run Lengths

Landscapes with a high autocorrelation will also show long *runs* in random walks. A run is a contiguous sequence of steps during which the free energy does not leave a certain window. While autocorrelation is a single number associated with a random walk or an entire landscape, the distribution of run lengths encountered on a walk gives a more resolved picture of long range variation of ruggedness. By tuning the window size, one can examine ruggedness on different scales.

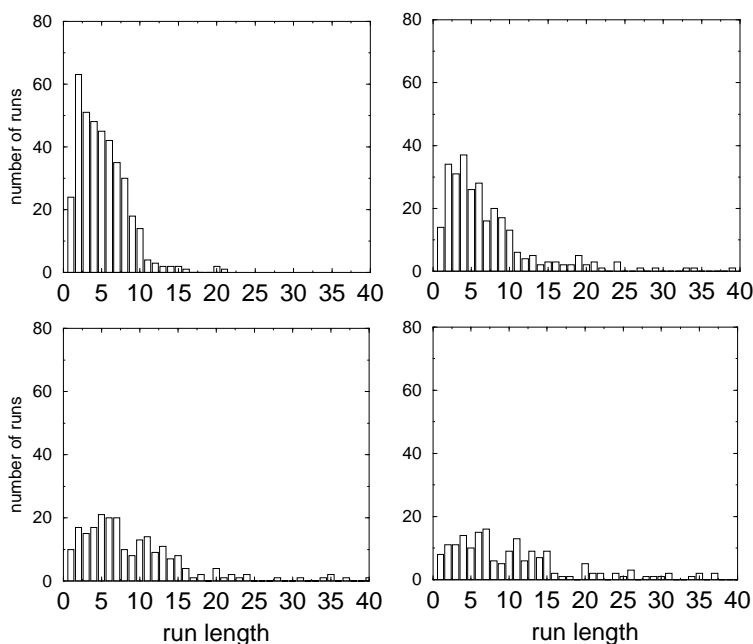


Figure 28: Distribution of runlengths encountered on a random walk in the energy landscapes of the networks ranked 6 (upper panel) and 10 (lower panel) according to network size. Left: window size 0.1 kcal/mol. Right: window size 0.5 kcal/mol.

The very fact that a whole distribution is associated with each network precludes a comprehensive presentation of all of them. Instead we will discuss some examples.

The first example deals with the pair of structures $'..(((...))).....'$ and $'..(((...))).....'$, respectively, which are ranked 6 and 10 according to network size. Both structures are part of the neutral network of the substructure $'(((...)))'$, the “simple hairpin” discussed in chapter 5. The autocorrelation of the free energy decays exponentially for both networks, the characteristic length exceeding 6 in both cases (data not shown). Yet it is already evident from the autocorrelation data that the two landscapes are not identical: $\rho(6) = 0.42$ for structure #6, while $\rho(6) = 0.7$ for structure #10. The distributions of run lengths (see figure 28) not only confirm this difference, but allow to describe it in more detail. The most probable run length is shorter for network #6 than for #10 (2 and 4 for the two window sizes, versus 5 and 7), as expected from the autocorrelation data. What

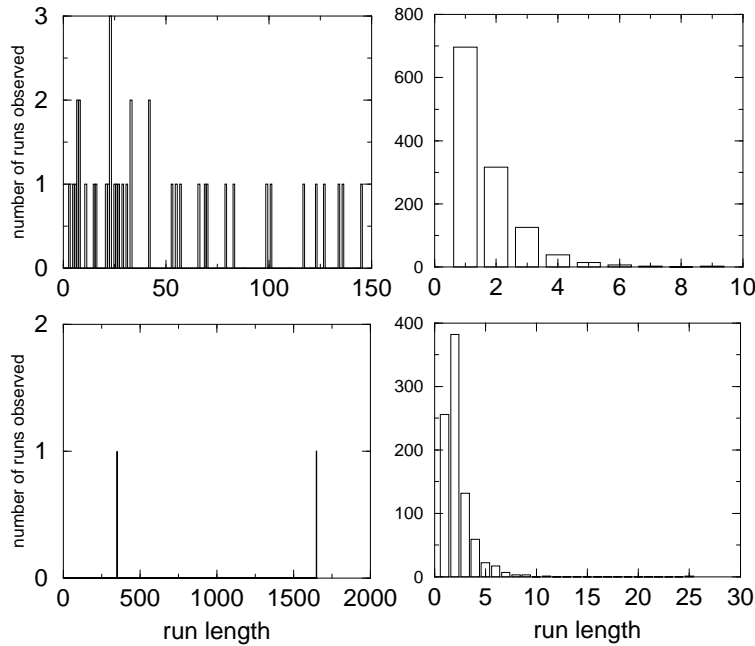


Figure 29: Distribution of runlengths encountered on a random walk in the energy landscapes of the structures
a) rank 30(((.....))) (upper left) b) rank 31 (((.....))).... (upper right) c) rank 82((.....))
(lower left) d) rank 97 ((.....))..... (lower right)

is not implicit in the autocorrelation is the fact that the higher variability of free energy values in network #6 to a good deal is due to small fluctuations, leading to noticeably longer runs when the window size is changed from 0.1 to 0.5 kcal/mol. In network #10, enlarging the window size has a much smaller effect. Thus in this case the runs at the smaller window size must have already been separated by relatively large jumps in energy. Another distinguishing feature of network #10 is the bimodality of the distribution (at both window sizes). This landscape seems to contain a separate regime in which the probability for longer runs is high. One feature is shared by both networks: network #6 at 0.5 kcal/mol, as well as network #10 at both window sizes, show an extended tail of long runs. It is possible (although not likely from a random starting point) to stay on a neutral path with respect to energy for up to 40 steps.

The differences between the landscapes of networks #6 and #10 are surprising because these

structures are so similar. Yet they are actually small compared to the differences between the landscapes of other networks. We conclude this section with four extreme examples (see figure 29).

The window size is 0.5 kcal/mol in all cases.

Structures #30 and #31, respectively, constitute another case of “similar structure but (very) different energy landscapes”. These structures are mirror images of each other. The runs encountered on a random walk of length 2000 in the energy landscape of #30 are of extremely variable length, most individual lengths occurring only a single time. The spectrum of run lengths ranges from one to 150. In the landscape of #31, on the other hand, nearly all runs are of length one. This very rugged landscape contrasts with that of #82, which is nearly completely flat (there are only two runs in the entire walk of length 2000). Structure #82 is one of the 2-stacks with strong constraints on the sequence level. With a fixed sequence composition of the stack, one would indeed expect similar free energies. Yet structure #97, which also consists of a 2-stack, behaves again differently: the most probable run length is 2, runs of length 6 being already extremely rare. In this case the difference can be explained: structure #97 possesses a tetraloop, with corresponding sequence constraints in the loop region (see conservation profile in the appendix). Whenever adjacent steps in the walk differ by a mutation in the loop, free energy is likely to change.

Summarizing, there is unexpected diversity among the energy landscapes of the networks of Q_{AUGC}^{16} . The neutral networks are definitely not random samples from the entire space with respect to their energy distributions. Yet the network specific energy landscapes are not always especially suited for optimization of stability by a gradient search. Depending on the relative importance of retaining a structure and stable folding, detours via other networks may be preferred by an evolving population in the space.

4 Probing Large Neutral Networks

4.1. Introduction

We have noticed in the preceding chapter that a sequence length of 16 is about the upper limit for an exhaustive enumeration of the entire sequence space, using the natural alphabet. Thus, for any biologically meaningful structure, the length of which probably will far exceed 16, there will be no complete information on global properties of its neutral network.

Global properties can both affect evolutionary processes on the network and can in principle be exploited by such processes, if the system is able to perceive them. While natural populations are quite blind in this respect, a human experimenter doing irrational design of sequences could adjust parameters like mutation rate and population size to, for example, the degree of connectivity and denseness of the network, would he know them.

We present an algorithm, which, given a network conforms to the random graph model presented in chapter 2, enables us to estimate whether it is a connected graph, using only a small sample of sequences from the network. The actual test for the existence of a connection between two sequences is linear in time in the sequence length (the entire algorithm in its present form is not, for reasons discussed below). The test presupposes a certain property of networks which are outcomes of the random graph process described in section 2.6.3.2 at superthreshold choosing probabilities, namely, local connectivity. It is not clear from the beginning that minimum free energy networks of finite length sequences will have this property. We show however that it holds well enough for networks in the space Q_{GC}^{30} of RNA sequences of length 30 over the alphabet $\{G, C\}$ in order to indeed predict connectivity from small sample sizes. In addition we propose a

method which can aid the decision in less well locally connected networks and discuss the relation of features of the underlying RNA secondary structure and the degree of local connectivity of the network (which can possibly be exploited for deciding beforehand how well the algorithm is going to work on a network).

4.2. Algorithm

4.2.1. Basic Idea

A graph is defined to be connected if there exists a connecting path between all pairs of its nodes (see sect. 2.2). Thus the number of components can in principle be estimated from a statistical sample of all pairs of nodes. In a general graph this is however precluded by the fact that adjacency information on all nodes is needed in order to determine whether a given pair is connected. Christian Reidys proved that in the limit of infinite sequence length n a subgraph which is randomly induced in \mathcal{Q}_α^n with a choosing probability $\lambda > 1 - \alpha^{-1}\sqrt{1/\alpha}$ possesses the following property:

Definition 11. (Locally connected graph). *A graph G is called locally connected if for any two vertices $v, v' \in v[G]$ there exists at least one path fulfilling the following criteria: the path starts at a vertex $v_1 \in B_1(v)$ and ends in $v'_1 \in B_1(v')$ ($B(v)$ being the set of neighbours of v) and it is completely contained in a cylinder of radius 1 with axis endpoints at v_1 and v'_1 . (For a formal definition see Reidys et al. (1997) [71].)*

The property is preserved if the underlying graph is a generalized Hamming graph $\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p}$ (because it holds both in $\mathcal{Q}_\alpha^{n_u}$ and in $\mathcal{Q}_\beta^{n_p}$ [70]) which means that neutral networks of RNA secondary structures are expected to be locally connected according to random graph theory. Thus

it should be feasible to estimate connectivity in the absence of complete sequence information on a neutral network.

4.2.2. Implementation

A pseudocode description of the connectivity test is given in table 5. It tries to connect neutral neighbours of two reference sequences v_1, v_2 by a simple backtrack algorithm with stopping points at the $ndiff$ positions at which the neighbours (n_1, n_2) differ. They are visited from left to right. The following information is available to the algorithm at each such point: first, the sequence of *intermediates* ($n_{2_1}, n_{2_2}, \dots, n_{2_{(i-1)}}, n_{1_i}, n_{1_{(i+1)}}, \dots, n_{1_{(ndiff-1)}}$), $i = 0, ndiff$, second, a current sequence *string* which is on the network and either is identical with the current intermediate or differs from it by a single compatible mutation, third, that stopping point at which the additional compatible mutation has been introduced, in case there is one (*laststop*), and fourth, a table of additional mutations which have already been tried out at each intermediate. The basic action is to flip the i th position of difference to the state which is observed in n_2 . If the result is not neutral, one of two actions is taken: if the current sequence does not contain an additional mutation (“bypass”), single mutations are randomly introduced into the current intermediate until either one is found which is both neutral and allows a flip of state without losing the structure (in this case the algorithm proceeds to the next stopping point) or all possible mutations are exhausted and it stops. If however a bypass is already present, it “backtracks” to the stopping point *laststop* at which this mutation has been introduced, setting the current sequence to the *laststop*th intermediate. The mutation tables of points $nstop > laststop$ are not reinitialized on such a backtrack, because it is already known that paths which pass through these mutations will not be successful.

The time complexity for testing the connectivity of a single pair of sequences (n_1, n_2) of length n at a fixed distance d is $O(n)$. This can be understood as follows: At every stopping point i there is only one current sequence for which there is more than a single possibility to elongate the current

path to the right, namely, the i th intermediate. This sequence can give rise to $O(n)$ different elongations. Of these, a single one is equal to the $i + 1$ th intermediate and is again free to vary, while the remaining ones are fixed. Therefore with a constant number of intermediates (fixed d) there are $O(n)$ different paths in the search space. Because the algorithm guarantees that no path is exactly repeated, it is enough to consider this upper limit (no need to exactly track the partial repetitions of paths by the backtracking). The same sequence may appear in different paths, but because the length of the paths itself is $O(d)$, which is a constant, the number of operations on sequences is still $O(n)$.

During the execution of the algorithm the function `fold(string)`, which maps the sequence *string* to its structure, is called over and over again. The time complexity of this function is $O(n^3)$ if it is implemented as an actual minimum free energy folding of the sequence[26], so that the entire algorithm cannot be linear in time in the sequence length.

4.2.3. *Delaying the Decision – Simultaneous Neutral Walks*

In a network in which one has reason to expect that local connectivity at high pair distances (which are best suited for detecting multiple components) is bad, we propose the following extension of the connectivity algorithm:

Start with a pair at a long distance. If it is not connected, choose a pair of neutral neighbors of the initial pair, and try to connect these. If the distance of the neighbours is forced to decrease, the probability of finding a connection (if one exists) will increase (see the data on the dependence of local connectivity on pair distance below). By this rule two simultaneous neutral walks in sequence space are constructed (see figure 30). They terminate for one of three reasons: a connection may be found, it may not be possible to find an elongation which meets the distance criteria, or the walk length reaches a predefined maximal length. For a connected initial pair, there is a probability to terminate the path by successful connection at every step, which should lead to shorter paths

```

connected = 0
for(n1=1 ... nneighbours(v1))      # all neutral neighbours of node v1
  for(n2=1 ... nneighbours(v2))    # all neutral neighbours of node v2
1:   intermediates[0] = n1
     for ( nstop = 1 ... ndiff)     # differences between n1 and n2
       intermediates[nstop] = intermediates[nstop-1]
       substr(intermediates[nstop],pos[nstop],1) = substr(n2,pos[nstop],1)
     end for
     string = n1; nstop = laststop = 0; state = GROUNDSTATE
     while (nstop < ndiff)
       position = pos[nstop]        # position of nstop'th difference
                                   # between n1 and n2, from left to right
       if ( (state == BYPASS) && UndoBypass(string)) then
         state = GROUNDSTATE
       endif
                                   # repair "nstop"th difference
       substr(string,nstop,1) = substr(n2,nstop,1)
       if (fold(string)) then
         ++nstop
       else
         if ( STATE == BYPASS ) then
           nstop = laststop         # return to the intermediate
                                   # at which the bypass was introduced
           string = intermediates[laststop]
         else
           if ( FindBypass(string)) then
             laststop = nstop; ++nstop
             state = BYPASS
           else
             connected = 0         # pair (n1, n2) is not connected
             break                 # next pair of sequences
           endif
         endif
       endif
     end while                       # nstop < ndiff
2:   connected = 1; STOP
     end for                         # n2 loop
end for                             # n1 loop

```

Table 5: Algorithm for testing for a local connection between sequences v1 and v2. Substr(s,i,j) returns the substring of string s which starts at position i and consists of j consecutive characters.

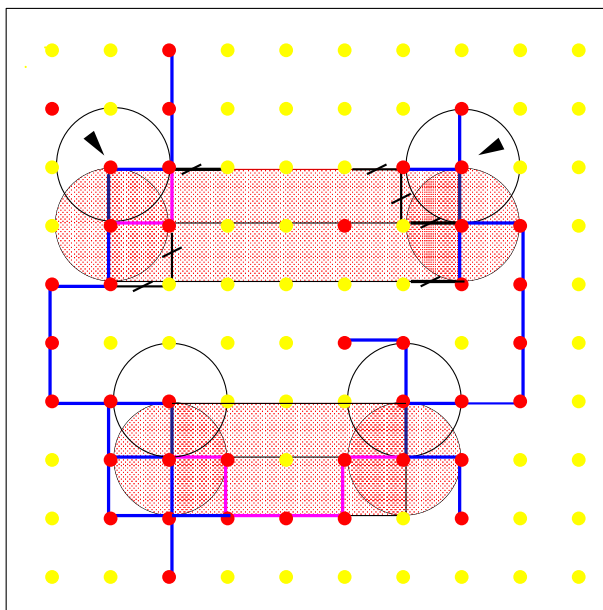


Figure 30: Sketch of a simultaneous walk of a pair of neutral sequences in sequence space: Dark grey balls denote neutral sequences, light grey ones indicate sequences which are not neutral and hence fold in structures different from the corresponding structure of the pair. Arrows show the initial pair. Open circles refer to their single error-class. If connection is not successful by a simultaneous neutral walk a new pair is chosen. This pair is again checked for connectivity.

on average. It is possible to classify the networks as connected (disconnected) by discriminant analysis if the length distribution of the paths is used as the feature vector.

4.3. Local Connectivity in Minimum Free Energy Neutral Networks of Q_{GC}^{30}

4.3.1. The Role of the Connectivity Algorithm in Implementing a Test

What can be measured in an unknown network is the number of locally connected pairs, n_{conn} , among the total number of pairs in the sample, N . Whether or not such a result can be meaningfully interpreted in terms of connectivity of the graph depends on how n_{conn}/N relates to the ratio of the number n_{conn} to the number of arbitrarily connected pairs, n_{same} . These fractions are expected to be network specific and are not known in advance (but they relate to observable

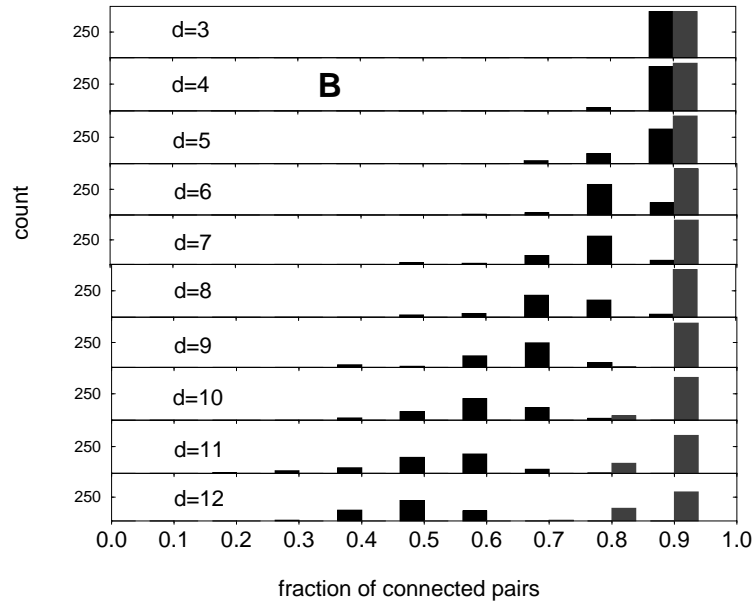
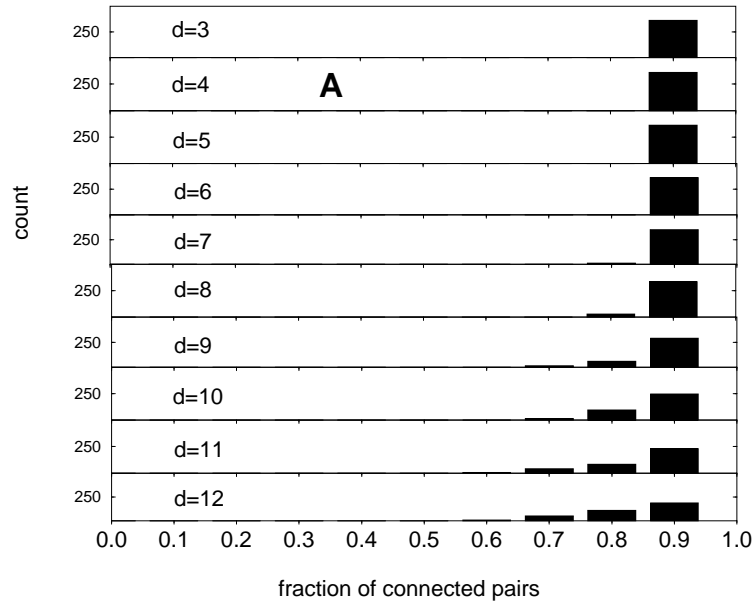


Figure 31: Distributions of the fractions n_{conn}/N (black bars) and n_{conn}/n_{same} (light bars) for test sets A and B, at different distances of the to be connected sequence pairs (for set A only n_{conn}/N is displayed, because in this set $n_{same} = N$). The number of networks in the set which show a fraction in a given bin is plotted on the y axis, while the x axis gives the fractions in bins of size 0.1. A label x at a bin denotes the interval $[x, x + 0.1]$.

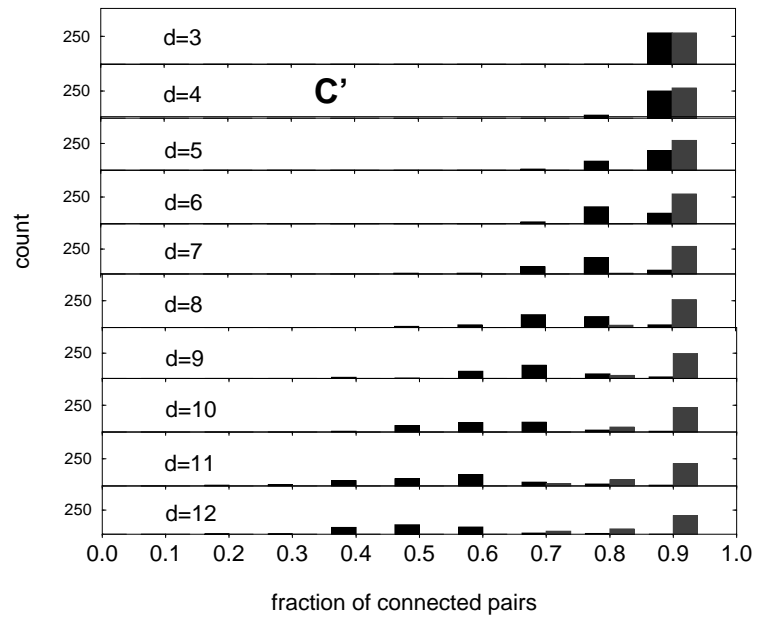
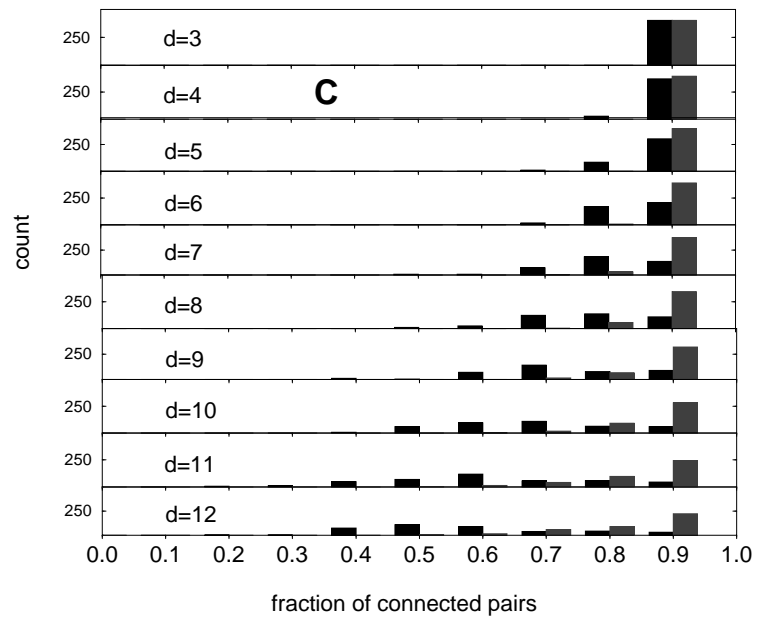


Figure 32: Distributions of the fractions n_{conn}/N (black bars) and $n_{\text{conn}}/n_{\text{same}}$ (light bars) for test sets C. Upper plot: complete set. Lower plot: networks which contain only a single big component are omitted.

properties of the reference secondary structures, see below). Assume they have been determined for a large enough number of networks so that the essentials of the distribution are most likely captured. An observed value $nconn/total$ is then interpreted with respect to this distribution: it indicates true disconnectedness with a statistical significance p which is equal to the area under the distribution's curve to the left of it (in the direction of smaller fractions). We have not pursued a rigorous treatment of this idea, but rather used an empirical threshold on $nconn/total$ in order to classify networks as connected or not (see below).

4.3.2. Test Sets

We have used three different sets of networks from the space \mathcal{Q}_{GC}^{30} , which has been exhaustively folded by Walter Grüner *et al.* (1996) [37, 38]. Set A consists of 361 connected networks with ranks according to network size in the range between 203 and 2595. The 499 networks of set B decompose into two components of about equal size. Their ranks cover the range from 149 to 2687. Set C finally is a random sample of size 401 from the rank range [417 ... 2103], containing both completely connected and very fragmented networks. All three sets are drawn from the common structures of \mathcal{Q}_{GC}^{30} , which comprise ranks 1 to 4907 [54].

The space \mathcal{Q}_{GC}^{30} has been chosen because its neutral networks have already been determined by exhaustive folding [37, 38]. Thus `fold(string)` in this case can be realized as a lookup in the output of this analysis, which because of a special data structure for storing the sequences of a network, *tries*, takes only $O(n)$ time [38].

4.3.3. Results: Simple Local Connectivity

Figures 32 and 32 summarize the rates of local connectivity in the test sets. The number of sequence pairs tested per network was 100 and 500 for sets A and B. For set C, the sample size had been adjusted so that each sample contained 250 truly connected pairs.

The figure shows two conflicting trends: sequence pairs at small distances (3 or 4) are very well locally connected ($nconn/nsame \in]0.9, 1.0]$). At these distances however the distributions $nconn/N$ and $nconn/nsame$ practically coincide: nearly all sampled pairs are in the same component even if the graph contains more than a single one, and thus do not yield information about the graph structure. The distribution $nconn/nsame$ spreads out slowly with increasing distance of the pairs, this behaviour being qualitatively identical in all three sets (the slightly larger spread in set A is probably due to the small sample size). Yet the peak of the distribution remains at the interval $]0.9, 1.0]$ in all sets up to a pair distance of 12. The peak of $nconn/N$, in contrast, quickly moves away from that interval with increasing distance if the set contains disconnected networks. It is located at $]0.5, 0.6]$ at a distance of 12 in both set B and set C. The distributions of $nconn/N$ and $nconn/nsame$ are well separated at this distance both in set B and set C (in set C there is a slight overlap of the tails). If a value of 0.75 of the observable measure $nconn/N$ is used as a threshold for prediction (labelling networks with a degree of connectivity less or equal to the threshold value as disconnected, and others as connected) the following distribution of false negatives, false positives, and success rate results for the three sets:

	false+	false-	percent correct
A	0	29	92
B	1	0	99
C	19	34	87

Thus in most networks it is possible to tell from a small sample of pairs of sequences whether the network as a whole is connected. What is not possible is to exactly determine the number of components. In figure 33 the fraction $nconn/N$ at a distance of 12 in the networks of set C is graphed against the number of components which are of size greater or equal than one third of the size of the biggest component (smaller ones are expected to be neither visible to the sampling

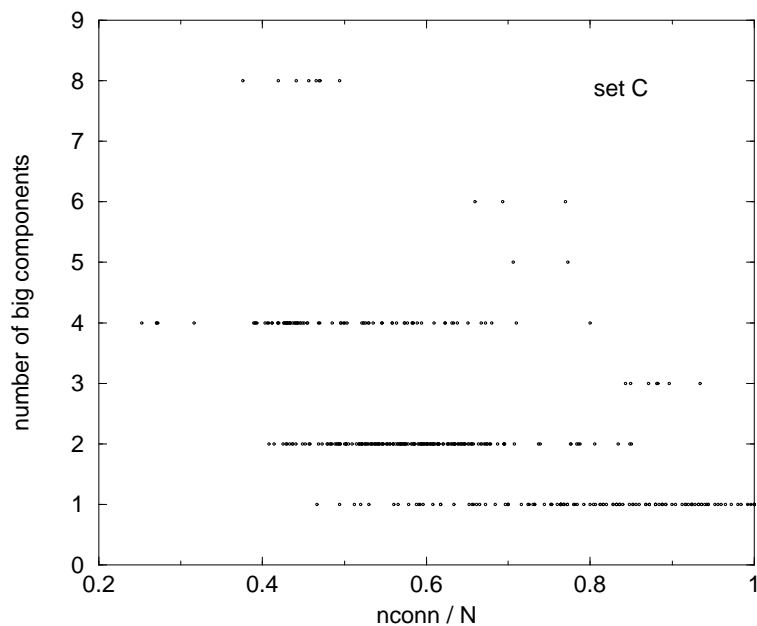


Figure 33: The fraction of locally connected pairs at a distance of 12 versus the number of big components in the networks of set C.

process nor to be of much importance for evolutionary processes on the network). We see that networks with 2, 4, and 8 big components show comparable values of $nconn/N$ (from about 0.4 to about 0.6). This means that in networks with more than two components even at a distance of 12 it is not completely equiprobable to hit any of them (otherwise the fraction should be 0.25 for a network with 4 components).

4.3.4. Results: Simultaneous Neutral Walks

The result of applying the method of simultaneous neutral walks to the networks of set C is shown in figure 34. Predicting networks on which the discriminant function takes values less or equal than 65.0 as disconnected, and others as connected, the following success rates result:

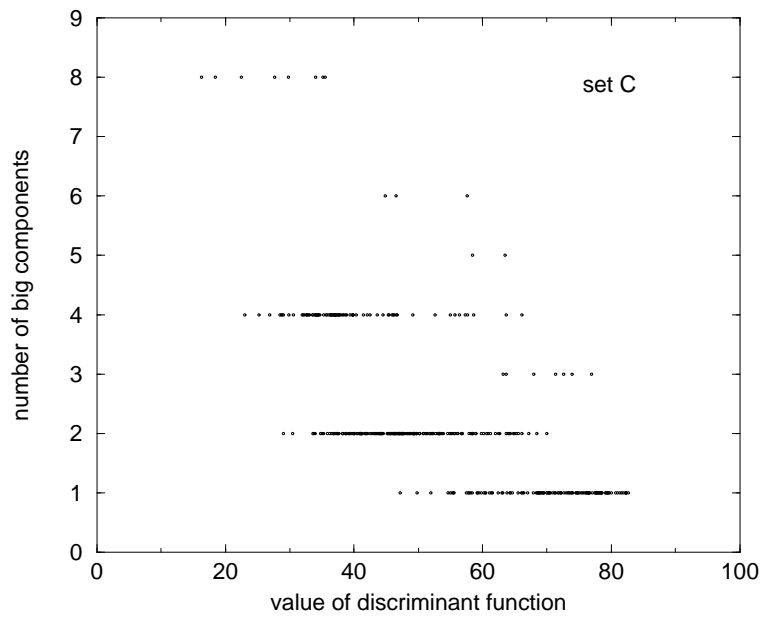


Figure 34: The fraction of locally connected pairs at a distance of 12 versus a linear discriminant function trained to distinguish between connected and disconnected networks.

	false+	false-	percent correct
C	12	30	89

In terms of the power to discriminate between connected networks and arbitrarily disconnected ones this is not much improvement over the simple test for local connectivity (which is much faster). There is however a slightly better correlation between the value of the discriminant function and the actual number of big components than is the case with simple connectivity. Perhaps one could pursue this fact in order to develop an algorithm which actually estimates the number of components.

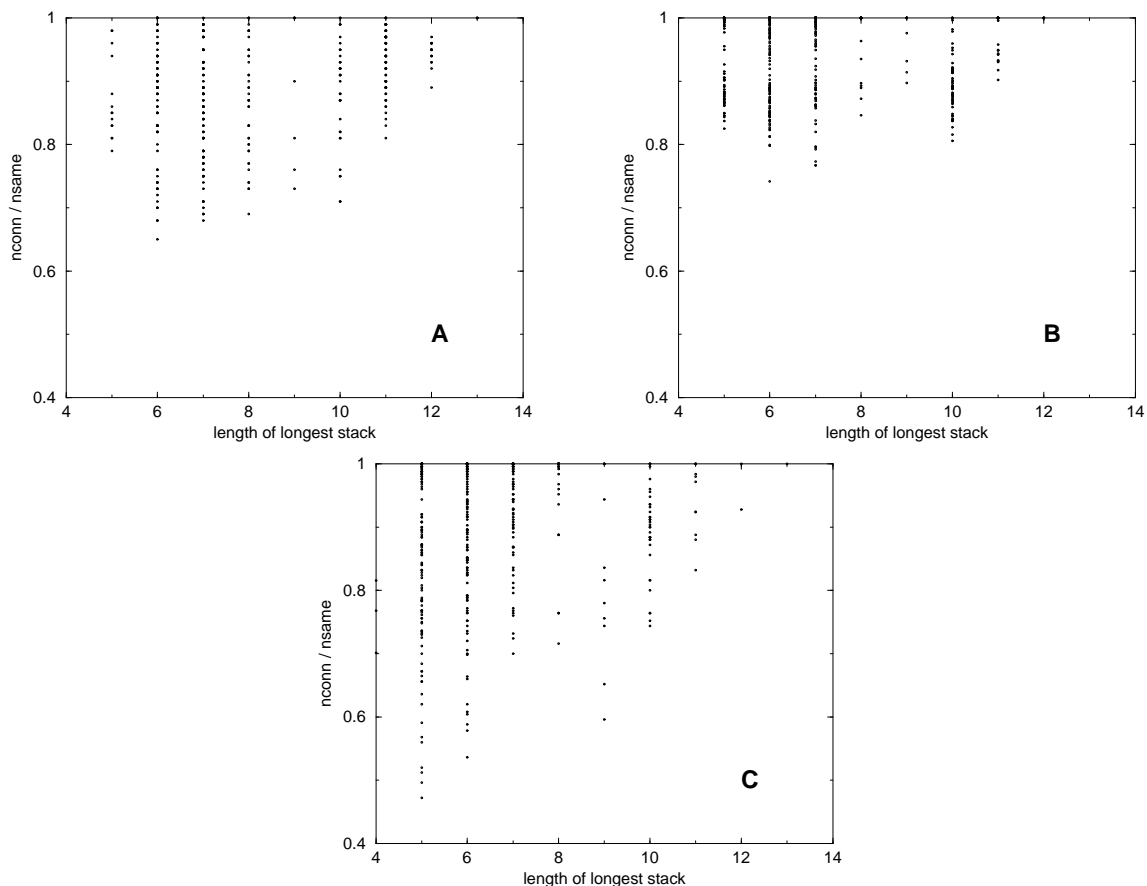


Figure 35: The networks of structures with long stacks exhibit high fractions of $nconn/nsame$ in all three test sets. For structures other than the most stable ones, the degree of local connectivity (and thus the expected performance of the connectivity test) is however not easy to predict.

4.3.5. Relation Between the Degree of Local Connectivity and Other Variables

The underlying secondary structure of a neutral network itself gives a hint to the reliability of the algorithm: the longest uninterrupted stacks are associated with high values of $nconn/nsame$ (see figure 35). Accordingly, these networks exhibit the highest absolute values of minimum free energies and the best *well-definedness* [50]. Prediction by these measures is however not better than by the much simpler to determine stack length (data not shown).

4.4. Outlook: Analyzing Large Sequence Spaces Over the Natural Alphabet

4.4.1. Performance Analysis

As first step towards assessing the feasibility of analyzing large sequence spaces by means of the connectivity test, we have studied the relation between the number of instructions spent in the test and the chain length, using the spaces Q_{AUGC}^{20} , Q_{AUGC}^{30} , Q_{AUGC}^{40} , Q_{AUGC}^{50} , Q_{AUGC}^{60} , and Q_{AUGC}^{70} . The piece of code analyzed corresponds to the pseudocode in table 5 between the labels 1: and 2: . It tests a single pair of points in sequence space (neighbours of the to be connected pair of sequences) for the existence of a path. (Due to a bug in the program, sometimes more than a single pair was tested. Although the scaled measure instructions/call is not affected by this, the chance of hitting a “difficult” pair is enhanced if there are more pairs.) The number of calls and the number of instructions spent in the function (from here on called “the test”) were recorded by the UNIX utility programs `pixie` and `prof`.

The basic experiment consisted of selecting a random network from the space, determining two pairs of folding sequences at each of the distances 3, 4, and 5, and running the test on each of the 6 pairs.

Figure 36 presents the results. According to our above argument, it should scale linearly in time with the chain length. It turns out that the average behaviour of the test on the different spaces is quite similar (at all chain lengths the peak of the distribution is in the bin of [1500 ... 2000] instructions per call), but the number of outliers in the righthand tail is increasing with chain length. These are the cases in which backtracking has to be done and the chain length dependent cardinality of the search space comes into play. The means of the distributions indeed increase very roughly linearly. In an actual application however the occasional outliers pose a serious problem.

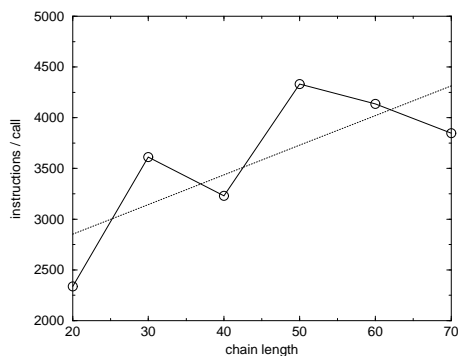
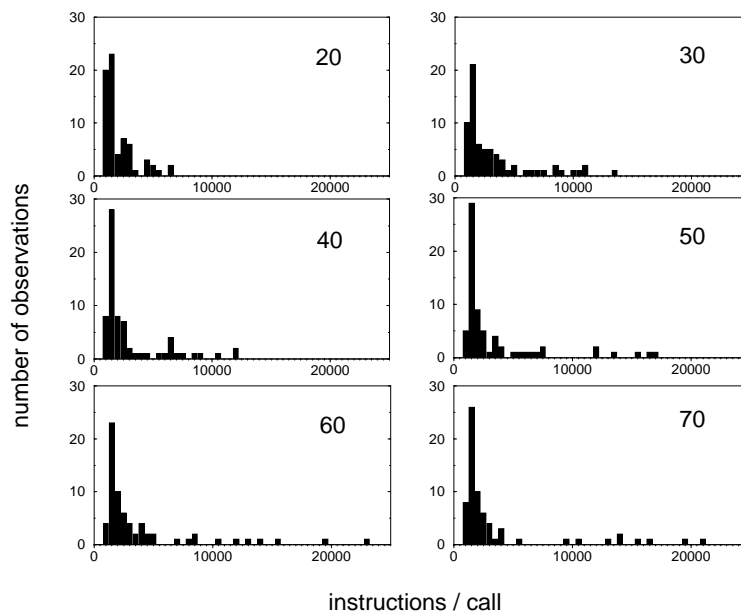


Figure 36: Performance of the connectivity test (code in table 5 between the labels 1 and 2) on Q_{AUGC}^{20} , Q_{AUGC}^{30} , Q_{AUGC}^{40} , Q_{AUGC}^{50} , Q_{AUGC}^{60} , and Q_{AUGC}^{70} . At the length of 50, the distribution contains a single additional count at 47033 instructions/call, which is not displayed.

4.4.2. Preliminary Data On Connectivity

In order to get a more realistic impression on the degree of connectivity in the six spaces, we ran the full algorithm given in table 5 on 125 sequence pairs from each of them. The pairs were required to have a (Hamming) distance in the range of 3 to 7 (due to the big cardinality of the

shape space in large sequence spaces over the natural alphabet it takes too long to find a fixed number of pairs on a given neutral network at a given exact distance).

The following values for $nconn/N$ were observed:

sequence length	d=3(#)	d=4(#)	d=5(#)	d=6(#)	d=7(#)	seconds
20	1.00(11)	1.00(19)	0.96(28)	1.00(39)	1.00(28)	308
30	1.00(12)	1.00(8)	1.00(27)	1.00(35)	1.00(43)	516
40	1.00(11)	1.00(15)	1.00(30)	1.00(36)	1.00(33)	5259
50	1.00(11)	1.00(20)	1.00(34)	1.00(33)	1.00(27)	2377
60, run 1	1.00(10)	1.00(18)	1.00(20)	1.00(39)	1.00(38)	4961
60, run 2	1.00(10)	1.00(12)	1.00(26)	1.00(31)	1.00(46)	5714
60, run 3	?	?	?	?	?	> 163262
70	1.00(17)	1.00(28)	1.00(31)	1.00(25)	1.00(24)	1616

The data refer to the networks of the following structures:

sequence length	structure
20(((.....))..
30	...((((((.....)))).....
40	...((((((.....)))).....(((.....))..
50	...((((((.....)))).....(((.....)).....
60((((((((.....))))))..(((((((.....((.....)).))))))..
70	...((((((.....)))).....(((.....)).....(((.....)).....

Again, the execution time depends much more on the individual network than on the sequence length. In all networks on which there are data all 125 pairs could be connected, with a single exception in Q_{AUGC}^{20} . There are however reasons to suspect that the network which was probed in Q_{AUGC}^{60} is not fully (locally) connected. The third run was aborted without output after 2 days, when only a total of 77 pairs had been tested. Most of this time had been spent with only three different pairs. These pairs might be truly disconnected. From such cases it is evident that performance would greatly benefit if instead of testing all neighbours of a poorly connected pair, one would already decide after a fixed number of attempts that the pair is not connected. It is however unclear how many true connections would be missed this way.

5 Neutrality With Respect To A Substructure

5.1. Introduction

The typical outcome of artificial selection of RNA molecules for a binding function is a family of sequences which share some common structural feature (e.g. a hairpin of a certain length), but which neither have completely identical structures nor usually the same sequence length. They are members of the substructure neutral network of the feature in question, as discussed in chapter 2. One can speculate that sequences at early stages of the origin of life had properties similar to those observed in today's selection experiments, which is one more reason to be interested in neutral sets of this sort.

In order to compare such a more loose concept of neutrality with the results by Reidys et al. [71] on neutral networks of completely defined secondary structures, one needs to consider two points. The first is the graph structure of the space of compatible sequences, which provides upper limits for the total size of a neutral network, the number of neutral neighbours per sequence, and the degree of connectivity. (Note that if the space of compatibles is not connected, no network embedded in it can ever be. That is the case with the compatibles of a fully defined structure viewed as a subspace of a simple Hamming space, as discussed in chapter 2.) The second point to be considered is the folding probability. While the number of compatibles grows very fast with the difference between total length L and substructure length l , the probability of a compatible sequence to actually adopt the substructure often decreases with total length (dependent on the extent to which the formation of the substructure is context sensitive). The evolutionary fate of sequences which are longer than the substructure will depend on which of the two aspects prevails.

5.2. Definitions and Abbreviations

In the sequel the following definitions and abbreviations are used:

l	The length of the substructure under consideration.
L	The total sequence length.
i	$L - l$
\mathcal{C}_i	The set of sequences at additional length i which are compatible with the substructure.
C_i	The number of sequences at additional length i which are compatible with the substructure.
frame	A subsequence of length l . The k th frame spans positions $k \dots k + l - 1$. There are $i + 1$ frames in a sequence. A compatible frame is a subsequence which is in \mathcal{C}_0 .
p	A word of length $i + 1$ over the alphabet $\{*, 1\}$. Every position corresponds to one frame. A 1 means that this frame is compatible with the substructure. No statement is made concerning frames which are marked by an $*$.
q	A word of length $i + 1$ over the alphabet $\{0, 1\}$. A 0 at position k means that the k th frame is <i>not</i> compatible with the substructure. A 1 means that it is.
$\mathcal{C}_{i,x}$	The set of sequences of additional length i which show the compatibility pattern x ($x \in \{p, q\}$). The $\mathcal{C}_{i,q}$ constitute a partition of \mathcal{C}_i , which is not true for the $\mathcal{C}_{i,p}$.
$C_{i,x}$	The cardinality of $\mathcal{C}_{i,x}$.
<i>nsymbols</i>	The number of bases in the alphabet.
<i>npairs</i>	The number of base pairs in the alphabet.
n_u	The number of unpaired positions in an RNA secondary structure or in a substructure thereof.
n_p	The number of base pairs in a structure or substructure.

5.3. The Compatible Sequences of a Substructure

5.3.1. Estimating the Number of Compatibles

A completely defined RNA secondary structure can be interpreted as a substructure at additional length zero, hence $C_0 = npairs^{n_p} \times nsymbols^{n_u}$. It would be nice if the C_i , $i > 0$ could be estimated from easy to access quantities like n_u, n_p . In the following we give an estimation formula which works well if sequences which are compatible in more than one frame are very rare (that is, the $C_{i,p}$ constitute essentially a partition of C_i). It breaks down if this is not the case.

Every sequence in C_0 is mapped to at least $nsymbols$ sequences in C_1 , by appending one of the possible symbols in turn. The result is $C_{1, "1*"} = C_{1, "10"} \cup C_{1, "11"}$. The only constellations which cannot be reached this way are those in $C_{1, "01"}$. These sequences, which are compatible in the last frame only, can only be generated by prepending a symbol to a sequence of C_0 . All remaining outcomes of the prepend operation have already been generated by means of appending.

Thus C_i is related to C_{i-1} by

$$C_i = nsymbols \times C_{i-1} + C_{i, "00...01"} \quad (1)$$

$C_{i, "00...01"}$ depends on the substructure under consideration. If l is large and the substructure is complex, then the probability of more than one frame being compatible is small and $C_{i, "00...01"}$ can be approximated by $C_{i, "*****1"}$, the number of sequences which are compatible *at least* in the last frame. This is equal to $npairs^{n_p} \times nsymbols^{n_u+L-l} = C_0 \times nsymbols^l$. Thus

$$C_i \approx nsymbols \times C_{i-1} + (C_0 \times nsymbols^l) \quad (2)$$

which nonrecursively simply reads as

$$C_i \approx (i + 1) \times C_0 \times nsymbols^l \quad (2')$$

5.3.2. The Exact Number of Compatibles

With short, simple substructures (e.g. short hairpins), sequences with more than one compatible frame cannot be neglected.

Ideally, one would like to know the $C_{i,q}$ for all q at the additional length i in question. Then not only C_i would follow as $\sum_q C_{i,q}$, but one could also derive more detailed information like for example the proportion of all compatible sequences with exactly n compatible frames.

In the following we give a method for computing the $C_{i,q}$, *which however applies only to binary alphabets*.

5.3.2.1. The Dependency Graph of a Set of Frames

Previous work on the relation between the compatible sequences of two completely defined secondary structures [100] has emphasised the importance of the multiple dependencies which are introduced into a sequence if it is to satisfy the pairing patterns of more than one secondary structure. With a single pairing pattern to be fulfilled, the maximal covarying sets of positions are of size two, namely, the base pairs. Two patterns are already enough to build up large chains of dependencies, in the following way: imagine two positions p_1, p_2 which form a base pair in the first structure. Depending on the structures involved, p_2 now may form a base pair with a *different* position p_3 in the second structure. This position in turn may be paired with yet another position p_4 in structure number one, and so on. Jacqueline Weber in [100] describes these chains of dependent positions by means of group theory. A single completely defined secondary structure is uniquely mapped onto a permutation, in which two positions are equivalent if they form a base pair. The orbits of the product of the permutations of two structures are exactly the chains of dependent positions which arise in sequences which have to fulfill the pairing patterns of both structures. From the *orbit decomposition* it is possible to compute the number of sequences which fulfill both patterns. These sequences have been called the *intersection* or the *overlap* of the two structures [100].

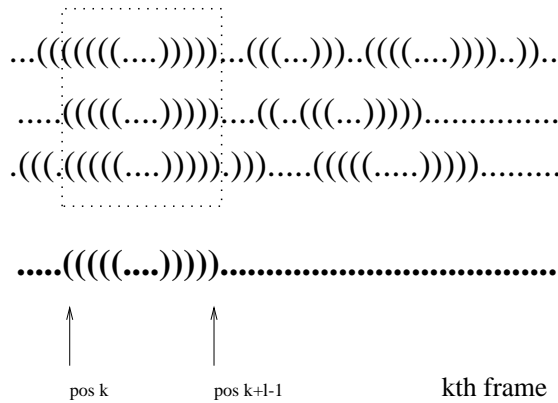


Figure 37: The pseudostructure which is associated with the k th compatible frame at a given total sequence length. If a sequence is compatible with the pseudostructure (according to the usual definition of compatibility for full length structures) then it is also compatible with the substructure in the k th frame. At positions outside the frame, its actual full length secondary structure may exhibit an arbitrary pairing pattern.

Substructure neutral networks can be viewed as a special case and an extension of the concept of the intersection of the compatibles of two completely defined structures. A sequence is compatible with the substructure in the k th frame if it is compatible with that full length structure which adopts the substructure beginning at position k , and is unpaired otherwise (cf. Fig. 37). Because the unpaired state outside the substructure is merely an ambiguous symbol which replaces the irrelevant true state, one could call such a catch-all structure a “pseudostructure”. *Sequences which are simultaneously compatible with two frames are members of the intersection of the pseudostructures of the respective frames.*

An important result concerning intersections, which was first stated by [70], is the following

Theorem 1. (Intersection Theorem) *Let s and s' be two arbitrary secondary structures. Then*

$$\mathcal{C}[s] \cap \mathcal{C}[s'] \neq \emptyset$$

where $\mathcal{C}[s]$ is the set of compatibles of structure s .

The group theoretical approach of [100] cannot be extended to the case of more than two structures. As a first step towards an alternative solution, we observe that the basic relation which is introduced

by any compatibility constraint is always pairwise, namely, “position i has to be able to form a pair with position j ”. A binary relation on a set of objects is equivalent to an undirected graph, with the objects as nodes and an edge joining two nodes if the relation is true.

Definition 12. (Dependency Graph). *Let $\sigma = \{s_1, s_2 \dots\}$ denote a set of RNA secondary structures of length L , with an arbitrary number of elements. By the dependency graph \mathcal{D}_σ of σ , we mean the following undirected simple graph on L nodes. The nodes are labelled $1, 2, \dots, L$ and correspond to the positions in the structures. There is an edge joining two positions i and j if these positions are paired in at least one of the structures in σ . In the special case in which the elements of σ are pseudostructures of a substructure at a given additional length i , the dependency graph is also denoted \mathcal{D}_q . Here, q is a word of length $i + 1$ over $\{0, 1\}$. The k th pseudostructure belongs to σ if the k th position in q is occupied by a “1”.*

From a simple property of \mathcal{D}_σ one can now in a first step deduce whether the simultaneous intersection of the structures in σ is empty or not.

Theorem 2. *Let \mathcal{D}_σ be the dependency graph of the set σ of RNA secondary structures. Let further $\mathcal{C}[s_i]$ denote the set of compatible sequences of structure $s_i \in \sigma$. If \mathcal{D}_σ is bipartite (i.e. its nodes can be partitioned into two sets such that no edge joins nodes which are in the same set)² then*

$$\bigcap_{i=1}^{|\sigma|} \mathcal{C}[s_i] \neq \emptyset$$

Proof. This is just a reformulation of the compatibility constraint on sequences over a binary alphabet. Let the alphabet be $\{A, B\}$ and the valid base pairs $\{A - B, B - A\}$. For a given sequence over this alphabet, denote by α the set of positions which are occupied by “A”, and by β the set of the remaining positions (occupied by “B”). Now every edge in \mathcal{D}_σ describes a base pair

²In practice a graph is tested for this property by arbitrarily assigning one node to one of the sets, then breadth first traversing the graph starting from this node. At every node encountered, all neighbours must be either unassigned or assigned to the complementary set of the node. As soon as a node is met for which this condition does not hold, the graph is not bipartite.

which has to be satisfied in a sequence which is compatible with all structures in σ . Thus such a sequence exists if there is a partition of the nodes of \mathcal{D}_σ into two sets α and β such that every edge joins an “A position” to a “B position”. ■

From \mathcal{D}_σ we can not only infer the (non)existence of simultaneously compatible sequences, but also their exact number.

Corollary 1. *Let \mathcal{D}_σ be a bipartite dependency graph with n_c components. Then*

$$\bigcap_{i=1}^{|\sigma|} \mathcal{C}[s_i] = 2^{n_c}.$$

For in any component of a bipartite graph, the only way of changing assignments of the nodes to the sets α and β is to flip the assignments of all nodes at once. Thus there are exactly two ways to allocate the symbols “A” and “B” to the positions which correspond to the nodes of a component. The allocations of different component are independent of each other. Therefore the total number of compatible sequences is equal to the cardinality of the cartesian product of the per-component allocations, 2^{n_c} .

In the context of the intersection theorem it is usually stated that it is unknown whether there exists a number J for which it holds that for all $j > J$ the intersection of the sets of compatibles $\bigcap_{i=1}^j \mathcal{C}[s_i] = \emptyset$ [54]. Without being able to make a rigorous statement concerning this open question, we can however observe the following: Imagine a substructure which consists of a single stack with a loop of even length. The corresponding pseudostructures of length L have a nonempty simultaneous intersection, because the alternating sequence is compatible with each of them. That is true for every $L = l \dots \infty$. Therefore $L - l + 1$, the number of pseudostructures at total length L , is a lower bound for J . Because this lower bound is not constant for all lengths L , J might not be constant, too.

It is not clear whether and how the dependency graph approach can be extended to non-binary alphabets and/or the case in which there is more than one pairing partner for a symbol. There is

certainly a connection to *coloring problems* in graph theory, which deal with the number of ways in which the nodes of a graph can be colored so that no edge joins nodes of the same color. The famous four-color theorem says that there is such a coloring with four colors for every planar graph [3, 4]. With base pairing however, the constraint is not just “different colors”, but “complementary bases”, which is most probably not always satisfiable. In addition, a dependency graph will usually not be planar.

5.3.2.2. The Exact Number of Sequences Which Are Compatible In At Least One Frame

In terms of the notation introduced at the beginning of this chapter, $|\cap_{j=1}^{|\sigma|} \mathcal{C}[s_j]| = C_{i,p}$, where σ is the set of those pseudostructures s_j which correspond to the frames which are marked by “1” in the word p over the alphabet $\{*, 1\}$. Thus by the method described above we are able to compute the $C_{i,p}$. In this subsection we show how these can be used in order to find the $C_{i,q}$, which are the quantities we are really interested in.

Let us focus on a particular $C_{i,q}$, for example $C_{i,{}''01100''}$. One can arrive at an inclusion exclusion type expression for $C_{i,{}''01100''}$ in the $C_{i,p}$ from two different starting points.

The first method is recursive and starts with observing that $C_{i,{}''01100''} = C_{i,{}''0110*''} - C_{i,{}''01101''}$: a “0” in the last coordinate is that which remains if patterns with a “1” in that coordinate are subtracted from those patterns with either a “0” or “1”. Now the index patterns on the right hand side of that equation contain one less “0”. Applying the same rule (split on the rightmost “0” into “*” and “1”) recursively to the terms on the right hand side will eventually lead to an expression in the $C_{i,p}$:

$$\begin{aligned} C_{i,{}''01100''} &= \\ (C_{i,{}''0110*''} - C_{i,{}''01101''}) &= \\ ((C_{i,{}''011**''} - C_{i,{}''0111*''}) - (C_{i,{}''011*1''} - C_{i,{}''01111''})) &= \end{aligned}$$

$$\begin{aligned}
 &(((C_{i, ''*11**''} - C_{i, ''111**''}) - (C_{i, ''*111*''} - C_{i, ''1111*''})) - \\
 &\quad ((C_{i, ''*11*1''} - C_{i, ''111*1''}) - (C_{i, ''*1111''} - C_{i, ''11111''}))) = \\
 &C_{i, ''*11**''} - C_{i, ''111**''} - C_{i, ''*111*''} - C_{i, ''*11*1''} + C_{i, ''1111*''} + C_{i, ''111*1''} + C_{i, ''*1111''} - C_{i, ''11111''}.
 \end{aligned}$$

Simple as this is as an algorithm, it is difficult to state in a closed form. The rearranged last line suggests however an alternative route to the same result. We see that the summands can be grouped according to the number of “1”s. All summands in such a group have the same sign and the signs alternate between groups. This is strongly reminiscent of the inclusion exclusion principle, by which the cardinality of the union of sets with nonempty intersection is computed [73]. And indeed the problem can be stated that way. In order to derive $C_{i, ''01100''}$ from $C_{i, ''*11**''}$, those sequences have to be excluded which are compatible in one or more of the starred frames. Assume there are n_0 zeros in the q for which we want to derive $C_{i, q}$. Denote by p_q that pattern in which all zeros are replaced by *s, and by $p_{q_1}(1) \dots p_{q_1}(n_0)$ copies of p_q in which the first, \dots , n_0 th * is flipped to a “1”. Then

$$C_{i, q} = C_{i, p_q} - |\cup_{j=1}^{n_0} C_{i, p_{q_1}(j)}|.$$

(Remember \mathcal{C} means the set of sequences, and C its cardinality.) The cardinality of the union on the right hand side can be described by the inclusion exclusion principle, by which

$$\begin{aligned}
 |\cup_{1 \leq j \leq n} A_j| &= \sum_{1 \leq j_1 \leq n} |A_{j_1}| - \sum_{1 \leq j_1 < j_2 \leq n} |A_{j_1} \cap A_{j_2}| + \sum_{1 \leq j_1 < j_2 < j_3 \leq n} |A_{j_1} \cap A_{j_2} \cap A_{j_3}| - \\
 &\quad \dots (-1)^{n+1} |\cap_{j=1}^n A_j|
 \end{aligned}$$

for n sets A_1, \dots, A_n . We observe that in the index patterns of the $C_{i, p_{q_1}(j_1)} \cap C_{i, p_{q_1}(j_2)}$ ($1 \leq j_1 < j_2 \leq n_0$) two of the *s in p_q are flipped to “1”, namely, the j_1 st and the j_2 nd. In general, in the intersection of k such sets, k of the *s are flipped in the index patterns. If we denote by $p_{q_k}(l)$ ($1 \leq k \leq n_0, 1 \leq l \leq \binom{n_0}{k}$) those copies of p_q in which k *s are flipped, the inclusion exclusion principle reads

$$|\cup_{j=1}^{n_0} C_{i, p_{q_1}(j)}| = \sum_{1 \leq k \leq n_0} (-1)^{k+1} \sum_{1 \leq l \leq \binom{n_0}{k}} C_{i, p_{q_k}(l)} = C_{i, p_q} - C_{i, q}.$$

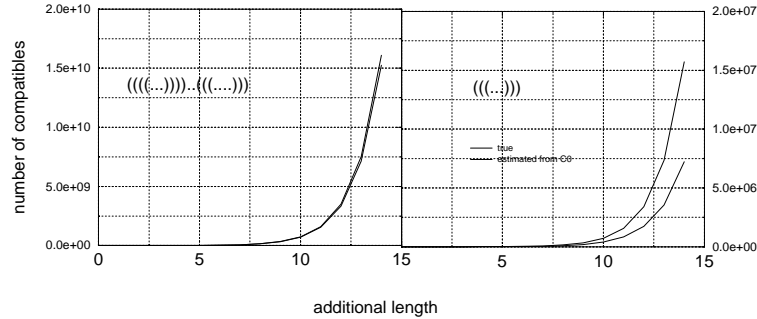


Figure 38: The number of binary sequences which are compatible with a substructure in at least one frame, estimated as $C_i \approx (i + 1) \times C_0 \times n_{symbols}^i$ (thin line), and its exact value (fat line). In the case of the long, complex substructure on the left side the estimation is very close to the true numbers. The reason for this is the fact that sequences which are compatible with more than one frame are very rare. If this condition is however not met, as in the case of the short hairpin on the right side, the estimation formula becomes increasingly more inaccurate for longer total sequence lengths.

Thus the $C_{i,q}$ can be computed from the know quantities $C_{i,p}$.

From the $C_{i,q}$, C_i follows as $\sum_q C_{i,q}$.

Fig. 38 compares the approximated cardinalities of the compatibles with the exact numbers, for the cases of a simple hairpin and a complex substructure. For the hairpin, the match is quite bad, as expected in a case in which it is not very unlikely to have more than one compatible frame in the same sequence. Thus it is obviously worthwhile the effort to compute the exact numbers.

5.3.2.3. The Density of Compatibles

The elements of \mathcal{C}_i are a subset of the $2^{(n_u + 2n_p + i)}$ binary sequences of length L . The ratio of C_i to the size of the sequence space is a measure of the availability of the substructure to selection. Of particular interest is its scaling with the additional length i , because an increased availability

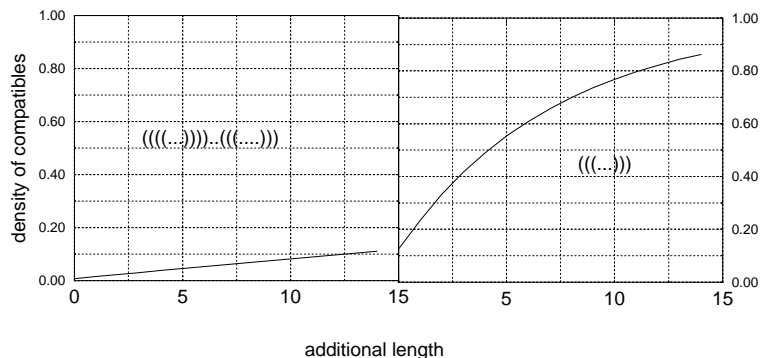


Figure 39: The density of compatibles $\sum_q C_{i,q}/2^{(n_u+2n_p+i)}$ as a function of the additional length i . In the case of the complex substructure, the rise is essentially linear up to $i = 14$, which is in agreement with the fact that the estimation formula works well for that substructure up to that length. In the case of the hairpin, the linear rise is clearly damped. Nevertheless an additional length of 14 is enough to take the density close to 1.0.

is one possible benefit of increasing the total sequence length.

From the estimation formula $C_i \approx (i + 1) \times C_0 \times nsymbols^i$ it would follow for a binary alphabet ($nsymbols = 2$):

$$d_i = \frac{(i + 1)2^{n_u} 2^{n_p} 2^i}{2^{(n_u+2n_p+i)}} = \frac{i + 1}{2^{n_p}}$$

This is a simple linear function of i which exceeds 1 for $i \geq 2^{n_p}$. Of course a density cannot be greater than 1, which means that the formula breaks down at latest at that value of i . In reality, the linear rise is counteracted by the increasing prevalence of sequences which are compatible with more than one frame. (A sequence which is compatible with k frames adds but a single count to C_i , instead of k counts which would be contributed by k independent sequences which are compatible with a single frame each.)

Fig. 39 compares the complex substructure and the hairpin of Fig. 38 with respect to their densities

of compatibles. Compatible sequences of the complex substructure are very rare at $i = 0$ (about 8 in 1000 random sequences), and their density rises very slowly but linearly to about 0.1 at $i = 14$. The linear rise suggests that we are in the regime in which the estimation formula still works, which is confirmed by Fig. 38. In contrast, the density of compatibles of the hairpin is equal to 0.125 already at $i = 0$, which exceeds the density of the complex substructure at $i = 14$. Its rise is damped from the beginning, which means that sequences in the intersections of multiple frames play a role already at small additional lengths. Nevertheless the hairpin reaches a density of $d_i = 0.86$ at $i = 14$. It is readily available “if needed” by selection. We will see in section 5.4.3 how this affects the population structure of a population of sequences which are kept on the substructure neutral network by selection for the hairpin.

5.3.2.4. The Probability Of Being Compatible In k Frames

From the $C_{i,q}$'s one can also compute the probabilities $p_1, p_2, \dots, p_{L-l+1}$ of a sequence to be compatible with exactly $1, 2, \dots, L - l + 1$ frames:

$$p_k = \frac{\sum_{q,q \text{ contains } k \text{ ''1''s}} C_{i,q}}{C_i}$$

These numbers are important if one were to describe a substructure neutral network by a random process on the set of compatibles (which we are not going to do). Assume there is a constant folding probability λ , with which a compatible frame actually adopts the substructure. Then the total folding probability of a sequence with k compatible frames is equal to the probability of at least one frame actually folding: $\lambda_k := \sum_{j=1}^k \binom{k}{j} \lambda^j (1 - \lambda)^{(k-j)}$. A random process probably would have to consist of two steps: the first “throw of the die” would decide about the actual folding probability of the sequence, which would be equal to λ_k with probability p_k . In the second step it would be decided whether the sequence belongs to the network, based on its folding probability.

Fig. 40 compares the p_k of a long, complex substructure and a short hairpin. In the former, $p_1 \approx 1.0$ even at an additional length i of 14. The network of this substructure could be modelled

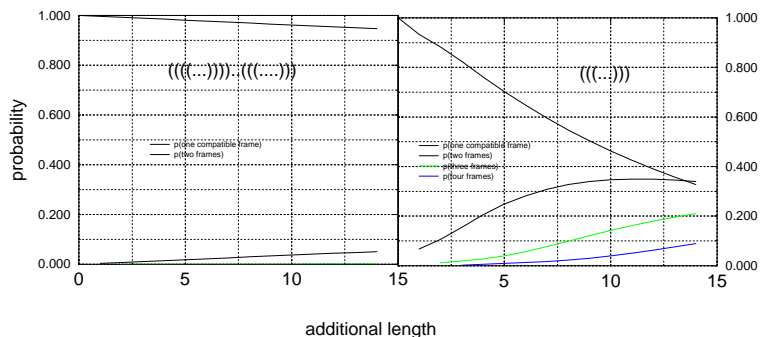


Figure 40: The probability to observe 1, 2, 3, or 4 compatible frames in a compatible sequence at different total lengths. For the complex substructure on the left side, the probability of one frame is overwhelmingly dominating. The only other probability which is not essentially zero is that of two frames, but this, too, is small compared to $p(1)$. With the simple hairpin on the right side the situation is very different. At $L - l = 14$, $p(2)$ already exceeds $p(1)$.

with a fixed λ . In the hairpin in contrast, p_2 is quickly rising with i , and has become the maximum of the p_k distribution at $i = 14$.

5.3.3. On Possible Graph Structures on the Set of Compatibles

5.3.3.1. Point Mutation Only Or Pair Neighbours ?

The compatible sequences for a given total length L are a subset of the $n\text{symbols}^L$ sequences which make up the Hamming space over that length and alphabet. Using point mutation as the only

mutational operator, the graph of compatibles would become a subgraph of the Hamming graph. If in addition insertion and deletion of single symbols are considered, the compatible sequences of different total lengths L are combined into a subgraph of the infinite Levenshtein graph defined in chapter 2. This is a very parsimonious way to describe the relation between the sequences, and it is close to biological reality. The possible objections against it are of technical nature: modelling the relation between the average number of neutral neighbours of a sequence and the likely connectedness and density of the whole network by random graph theory requires the graph of compatibles to be connected and regular. The second property surely does not hold for the above defined subgraph of the Levenshtein graph (let us call it graph A). It depends on the actual substructure whether or not the first one is true.

An alternative approach starts out from the set of pseudostructures which is derived from a substructure at some fixed total length L . There is a distinct pattern of base pairing for each pseudostructure, and thus it makes sense to speak of point neighbours and pair neighbours of a compatible sequence, quite like in the case of completely defined structures. The compatible sequences of a given pseudostructure can accordingly be organized into a generalized Hamming graph of the form $Q_{A,B}^{n_u} \times Q_{AB,BA}^{n_p}$ (assuming we are dealing with a binary alphabet). If a sequence is compatible in more than a single frame, it belongs to the intersection of the respective pseudostructures (cf. above). In accordance with Jacqueline Weber [100], one can define the set of neighbours of such a sequence to be the union of its neighbours in each of the generalized Hamming spaces. Then, a given position has as many different pair neighbours as there are compatible frames of the current sequence in which it is paired. In addition it has a single point neighbour if it is unpaired in at least one of the compatible frames. Let us call the graph which is induced on \mathcal{C}_i by this relation graph B.

In the following we will discuss properties of both graphs.

5.3.3.2. Connectivity of the Space of Compatibles

\mathcal{C}_i is naturally partitioned into the sets $\mathcal{C}_{i,q}$. Each such set is characterized by a unique set of constraints on mutations, given by the base pairs of its active pseudostructures and described by the dependency graph \mathcal{D}_q . The sequences of the set constitute the (multiple) intersection of the pseudostructures (cf. above).

Based on this partition, it is possible to state the problem of the connectivity of \mathcal{C}_i under a given set of move operators on a more abstract level than on the level of the sequences themselves.

Within a set $\mathcal{C}_{i,q}$, each component of \mathcal{D}_q constitutes one covarying group of positions. It depends on the move operators whether or not sequences within the same set are connected by a path which runs entirely within the set: A set $\mathcal{C}_{i,q}$ decomposes into 2^m components, where m is the number of components of \mathcal{D}_q with more than 1 element (in the case of point mutation only) resp. more than 2 elements (if pair exchanges are possible).

Sequences within the constituent components of the $\mathcal{C}_{i,q}$ are by definition connected. Thus the basic objects for which connectivity has to be proven are these components themselves.

For the case of point and pair mutation, a similar scenario has been described by [100]. She has shown that the sequences in the intersection of two completely defined secondary structures decompose into islands which are connected by paths in the generalized hypercube of structure I or II. These “islands” are the components of the set of intersection sequences, and the connecting paths run via one or the other of the two sets which together with the intersection partition the entire space: namely, those sequences which are compatible only with structure I or structure II, resp.

Thus if the graph structure on \mathcal{C}_i is that of graph B above (point and pair mutation), then the components into which a single $\mathcal{C}_{i,q}$ decomposes are connected by paths in the generalized hypercubes of one or more of the active pseudostructures of the set. Therefore in this case it is not even necessary to distinguish between the components. The relevant graph Γ_1 is defined

on the nonempty sets $\mathcal{C}_{i,q}$ themselves, and two such sets are joined by an undirected edge if the intersection of its constituent pseudostructures is nonempty (for if this is the case, any sequence in one set is connected with any sequence in the other set by a path in the generalized hypercube of one of the shared pseudostructures). Sequences in sets which do not share pseudostructures are connected if there is a path in Γ_1 connecting the sets. This is however by the intersection theorem is always possible: Imagine two sets of frames q, q' which do not share an element. Then the singletons $q'' = \{x\}$ and $q''' = \{y\}$ exist for arbitrary $x \in q, y \in q'$. By the intersection theorem $q'''' = \{x, y\}$ is not empty. Therefore there is a path (q, q'', q''', q''', q) connecting the two sets of frames. *It follows that under point and pair mutation the graph of compatibles of a substructure is connected.*

The situation is more complicated if the move set consists of point mutation only. Now one has to deal with the explicit components of the sets $\mathcal{C}_{i,q}$, because it is not guaranteed that connecting detours exist. There is an edge between two components if there exists a sequence in one set which by a point mutation is converted into an element of the other set. Verifying the existence of an edge between specific components of different sets $\mathcal{C}_{i,q}$ requires detailed sequence information, which is why it may indeed be easier in this case to do a component decomposition on the sequences themselves.

5.3.3.3. On the Number of Compatible Neighbours

A mutation which “leads out of” $\mathcal{C}_{i,q}$ is not necessarily forbidden. This is only the case if the result is incompatible in *all* frames. Outcomes in which at least one compatible frame remains belong to some different $\mathcal{C}_{i,q'} \subset \mathcal{C}_i$.

A mutation which results in an incompatible sequence surely has to destroy all current compatible frames of a sequence. This can only happen if it hits a position which is paired in all frames.

The number of compatible point neighbours of a sequence in \mathcal{C}_i is at least equal to L minus the

number of positions which are simultaneously paired in all compatible frames of the sequence.

The actual number of point neighbours may be larger, because formerly incompatible frames (if they exist) may become compatible by the mutation, so that a compatible sequence may still result.

There is a threshold number $n_f^* = r + 1$ of compatible frames, so that for $n_f \geq n_f^*$ the number of simultaneously paired positions must zero, and the number of compatible point neighbours equals L . Here, r is the length of the longest run of paired positions in the substructure.

The number of compatible neighbours in graph B is at least equal to L minus half the number of simultaneously paired positions (and thus greater than the number of neighbours in graph A).

This is so because every position which is unpaired in one or more compatible frames has *at least* one neighbour (in graph A, it has exactly one). At a simultaneously paired position, as many pair exchanges are possible as there are frames, while such a position may not be mutated in graph A. Of the pair exchanges, at most half are redundant (if both partners of a pair are in the set of simultaneously paired positions, only one exchange counts). The actual number of compatible neighbours depends on the active pseudostructures and can be much larger than this lower limit.

5.4. Substructure Neutral Networks

5.4.1. The Folding Probability of a Single Compatible Frame

Fig. 41 illustrates the following setup: n_1 different compatible sequences of length l are chosen for the respective substructure (hairpin or complex). Because there are only 64 compatibles of the hairpin, $n_1 = 64$ in this case. For the complex substructure, $n_1 = 5000$. Next for each of the compatible sequences, n_2 random contexts of length $i = L - l$ are chosen. In order to compensate for the bad statistics with 64 compatible sequences only, $n_2 = 100$ for the hairpin (for values of i

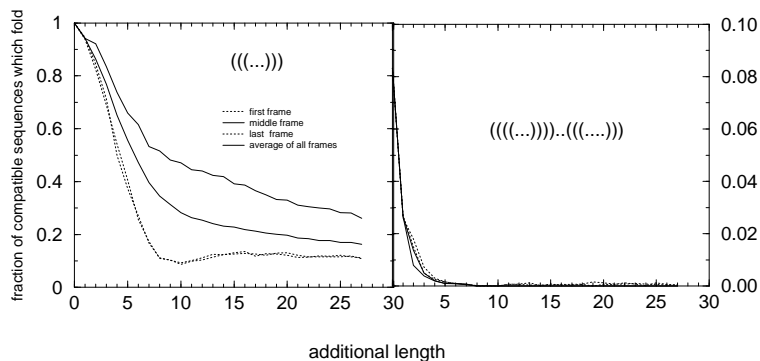


Figure 41: The probability of a compatible frame to actually adopt the substructure. Note the different scales of the y axes.

which are too small to permit 100 different realizations, n_2 is set to the maximal number of different contexts). In the complex substructure, $n_2 = 1$. Each compatible sequence is inserted into each of its random contexts in all possible frames, and the number of times the substructure is realized *in this frame* is recorded. (The total sequence may by chance contain one or more additional compatible frames, and the substructure may be adopted on one of those. Such an outcome counts as negative.) Finally, the fraction of successes, pooled over all compatible sequences and all random contexts, is reported for each frame. The average of this measure over all frames is an estimate of the per-frame folding probability $\lambda_{i,1}$. It is this parameter on which it hinges whether or not the

increased density of compatibles at $i > 0$ results in an increased density of neutral sequences or not.

The two substructures behave very differently in several respects. At $i = 0$, all of the compatible sequences of the hairpin actually fold, while only less than 1/10 of those of the complex substructure do. This is not due to a higher stability of the hairpin: with a $\langle \Delta G \rangle$ of -2.4 kcal/mol, it is actually less stable than the complex substructure ($\langle \Delta G \rangle = -7.5$ kcal/mol). The difference is in the extent to which the two different substructures exploit the available length with respect to stability: the hairpin is certainly the most stable possible structure of a sequence which is compatible with it. That is not necessarily true for the complex substructure. The size of its longest stack, 4, is small compared to a stack length of 10 which is possible at a sequence length of 23. Many compatible sequences of this substructure are capable of forming longer stacks: an example is the sequence “GGGGCCCCCGGGGGGCCCCC”, the minimum free energy structure of which is “((((((((((.....)))))))))”, with a ΔG of -20.53 kcal/mol. Note that the simple combinatorial argument “the longer the length of the substructure, the more alternative structures it has to compete with” does not hold. What counts is the average rank of the substructure in the list of suboptimal structures of the compatible sequences (which we are not able to compute here).

Proceeding to $i > 0$, we observe the following. First, the average folding probability of a compatible frame decreases with increasing i in both cases. The rate of this decrease is very different in the two substructures: at $i = 1$, the folding probability of the complex substructure is less than one third of its initial value, while it is only reduced by about 6% in the hairpin. There are several possible reasons for this. At $i = 1$, the hairpin is still the energetically “best possible use of a compatible sequence” [37]. At $i > 1$, longer stacks are possible. Yet, the hairpin is a substructure of all stacks with a 3-loop, so these longer stacks can still be substructure neutral. The role of this effect is evident from the fact that the location of the compatible frame matters a lot in the case of the hairpin: frames in the middle of the sequence have the highest folding

probability. It is these frames that may be extended to the longest stacks. (In addition, the contiguous stretches of random context are smallest if the frame is located in the middle of the sequence. Longer such contiguous stretches are sterically more free to interact with positions in the compatible frame, possibly resulting in an actual structure in which positions in the frame are paired with outside positions.) The stacks of the complex substructure, in contrast, cannot be bulge-free extended. With increasing i , the stability of the substructure will increasingly fall behind the increased stability of competing long stacks. (An example is the compatible sequence GGGGGGGGGGGGGCCCCCCCCCCCCCGGGC at $i = 7$. Its minimum free energy structure is (((((((((((((.....))))))))))....), with a ΔG of -27.50 kcal/mol.) There is no pronounced difference between terminal and middle frames in this case.

The folding probability of the complex substructure quickly approaches zero for all frames. Therefore in a scenario in which there is selection for the substructure but the total length is allowed to vary, one would expect that there is a big advantage in reducing the additional length to zero. We will discuss such a scenario in section 5.4.3. The situation is different for the hairpin. After having reached a minimum at $i = 9$, the probability of the terminal frames slightly rises again and then seems to rest at a plateau of about 0.1 (to be sure about the stability of the plateau, it would be necessary to run the experiment for $i > 27$). Probably this is the additional length at which the context sequence is long enough to form a folding unit of its own (note that the length of the hairpin is equal to 9). The average and middle frame probability are still decreasing at $i = 27$, although very much more slowly so than in the case of complex substructure. The former is at 0.16 at $i = 27$, the latter at 0.26 (compare that with the average probability of 0.000042 in the complex substructure at that additional length).

5.4.2. From $\lambda_{i,1}$ to the Folding Probability of a Random Sequence

In the last subsection we have described $\lambda_{i,1}$, the average single frame folding probability at

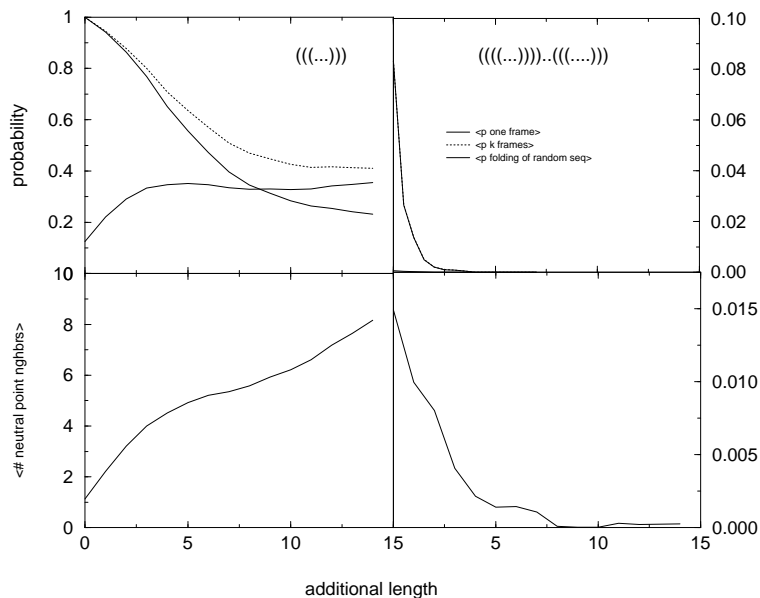


Figure 42: Upper panel: The average single frame folding probability $\lambda_{i,1}$ (fat line), and the derived measures $\lambda_{\langle i \rangle}$ (the probability of at least one compatible frame folding, dotted line), and $\lambda_{i,r}$ (folding probability of a random sequence, light line). In computing $\lambda_{\langle i \rangle}$, the probabilities of up to 6 frames were taken into account. **Lower panel:** the expected number of neutral point neighbours, estimated as $(i + l) \times \lambda_{i,r}$. Longer sequences are advantageous for the hairpin, while $i = 0$ is clearly the optimum for the complex substructure.

additional length i . Together with the probability of k frames being compatible, $p_{i,k}$ and the density of compatible sequences, d_i , it allows to compute the probability of folding of a random sequence of length $L = l + i$. Of course this probability could have been determined experimentally by folding a large number of random sequences, quite as $\lambda_{i,1}$ has been. However by expressing it in terms of $p_{i,k}$ and d_i , which need not be statistically determined but are exactly known, one gets a better understanding of what causes the values to be in the range they are. In addition, one could conceivably systematically study the influence of varying values of $\lambda_{i,1}$ (which we have not done).

If there are k compatible frames in a compatible sequence, the total folding probability is equal to $\lambda_{i,k} := \sum_{j=1}^k \binom{k}{j} \lambda_{i,1}^j (1 - \lambda_{i,1})^{(k-j)}$, the probability of at least one frame actually folding. The expected folding probability of a random compatible sequence, considering up to n frames, is $\lambda_{<i>} := \sum_{k=1}^n p_k \times \lambda_{i,k}$. The probability of a random sequence to adopt the substructure is equal to the product of the density of compatibles and the probability of folding: $\lambda_{i,r} := \lambda_{<i>} \times d_i$. It is this measure which determines whether or not it “pays” to elongate the sequences beyond $i = 0$. The neutral network of a substructure is likely to be first found (under selection for a task which can be carried out by the substructure) at this additional length i for which $\lambda_{i,r}$ is maximal. Once a population is on the network, the scaling of the number of neutral neighbours with i is expected to be an important determinant of the growing or shrinking of the total sequence length. As Martijn Huynen and Erik van Nimwegen [94] have shown, a population on a neutral network accumulates at those nodes which have the most neutral neighbours (the so called buffering). A simple approximation of the number of substructure neutral neighbours is given by $(l + i) \times \lambda_{i,r}$: this considers only point neighbours, and assumes that each such neighbour behaves like a random sequence.

Fig. 42 contrasts the different measures. As we have seen in the previous subsection, $\lambda_{i,1}$ of the hairpin decreases monotonically on $i \in [0, 14]$. $\lambda_{<i>}$ is clearly different from $\lambda_{i,1}$ already at $i = 1$. Between $i = 11$ and $i = 14$ it goes on a semi-plateau on which it shows only a minimal decrease from 0.415 to 0.411. The decay of $\lambda_{<i>}$ and the rise of d_i combine into a $\lambda_{i,r}$ which, after an initial rise from 0.125 to 0.33 (between $i = 0$ and $i = 3$) stays nearly constant. (There is a shallow minimum between $i = 8$ and $i = 10$. From then on $\lambda_{i,r}$ is very slowly but monotonically increasing again.)

The expected number of neutral point neighbours, estimated as $(l + i) \times \lambda_{i,r}$, is monotonically increasing for the hairpin.

We have seen above that compatible sequences of the complex substructure with more than

a single compatible frame are extremely rare. Consequently $\lambda_{<i>} \approx \lambda_{i,1}$ (the two curves are indistinguishable in Fig. 42). The very small total folding probability of a compatible sequence renders the folding probability of a random sequence even smaller: $\lambda_{i,r} \leq 0.000649$. The expected number of neutral point neighbours is much less than 1 even at $i = 0$, and it is further decreasing until $i = 14$. All these point into the same direction: this substructure is evolutionarily most stable when it is a completely defined structure with no additional sequence portions.

5.4.3. A Population on a Substructure Neutral Network

5.4.3.1. Background

We have seen above that the density of compatibles and the folding probability of a substructure vary (sometimes dramatically) with the additional sequence length i . Assume a population of sequences is evolving under point mutation, 1-insertion, and 1-deletion according to the protocol of Gillespie [36]. Sequences on the substructure neutral network have a constant fitness (irrespective of their length and the number frames which may simultaneously fold). The fitness of sequences off the network depends on the local string alignment score of the substructure to the complete structure of the sequence (thus there is not only “on” and “off” the network, but also “close to” and “far away”). If the population is seeded by M copies of a sequence with a single compatible frame, it depends on the density of compatibles at a *total* length of i whether or not it is likely that additional frames emerge in the random context. It is crucial for the genetic makeup of the population whether or not they do: the emergence and selection of a new frame is equivalent to the de novo creation of a sequence, independent of ancestors in the population. But new frames emerge in the context of old frames, with which they remain joined in a single replicating unit. They do not compete with frames in the same sequence, but jointly enhance the resistance of the sequence

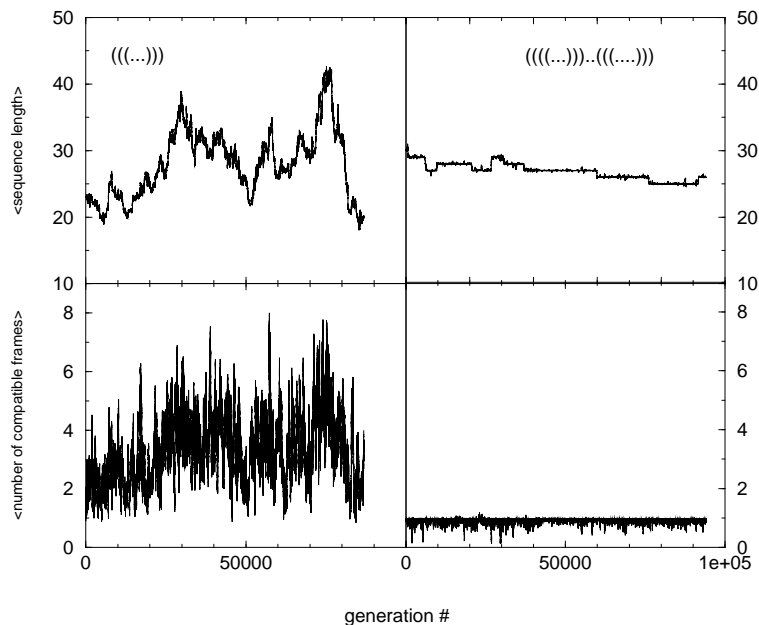


Figure 43: The average sequence length (upper panel) and the average number of compatible frames per sequence (lower panel) in a population of size 250 which is drifting on the substructure neutral network under point mutation, 1-insertion, and 1-deletion. The one-digit mutation rates for the three operators are 0.001 (point mutation), 0.005 (1-insertion), and 0.005 (1-deletion).

to mutation. For if there are multiple compatible frames, and a mutation destroys one of them, the substructure may still form on a different one, and the fitness of the sequence stay unchanged. Like a pseudogene in a genome, a frame which has been rendered incompatible by a mutation is further carried along, and has the chance to revert to compatibility by a compensatory mutation. If however the de novo formation of a frame is very unlikely, then the initial compatible frame will be the ancestor of a single true phylogeny. The role of the context reduces to its influence on the folding probability of the frame. If the influence is negative, as with the complex substructure described above, one expects that the additional length is reduced to zero.

5.4.3.2. Setup

The experiment was run for the hairpin and the complex substructure discussed above. The initial sequence length L_0 was equal to 23 for the hairpin and equal to 30 for the complex substructure. In both cases the population size was 250 and the sequence which seeded the population contained a single compatible frame. The sequence of the hairpin did *not* adopt the substructure on this frame, but the network was found already in the second generation. Because it turned out that the network of the complex substructure would not be found in reasonable time if the initial sequence did not fold (data not shown), in this case a folding sequence was deliberately selected.

The error rate of the mutational operators were set to 0.001 (point mutation), 0.005 (1-insertion), and 0.005 (1-deletion). The reason for the higher rates of insertion and deletion was the fact that we were interested in whether or not there were trends in the change of the total sequence length. Higher rates of length-altering operators reduce the waiting time until a selectively favored length mutant arises.

The fitness of a sequence was computed as $10^{(score/best\ score)}$. The scores in this formula refer to a local alignment[84] of the substructure to the complete secondary structure of the sequence, using the following symbol similarity table:

$$\begin{array}{cccc} & (&) & \cdot \\ (& 5. & -2. & 0. \\) & -2. & 5. & 0. \\ \cdot & 0. & 0. & 3. \end{array}$$

bestscore is the score of an exact match. Thus the score of a sequence on the network is 10.0, and sequences off the network have smaller, similarity dependent scores. The penalty for opening and elongating a gap both were equal to 1.0.

5.4.3.3. Results

In the course of 100000 generations, the average sequence length of the hairpin population shows big and often very rapid alterations (see Fig. 43). There seems to be no general trend towards longer sequences, although a mean value of nearly twice the start length temporarily occurred. The

average number of frames per sequence is high (about 3 averaged over the entire run, but at times equal to 8). It follows the mean sequence length: the density of compatible frames in the sequences is not (markedly) increasing. What is true is that the average number of frames rises from 1 to 2 during the first 100 generations (during which the average sequence length is unchanged), and that it never drops below 2 again except for very short time intervals.

In the complex substructure, there is a general trend towards a shorter length, as expected. In contrast to the hairpin, length alterations are not easily accepted: there are long *epochs* [95, 96, 93, 92] of constant length. During the whole run, the average number of compatible frames is less or equal to one. However in this case the density of frames is slightly increasing, due to the reduced number of sequences in the population with zero compatible frames.

5.4.3.4. Interpretation of Results

In the hairpin, recurrent mutations which create compatible frames surely play a role, and probably more so than the “ordinary” passage of genetic information from ancestor to offspring via replication. That would explain why there is no preferred sequence length: the folding probability of a random sequence is indeed length independent for the hairpin. The density of frames per sequence of ≥ 2 is in accordance with the fact that $p_2 > p_1$ from an additional length of 14 on ($L_0 = 23$ corresponds exactly to $i = 14$). The fast acquirement of a second frame during the first 100 generations thus may *not* be due to selection for the sheltering effect of multiple frames.

The epochs in the time evolution of the sequence length in the case of the complex substructure point to a strong context dependency of the folding of the substructure: in most contexts a length alteration is not permitted. For the same reason the length decreases only very slowly, sometimes even increasing again for a short time.

6 Discussion

That phenotypically neutral mutants of the genotype contribute to both robustness of the genotype-phenotype map against mutations and to an enhanced search potential for better adaptations is theoretically evident [48, 49, 78, 72, 96, 91]. It is also known that the secondary and tertiary structures of biological macromolecules are often unchanged by extensive alterations to the primary sequence [77, 39, 24]. What remains open is how the graph structure of the complete *neutral networks* of the natural sequence structure mapping look like. That they contain large connected components or that different components are at least close to each other (density) is a prerequisite if neutrality is to have a major impact on search capacity at mutation rates which are low enough so that the genotypes can form a single or a few stable quasispecies [80, 49, 78]. The existence of largely connected (and sometimes dense) networks has been shown analytically and/or by simulation for a variety of simplifications of the sequence structure mapping (random graph model [71], lattice proteins [59, 7], RNA secondary structures over binary alphabets [37, 38]). In this thesis we have further approached the natural situation from two directions: first, we have not only explicitly determined the neutral networks with respect to secondary structure in an RNA sequence space over the natural alphabet but also developed an algorithm which allows an estimation of whether or not a network is connected from a statistical sample of its sequences and thus can be applied to the networks in sequence spaces over the natural alphabet for sequence lengths which do not permit exhaustive folding. Second, we have done the first steps in the direction of introducing a new concept of structural neutrality, *neutrality with respect to a substructure*. This takes into account the fact that the strength of functional selection pressure often is not the same everywhere in the 2D or 3D structure of a biological macromolecule [90, 42]. Thus a mutation which alters the

structure in a less relevant part may still be neutral with respect to the function of the molecule. The most outstanding feature of the sequence structure map in the space Q_{AUGC}^{16} of all RNA sequences of length 16 over the alphabet $\{A, U, G, C\}$ is the huge size of the neutral sets. With $4^{16} = 4294967296$ sequences which are mapped to only 274 structures the average size of a neutral network is 15675063 sequences. The space is dominated by the set of the open structure, which with a count of 2709569048 comprises about 63% of the entire space and occurs in the 1-mutation boundaries of nearly all structures with high frequency. The sizes of the networks roughly follow a generalized Zipf's law if the open structure is excluded. In particular, the 70 highest ranking sets are comparable in size and much bigger than most of the remaining ones, so that one can distinguish between common and rare structures [81]. The networks of the common structures exhibit the properties which are most advantageous both in terms of robustness against mutations and of search capacity: they are large connected graphs and in their 1-mutation boundaries contain nearly all common structures other than the reference structure [31]. The common structures are also those which cover an appreciable amount of their compatible sequences (the compatible sequences of most other structures do not actually adopt the structure) and which have neutral sets which are approximately uniformly distributed over the entire sequence space. Yet also in these structures the fraction of neutral neighbours of a sequence on the network is considerably different from and much higher than the coverage of compatibles, indicating that there is a correlation of the propensity to be on the network between neighbouring sequences and thus a random graph model which independently chooses sequences cannot be fully appropriate. The random graph model according to Reidys *et al.* (1997) [71] predicts the networks to be dense and connected up about rank 125 and disconnected and not dense in their compatibles otherwise. In contrast to the prediction, nearly all networks in the space are connected (229 out of 273, not counting the open structure) but no network is dense in the sense that every compatible sequence which is not on the network has at least one neighbour which is. The fraction of non neutral compatible sequences

which have this property is correlated with the size of the network. The lack of full density even of the common structures can be explained with the overwhelming count of the open structure in this space, which displaces other networks. The larger than expected connectivity of the rare networks in contrast may not be an artifact of the small chain length but a genuine property of rare networks in general: rare structures in \mathcal{Q}_{AUGC}^{16} , and presumably in general, often show a dependency on a special sequence context of the formation of certain structural elements. This leads to long range correlations in sequence space of the folding probability for the structure, which in turn will cause a concentration of the neutral network of the structure in regions of high folding probability. If there is only a single such region, a fully connected, localized graph may result. In addition, the small distances between elements of \mathcal{Q}_{AUGC}^{16} may lead to a merging of components [82], which again would be a sequence length dependent effect.

When talking about a neutral network of an RNA secondary structure one usually thinks of a fitness landscape which takes a constant value on the sequences of the network and a much smaller, but likewise constant value on the remaining sequences of the space [72]. It turned out that the sequences of many networks of \mathcal{Q}_{AUGC}^{16} , differ greatly with respect to the free energy of folding into the reference structure. The free energy of folding may well contribute to fitness, for example if the sequence needs to unfold in order to be replicated [29] or else if an especially stable fold is required in an extreme environment. Neither the mean free energy (or its variance) nor the topology of the energy landscape over the sequences of a network (mostly showing an exponentially decreasing autocorrelation function, yet with very different slopes in different networks) is correlated with the size of the network. This underscores the fact that it is not the absolute energy gain on folding into a particular structure which determines the folding probability of a particular sequence. Rather, it is the differential gain compared to other structures with which the sequence is compatible, thus introducing a combinatorial element into the sequence structure mapping. The energy landscapes of very similar structures (such as those ranked 30 and 31, which are mirror images of each other)

can be extremely different. While the latter has a mean free energy which is about twice that of the former, its energy landscape contains more different values and has a much smaller autocorrelation. Thus the two structures provide to selection two versions of the same tool (assuming that they are basically equally well suited for any static intermolecular interaction) made of very different material, each of which may be superior in some situation.

Sequence spaces for lengths greater than 16 over the natural alphabet cannot be exhaustively folded in reasonable time. Longer sequences are however expected to better conform to the random graph model of Reidys *et al.* (1997) [71] (this is inherent in the model itself, which makes strict predictions only for infinite sequence lengths – but see Stephan Kopp on finite lengths [54]). One prediction of the model, namely, the existence of a path between any two connected sequences in a defined, small hypervolume in sequence space, is especially interesting in connection with the issue of analyzing the graph structures of large neutral networks. If it holds for a network then the task of testing whether a pair of sequences is connected is greatly simplified. Assuming that there is no extreme clustering of components of the network in sequence space, total connectivity can then be estimated from the number of connected pairs observed in a statistical sample of sequences. This can not only be done for the natural alphabet *in silico*, but in principle also for natural sequences and structures. (One is then not confronted with the infeasible task of determining the complete neutral network of a natural structure, but needs only find comparatively few sequences which adopt it. The idea of an *inverse SELEX* procedure, which would start from a single molecular species which already has a desired function and aims at increasing molecular diversity while not losing the function (thus exploring part of its neutral network) has been pursued by Dr. Michael Gebinoga (personal communication).)

Based on the above defined property, local connectivity, we were able to implement an algorithm which can classify common networks of the space Q_{GC}^{30} as connected or disconnected with a success rate of $\geq 89\%$. Preliminary data on larger spaces over the natural alphabet suggest a high degree

of connectivity. Because of its implications for the evolvability of RNA sequences this definitely deserves further investigation.

Neutral networks of RNA secondary structures have become a preferred object of study with respect to the questions of neutral evolution because (a) the combinatorics of the sequence structure map is comparatively simple and well understood [26] and (b) RNA sequences are probably close to the root of life, and may indeed once have functioned as “naked replicators” [35]. In this case properties of their neutral networks have of course been of utmost importance. The outcomes of artificial selection experiments [90] suggest that with respect to many tasks of molecular recognition the degree of neutrality in the sequence-task map may be even greater than that in the mapping of sequences into secondary structures: oftentimes the presence of some substructure somewhere in a sequence of variable length is enough to create a fully functional molecule [90, 76, 75, 74, 64, 42]. Such a loosening of the constraints on the structures, realistic as it may be, spoils some of the most convenient aspects of the relation between RNA sequences and their secondary structures. For example, a sequence is no longer either compatible with a structure or not: it can contain more than a single subsequence which is compatible with the substructure, and, if such subsequences do not overlap, it can even adopt the substructure more than a single time. Multiple compatible subsequences (“frames”) constitute both a constraint and a shelter: if compatibility with all of them is to be retained by a mutation, there might not be much freedom for change. We have given a formula for computing the number of sequences at a given total length which contain a particular constellation of compatible subsequences, using the newly introduced concept of a *dependency graph* (which is related to the *contact graph* discussed in [82]). If it is however only required that neutrality is retained (that is, a single compatible frame is kept), then the more mutations can be tolerated in a sequence the more frames it contains. Indeed sequences with a high number of frames temporarily occurred during the simulated evolution of a population of strings under point mutation, 1-insertion, and 1-deletion on the network of a simple substructure.









Whether or not sequences which are longer than the relevant structure “pay” in evolution depends on a) how well the substructure is able to autonomously fold in a large context of additional sequence and b) how probable the actual formation of additional compatible frames is. The latter question can be answered by the above formula, but on the former more extensive work needs to be done. Stable substructures which are simple enough to allow for de novo formation of compatible frames generated a broad spectrum of sequence lengths in a simulation. In nature this effect may contribute to the evolution of complex structures, by first generating additional sequence portions which then can be put to use by selection.

Another aspect which poses more difficulties in substructure neutral networks compared to full length structures is the graph structure to be imposed on the set of compatible sequences. Although we show that it is in principle possible for the set of compatible sequences of a substructure at a given length to be connected under point mutation (in contrast to the compatibles of a full length structure, which never are, except for the open structure), it seems to be preferable to take the correlations due to the base pairs into account. In accordance with Dr. Jacqueline Weber [100] we propose to mutate a pair of positions in a compatible sequence in a single step whenever they form a base pair in one of the compatible frames of the sequence. This may result in a position having more than a single pair exchange neighbour. The advantage of this construction is the fact that the resultant graph is connected. Correspondingly, it should be possible to describe a substructure neutral network as a random graph induced in the graph of compatibles. This will certainly be a most promising task for the future.



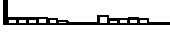

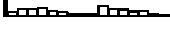
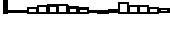













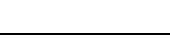
Appendix: SOCs of the Structures of AUGC₁₆







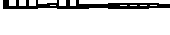













This appendix lists the structures which are realized by the sequences of Q_{AUGC}^{16} . Columns Rank, Size, and Structure contain the position in the size ordered rank list of all structures, the number of sequences which belong to the network, and the structure in bracket dot notation. Column SOC contains the Sequence Of Components: for every component of the network, there is one line giving the number of sequences it contains. Also for every component separately, column Conservation Profile depicts the degree of sequence conservation at the individual sequence positions, as described in section 3.4.1 of this work. The leftmost filled bar is a scale which marks complete conservation (only one base permitted). The open bars correspond to the single positions in the structure.





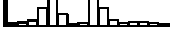
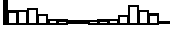


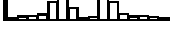
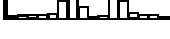



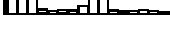

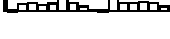




For the structure ranked 3 the conservation profile is missing because due to a system failure the sequence information on this network was lost. The component decomposition was already completed at this time. For the open structure (rank 1) no component decomposition was done because of the huge size of this network.

Rank	Size	Structure	SOC	Conservation Profile
2	52505831	(((...))).....	52505831	
3	52376319(((...)))	52376319	
4	44544114(((...)))	44544114	
5	44273746	((((...))).....	44273746	
6	33131192	..(((...))).....	33131192	
7	32883686(((...)))..	32883686	
8	32878614	..(((...))).....	32878614	
9	32800711(((...))).	32800711	
10	31738681	...(((...)))....	31738681	

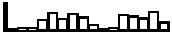



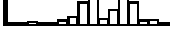





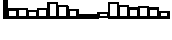






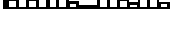


Rank	Size	Structure	SOC	Conservation Profile
11	31720954(((...)))...	31720954	
12	27886795	..((((...)))...	27886795	
13	27835512	..((((...)))....	27835512	
14	27791612((((...)))..	27791612	
15	27778147	...((((...)))..	27778147	
16	26952613((((...)))	26952613	
17	26723146	(((((...)))....	26723146	
18	24213789	...((((...)))	24213789	
19	24047941	(((((...)))...	24047941	
20	23939940	...((((...)))	23939940	
21	23718569	(((((...)))...	23718569	
22	23403940((((...)))	23403940	
23	23381252	(((((...)))....	23381252	
24	23298936	..((((...)))	23298936	
25	23090344	(((((...)))..	23090344	
26	22549750((((...)))	22549750	
27	22392809	(((((...)))....	22392809	
28	20122591	..((((...)))	20122591	
29	20007293	(((((...)))..	20007293	
30	19504439((((...)))	19504439	

Rank	Size	Structure	SOC	Conservation Profile
31	19317742	((.....))....	19317742	
32	17483821	..((((.....)))))	17483821	
33	17350407	(((((.....))))).)	17350407	
34	16835453	..((((.....)))).	16835453	
35	16820867	..((((.....)))).	16820867	
36	16806328	...((((.....)))).	16806328	
37	16302709	(((((.....))))))	16302709	
38	16070603	...((((.....)))).	16070603	
39	16018755	((.....))....	16018755	
40	15189906	..((((.....)))).	15189906	
41	15149248	..((((.....)))).	15149248	
42	14944802	..((((.....)))).	14944802	
43	14907515	..((((.....)))).	14907515	
44	14618572	..((((.....)))).	14618572	
45	14571453	..((((.....)))).	14571453	
46	14555798	..((((.....)))).	14555798	
47	14481816	..((((.....)))).	14481816	
48	14470119	..((((.....)))).	14470119	
49	14445167	(((((.....))))))	14445167	
50	14137170	..((((.....)))).	14137170	

Rank	Size	Structure	SOC	Conservation Profile
51	14102958	..(((.....))....	14102958	
52	13966925(((.....))..	13966925	
53	13955566	..((((.....))))	13955566	
54	13940500(((.....)).	13940500	
55	13879652	(((((.....))))).	13879652	
56	13597429	...(((.....))....	13597429	
57	13387645	((...)).....	13387645	
58	13312469((...))	13312469	
59	12148436	..(((.....)))	12148436	
60	12039830	..(((.....))....	12039830	
61	12010780	((.....))..	12010780	
62	11984481	..(((.....))..	11984481	
63	11911783	...(((.....)).	11911783	
64	10966289	..((((.....)))).	10966289	
65	10813722	..((((.....))))	10813722	
66	10775407	(((((.....))))).	10775407	
67	9910874	..(((.....))..	9910874	
68	9890910	..(((.....)).	9890910	
69	8412124	(((((.....))))	8412124	
70	8240900	..((...)).....	8240900	











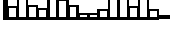
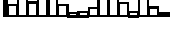

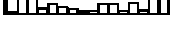
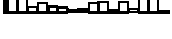





Rank	Size	Structure	SOC	Conservation Profile
71	8233388	..((...)).....	8233388	
72	7926178((...))..	7926178	
73	7913718((...)).	7913718	
74	7871294	.(((.....)))	7871294	
75	7787794	...((...)).....	7787794	
76	7748154	((.....)).	7748154	
77	7671761((...))...	7671761	
78	7550223	.(((.....))).	7550223	
79	7429530((...)).....	7429530	
80	7417372((...))....	7417372	
81	6426309	((.....))	6426309	
82	3757560((.....))	3757560	
83	3661228	((.....)).....	3661228	
84	3360679	((.....)).....	3360679	
85	3315282((.....))	3315282	
86	2920428	.(((.....))).	2920428	
87	2907344	((.....)).	2907344	
88	2567934((.....))	2567934	
89	2506239	((.....)).....	2506239	
90	2446866	..((.....)).....	2446866	








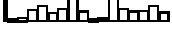



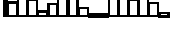
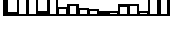



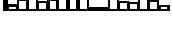



Rank	Size	Structure	SOC	Conservation Profile
91	2416639	.((.....)).....	2416639	
92	2359424((.....))..	2359424	
93	2329003((.....))	2034559	
			294444	
94	2327028	...((.....))...	2327028	
95	2320403((.....))...	2320403	
96	2286335((.....)).	2286335	
97	2254841	((.....)).....	1906756	
			348085	
98	2247197	..((.....))....	2247197	
99	2183285((.....))..	2183285	
100	2147037	.((.....)).....	2147037	
101	2119422	...((.....))...	2119422	
102	2035408((.....)).	2035408	
103	1930617	.(((.....)))..	1930617	
104	1916748	(((((.....))))..	1916748	
105	1724574	..((.....))...	1724574	
106	1687071	.((.....)).....	1687071	
107	1667250	...((.....))..	1667250	
108	1596729((.....)).	1596729	

Rank	Size	Structure	SOC	Conservation Profile
109	1459865	...((...(...)).)	1459865	
110	1442776	((...(...)).)...	1442776	
111	1392308((....)).	1214658	
			177650	
112	1391908((....))..	1208600	
			183308	
113	1376823	..((....)).....	1188621	
			188202	
114	1347136	..((....)).....	1160028	
			187108	
115	1327087	((...(....)).))	1327087	
116	1314943((....))...	1139018	
			175925	
117	1310120	...((....)).....	1130163	
			179957	
118	1279225((....))....	1104691	
			174534	
119	1188212	((...(....)).))	1188212	
120	1186672	((...(....)).))	1186672	
121	1079078	((...(....)).))	1079078	





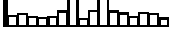






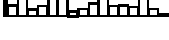





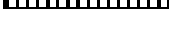


Rank	Size	Structure	SOC	Conservation Profile
122	871125	..(((.....)).)	871125	
123	870232	.(((.....)).)..	870232	
124	643992	(((((.....))).)	643992	
125	628013	((.....)).)	628013	
126	614810	.(((.....)).)	614810	
127	610722	((.....)).)	610722	
128	538933	..(((.....)).)	538933	
129	527991	((.....)).)..	527991	
130	520050	((.....)).)	520050	
131	513995	(((((.....))).)	513995	
132	506500	.(((.....))))	506500	
133	499795	.(((.....)).)	499795	
134	498944	((.....)).)	498944	
135	402368	..(((.....)).)	402368	
136	397521	..(((.....)).)	397521	
137	393097	((.....)).)..	393097	
138	391159	((.....)).)..	391159	
139	325629	.(((.....)).)	325629	
140	294304	((.....))))	294304	
141	279378	(((((.....))).)	279378	



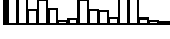











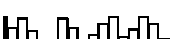




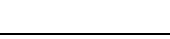
Rank	Size	Structure	SOC	Conservation Profile
142	269206	.((..((...))..))	269206	
143	261474	((..((...))..).	261474	
144	257506	((...))((...))	257506	
145	254456	((...))((...))	254456	
146	242619	.((..((...))..).	242619	
147	239708	(((((...)))..))	239708	
148	239684	.((..((...))..).	239684	
149	228354	(((((((...))))))	228354	
150	215088	(((((...)))..).	215088	
151	214178	.((((...)))..))	214178	
152	176046	(((((((...))))))	176046	
153	170988	((((...)))..)	170988	
154	164703	...(((...)))	164703	
155	164508	((..((....))..)	164434	
			74	
156	162629	(((((((....))))..))	162629	
157	162130	(((((....)))..))	162130	
158	156076	.((((....)))..)	156076	
159	150978	((((....)))..).	150978	
160	142764	(((((....)))..))	123564	

Rank	Size	Structure	SOC	Conservation Profile
			18455	
			745	
161	142052	..(((.(....).)))	125206	
			16363	
			483	
162	138627	((((.(....).)))..	121385	
			16695	
			547	
163	132525	..((.(....).))	132525	
164	130052	..(((.(....).))	130052	
165	128663	((((.(....).)).	128663	
166	128370	((((.(....).)).	128370	
167	121216	(((((.(...))))))	121216	
168	111269	..(((.(...)).))	111269	
169	110554	(((((.(...)).)).	110554	
170	104811	..(((.(...).)))..	104811	
171	102696	..(((.(...).))).	102696	
172	98232	(((((.(...))))))	98232	
173	96753	((((.(....).)).	96753	
174	87295	..(((.(....).))).	76755	





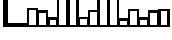











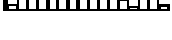



Rank	Size	Structure	SOC	Conservation Profile
			10222	
			318	
175	80707	((((...))))..	80707	
176	80001	..(((...)))	80001	
177	79411	((...(...))..)	79411	
178	79167	((...(...))..)	79167	
179	77950	(((((...))..))..)	77950	
180	73729	..((((...))..))	73729	
181	70180	..(((...))..)	70180	
182	69842	((...(...))..)	69842	
183	68986	..(((...))..)	68986	
184	67770	((...(...))..)	67770	
185	66447	..(((...)))	66447	
186	65322	((...(...))..)	65322	
187	64397	((...(...)))	64397	
188	57398	((...))..((...))	57398	
189	49885	..(((...)))	49885	
190	47840	(((((...)))..))	47840	
191	47610	..((((...)))	47610	
192	39277((...))	39277	





















Rank	Size	Structure	SOC	Conservation Profile
193	36267((.(....).))	34691	
			1576	
194	34175	((.(...).)....	34175	
195	33908	((.(....).)....	32055	
			1853	
196	32528	..((...).((...))	32528	
197	31533	..((...))((...))	31533	
198	31429	((...).((...)).	31429	
199	30367	((...))((...))..	30367	
200	30025	(((.(....))))	29934	
			91	
201	29699	.(((..(....).)))	26991	
			2708	
202	29428	(((..(....).))).	26777	
			2651	
203	28217	(((((....).)))	28217	
204	27475	((..(((....))))	27475	
205	27045	.(((.(....).)))	24890	
			2135	
			20	



Rank	Size	Structure	SOC	Conservation Profile
206	27012	(((.(....).)).)	24891	
			2102	
			19	
207	26436	((...(....).))	26436	
208	25285	(((.(....).)).)	23044	
			2240	
			1	
209	24925	..((.(....).))...	24925	
210	24267	...((.(....).))..	24267	
211	24003((.(....).)).	24003	
212	23712	..((.(....).))....	23712	
213	22954	((((.(....).))).)	22868	
			82	
			4	
214	22714	..(((.(....))))	22714	
215	22520	.((((.(....).))).)	22438	
			80	
			2	
216	22507	((((.(....))))).	22507	
217	22184	((((.(....).))...))	22184	

Rank	Size	Structure	SOC	Conservation Profile
218	20205	..((.(....).))..	19165	
			1040	
219	20203	(((....).))...	20203	
220	20181	...((.(....).)).	19217	
			964	
221	19764	.((.(....).))...	18733	
			1031	
222	16211	...(((....).))	16211	
223	15048	((....))((....))	15048	
224	14625	((....))((....))	13968	
			657	
225	14497	.((....))((....)).	14497	
226	13308	((....((....).))	13308	
227	12436	...((.(....)))	12436	
228	11647	((.(....))....)	11647	
229	11518	((....)).((....))	10226	
			1292	
230	11506	.(((....).))..	11506	
231	10880	..(((....).)).	10880	
232	10781	((..((....))))...	10781	

Rank	Size	Structure	SOC	Conservation Profile
233	10726	(((...)).(...))	8880	
			1846	
234	9955	(((((...).))))	9955	
235	7784	..((...)).	7784	
236	7318	((...))((...)).	6590	
			728	
237	7099	..((...))..	7099	
238	6885	((...))((...)).	5822	
			1063	
239	6739	..((...))((...))	6329	
			217	
			183	
			10	
240	6662	(((.....)).)	6662	
241	6466	..((...))((...))	5423	
			1043	
242	5765	(((....).))	5765	
243	5510	(((....).))	5510	
244	5455	(((.....)).)	5455	
245	5314	(((....).)..	4432	

Rank	Size	Structure	SOC	Conservation Profile
			882	
246	5234	(((.....))..)	5234	
247	5171	..(((.....))..)	4415	
			756	
248	4993	..(((.....))..)	4993	
249	4953	..(((.....))..)	4953	
250	4247	..(((.....))..)	3872	
			375	
251	4003	..(((.....))..)	2164	
			1839	
252	3905	(((.....))..)	1959	
			1946	
253	3830	...(((.....))..)	3830	
254	3767	(((.....))..)	3767	
255	3668	...(((.....))..)	3668	
256	3468	..(((.....))..)	2910	
			558	
257	3370	(((.....))..)	3370	
258	3283	(((.....))..)	3283	
259	3260	(((.....))..)	2903	

Rank	Size	Structure	SOC	Conservation Profile
			357	
260	3223	..((..((.....))))	3223	
261	3149	..(((.....)..)).	3149	
262	3052	((..((.....)))).	3052	
263	2581	..((..((.....)..))	2581	
264	2355	((..((.....)..)..	2355	
265	2309	..((..((.....)..)).	2309	
266	2252	..((..((.....)..)).	2252	
267	2219	..((..((.....)))).	2009	
			210	
268	2163	..((..((.....)..)..	2163	
269	2141	..((..((.....)..)..	2141	
270	1837	((.....))((.....))	1344	
			245	
			204	
			41	
			3	
271	1562	..((..((.....)..)).	1562	
272	795	(((((.....)...)))	795	
273	780	..(((.....)...)))	780	

Rank	Size	Structure	SOC	Conservation Profile
274	246	((.(.(...)).))	244	
			2	

Bibliography

- [1] Berkeley DB (version 2.4.14) <http://www.sleepycat.com/docs/ref/refs/refs.html>.
- [2] G. M. Adel'son-Vel'skii and A. E. M. Landis. An algorithm for the organization of information. *Soviet Mathematics Doklady*, 3:1259–1263, 1962.
- [3] K. Appel and W. Haken. Every planar map is four colorable. *Bulletin of the American Mathematical Society*, 82:711–712, 1976.
- [4] K. Appel and W. Haken. Every planar map is four colorable. *Illinois Journal of Mathematics*, 21:429–567, 1977.
- [5] D. Bashford, C. Chothia, and A. M. Lesk. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J Mol Biol*, 196:1:199–216, 1987.
- [6] S. Baskaran, P. F. Stadler, and P. Schuster. Approximate scaling properties of RNA free energy landscape. *J. Theor.Biol.*, 181:299–310, 1996.
- [7] E. Bornberg-Bauer. Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci U S A*, 96:19:10689–94, 1999.
- [8] A. Brennicke, A. Marchfelder, and S. Binder. RNA editing. *FEMS Microbiol Rev*, 23:3:297–316, June 1999.
- [9] F. Buckley and F. Harary. *Distances in Graphs*. Addison-Wesley, Reading, Ma., 1990.
- [10] T. Cech. Self-splicing of group I introns. *Ann. Rev. Biochem.*, 59:543, 1990.
- [11] T. Cech. Ribozyme engineering. *Current Opinion in Structural Biology*, 2:605, 1992.
- [12] C. Haslinger and P. F. Stadler. RNA structures with pseudo-knots. *Bull.Math.Biol.*, 61:437–467, 1999.
- [13] C. Darwin. *The Origin of Species*. reprinted in Penguin Classics, 1859.
- [14] J.-B. de Lamarck. *Philosophie zoologique, ou exposition des considérations relatives à l'histoire naturelle des animaux*. Paris, 1809.
- [15] M. Delbrück. A physicist looks at biology. *Trans. Conn. Acad. Arts Sci.*, 38:173–190, 1949.
- [16] M. Delbrück and S. E. Luria. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28:491–511, 1943.
- [17] R. Denman, C. Weitzmann, R. R. Cunningham, D. Negre, K. Nurse, J. Colgan, Y.-C. Pan, M. Miedel, and J. Ofengand. In vitro assembly of 30s and 70s bacterial ribosomes from 16s RNA containing single base substitutions, insertions, and deletions around the decoding site (c1400). *Biochemistry*, 28:1002–1011, 1989.
- [18] H. Driesch. *Der Vitalismus als Geschichte und Lehre*. J.A.Barth, Leipzig, 1905.
- [19] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 10:465–523, 1971.
- [20] M. Eigen, J. McCaskill, and P. Schuster. Molecular Quasi-Species. *Journal of Physical Chemistry*, 92:6881–6891, 1988.
- [21] M. Eigen, J. McCaskill, and P. Schuster. The molecular Quasispecies. *Adv. Chem. Phys.*, 75:149 – 263, 1989.

- [22] M. Eigen and P. Schuster. The hypercycle A: A principle of natural self-organization: Emergence of the hypercycle. *Naturwissenschaften*, 64:541–565, 1977.
- [23] M. Eigen and P. Schuster. *The Hypercycle: a principle of natural self-organization*. Springer, Berlin, 1979 (ZBP:234).
- [24] A. D. Farris, G. Koelsch, G. J. Pruijn, W. J. van Venrooij, and J. B. Harley. Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis. *Nucleic Acids*, 27:4:1070–8, 1999.
- [25] R. A. Fisher. *The genetical theory of natural selection*. Oxford: Clarendon Press., 1930.
- [26] C. Flamm, I. L. Hofacker, and P. Stadler. RNA in silico: The computational biology of RNA secondary structures. *Adv. Complex Syst.*, 2:65–90, 1999.
- [27] 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. *Monatshefte der Chemie*, 122:795–819, 1991.
- [28] W. Fontana, D. Konings, P. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33(9):1389, 1992.
- [29] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaption. *Physical Review A*, 40(6):3301–3321, 1989.
- [30] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophysical Chemistry*, 26:123–147, 1987.
- [31] W. Fontana and P. Schuster. RNA shaping space: The possible and the attainable in RNA genotype-phenotype mapping. *J.Theor.Biol.*, 194:491–515, 1998.
- [32] 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 combinatorial landscapes. *Phys. Rev. E*, 47(3):2083 – 2099, March 1993.
- [33] C. F. 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 Chacón, editors, *Advances in Artificial Life*. Springer Verlag, 1995.
- [34] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA*, 83:9373–9377, Dec. 1986.
- [35] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.
- [36] D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Chem. Phys.*, 81:2340–2361, 1977.
- [37] 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. *Monatsh. Chem.*, 127:355–374, 1996.
- [38] 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. *Monatsh. Chem.*, 127:375–389, 1996.
- [39] D. M. Halaby, A. Poupon, and J. Mornon. The immunoglobulin fold family: Sequence analysis and 3D structure comparisons. *Protein Eng.*, 12:7:563–71, 1999.

- [40] J. Haldane. A mathematical theory of natural and artificial selection. *Transactions of the Cambridge Philosophical Society*, 23:19–41, 1924.
- [41] R. W. Hamming. Error detecting and error correcting codes. *Bell Syst. Tech. J.*, 29:147–160, 1950.
- [42] L. H. Hansen, B. Vester, and S. Douthwaite. Core sequence in the rna motif recognized by the erme methyltransferase revealed by relaxing the fidelity of the enzyme for its target. *RNA*, 5:1:93–101, 1999.
- [43] P. G. Higgs. Compensatory neutral mutations and the evolution of RNA. *Genetica*, 1999. (in press).
- [44] I. L. Hofacker. *A Statistical Characterization of the Sequence to Structure Mapping in RNA*. PhD thesis, Universität Wien, 1994.
- [45] I. L. Hofacker, W. Fontana, P. F. S. L. S. Bonhoeffer, M. Tacker, and P. Schuster. *Vienna RNA Package*. <http://www.tbi.univie.ac.at/~ivo/RNA/>. (Public Domain Software).
- [46] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, 125(2):167–188, 1994.
- [47] I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 88:207–237, 1998.
- [48] M. Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43:165–169, 1996.
- [49] M. 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.
- [50] M. A. Huynen, A. S. Perelson, W. A. Vieira, and P. F. Stadler. Base pairing probabilities in a complete HIV-1 RNA. *J. Comp. Biol.*, 3:253–274, 1996.
- [51] P. Khaitovich, T. Tenson, A. S. Mankin, and R. Green. Peptidyl transferase activity catalyzed by protein-free 23S ribosomal rna remains elusive [letter]. *RNA*, 5:5:605–8, 1999.
- [52] M. Kimura and J. F. Crow. Natural selection and gene substitution. *Genet Res*, 13:2:127–41, 1969.
- [53] S. J. Klug. All you wanted to know about selex. *Mol Biol Rep*, 20:2:97–107, 1994.
- [54] S. Kopp. *RNA sequence to structure mapping*. PhD thesis, Universität Wien, 1998.
- [55] F. R. Kramer, D. R. Mills, P. E. Cole, T. Nishihara, and S. Spiegelman. Evolution *in vitro*: sequence and phenotype of a mutant RNA resistant to ethidium bromide. *J. Mol. Biol.*, 89:719–736, 1974.
- [56] T. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1962.
- [57] N. Leulliot, V. Baumruk, M. Abdelkafi, P. Y. Turpin, A. Namane, C. Gouyette, T. Huynh-Dinh, and M. Ghomi. Unusual nucleotide conformations in GNRA and UNCG type tetraloop hairpins: evidence from Raman markers assignments. *Nucleic Acids Res*, 27:5:1398–404, 1999.
- [58] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals (Russian). *Doklady Akademii Nauk SSSR*, 163(4):845–848, 1965.

- [59] D. J. Lipman and W. J. Wilbur. Modelling neutral and selective evolution of protein folding. *Proc R Soc Lond B Biol Sci*, 1991.
- [60] T. R. Malthus. *An Essay on the Principle of Population, as as It Affects the Future Improvement of Society*. J. Johnson, London, 1798. Darwin actually read the 6th ed. [1826]. London:Murray.
- [61] B. Mandelbrot. On the theory of word frequencies and on related markovian models of discourse. In R. Jakobson, editor, *The Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics*, volume 12, Providence, Rhode Island, 1961. American Mathematical Society.
- [62] E. Mayr. *The Growth of Biological Thought*. The Belknap Press of Harvard University Press, 1982.
- [63] G. Mendel. Versuche über Pflanzenhybriden. *Verh. Natur. Vereins Brünn*, 4:3–57, 1866.
- [64] E. D. Miller, K. H. Kim, and C. Hemenway. Restoration of a stem-loop structure required for potato virus X RNA accumulation indicates selection for a mismatch and a GNRA tetraloop. *Virology*, 260:2:342–53, 1999.
- [65] H. F. Noller. Ribosomal rna and translation. *Ann. Rev. Biochem.*, 60:191–227, 1991.
- [66] H. F. Noller, V. Hoffarth, and L. Zimniak. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science*, 256:1416–1419, 1992.
- [67] B. Palmquist. Heterogeneity and dispersion in the beta-binomial model. In *Annual Meeting of the American Political Science Association, Aug 27-31, 1997*, 1997.
- [68] R. L. Prentice. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Am. Stat. Assoc.*, 81:321–327, 1986.
- [69] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C; The art of scientific computing*. Cambridge University Press, 1993.
- [70] C. Reidys. *Neutral Networks of RNA Secondary Structures*. PhD thesis, Friedrich Schiller Universität Jena, Germany, 1995.
- [71] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.
- [72] C. M. Reidys, C. V. Forst, and P. Schuster. Replication and mutation on neutral networks. *Bull. Math. Biol.*, 1998. submitted.
- [73] J. Riordan. *An Introduction to Combinatorial Analysis*. Princeton University Press, 1978.
- [74] M. Sassanfar and J. W. Szostak. An RNA motif that binds ATP. *Nature*, 364:550–553, 1993.
- [75] D. Schneider, L. Gold, and T. Platt. Selective enrichment of RNA species for tight binding to *escherichia coli* rho factor. *FASEB J.*, 7:201–207, 1993.
- [76] D. Schneider, C. Tuerk, and L. Gold. Selection of high affinity RNA ligands to the bacteriophage R17 coat protein. *J. Mol. Biol.*, 228:862–869, 1992.
- [77] R. Schneider and C. Sander. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res*, 24:201–5, 1996.
- [78] P. Schuster. Landscapes and molecular evolution. *Physica D.*, 107:351–365, 1997.

-
- [79] 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.
- [80] P. Schuster and P. F. Stadler. Landscapes: Complex optimization problems and biomolecular structures. *Computers Chem.*, 18:295 – 324, 1994.
- [81] P. Schuster and P. F. Stadler. Sequence redundancy in biopolymers. A study on RNA and protein structures. In *Viral Regulatory Structures*, volume XXVIII of *SFI Studies in the Sciences of Complexity. Advances in HIV and HPV virus research.*, pages 163–186. Addison-Wesley, 1998.
- [82] P. Schuster and P. F. Stadler. Discrete models of biopolymers. TBI preprint 99-pks-012, 1999.
- [83] W. G. Scott. RNA catalysis. *Curr Opin Struct Biol*, 8:6:720–6, 1998.
- [84] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.
- [85] T. F. Smith, M. S. Waterman, and W. M. Fitch. Comparative biosequence metric. *J. Mol. Evol.*, 18:1:38–46, 1981.
- [86] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 17:213, 1971.
- [87] P. F. Stadler. Towards a theory of landscape. In R. López-Peña, R. Capovilla, R. García-Pelayo, H. Waelbroeck, and F. Zertuche, editors, *Complex Systems and Binary Networks (Proceeding of the Guanajuato Lectures)*, pages 77–163. Springer-Verlag, 1996.
- [88] M. Tacker. *Robust Properties of RNA Secondary Structure Folding Algorithms*. PhD thesis, University of Vienna, 1993.
- [89] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. l. Hofacker, and P. Schuster. Algorithm independent properties of RNA secondary structure predictions. *Eur. Biophys. J.*, 25:115–130, 1996.
- [90] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249:505, 1990.
- [91] E. van Nimwegen and J. P. Crutchfield. Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths ? SFI Working Paper 99-07-041. Submitted to Bull. Math.
- [92] E. van Nimwegen and J. P. Crutchfield. Optimizing epochal evolutionary search: Population-size dependent theory. SFI Working Paper 98-10-090.
- [93] E. van Nimwegen and J. P. Crutchfield. Optimizing epochal evolutionary search: Population-size independent theory. SFI Working Paper 98-06-046.
- [94] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. SFI Working Paper 99-03-021.
- [95] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell. Statistical dynamics of the royal road genetic algorithm. Santa Fe Institute Preprint 97-04-035.
- [96] F. W and S. P. Continuity in evolution: on the nature of transitions. *Science*, 280(5368):1451–5, May 1998.

-
- [97] A. Wagner and P. F. Stadler. Viral RNA and evolved mutational robustness. *J.Exp.Zool./MDE*, 285:119–127, 1999.
- [98] G. P. Wagner. Der dialog zwischen evolutionsforschung und computerwissenschaft. In *Die Evolution der Evolutionstheorie*, chapter 8, pages 221–233. Spektrum Akademischer Verlag, Heidelberg, Berlin, Oxford, 1994.
- [99] A. Walter, D. Turner, J. Kim, M. Lyttle, P. Mueller, D. Mathews, and M. Zuker. Coaxial stacking of helices enhances binding of oligoribonucleotides.. *PNAS*, 91:9218–9222, 1994.
- [100] J. Weber. *Dynamics of Neutral Evolution*. PhD thesis, Friedrich-Schiller-Universität Jena, 1997.
- [101] W. Wieser, editor. *Die Evolution der Evolutionstheorie*. Spektrum Akademischer Verlag, Heidelberg, Berlin, Oxford, 1994.
- [102] D. A. Williams. The analysis of binary responses from toxicological experiments involving reproduction and teratogeneity. *Biometrics*, 31:949–952, 1975.
- [103] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In D. F. Jones, editor, *int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.
- [104] S. Wright. Random drift and the shifting balance theory of evolution. In K. Kojima, editor, *Mathematical Topics in Population Genetics*, pages 1 – 31. Springer Verlag, Berlin, 1970.
- [105] F. M. Wuketits. *Biologie und Kausalität. Biologische Ansätze zur Kausalität, Determination und Freiheit*. Parey, Berlin, Hamburg, 1981.
- [106] B. Zhang and T. Cech. Peptidyl-transferase ribozymes: Trans reactions, structural characterization and ribosomal rna-like features. *Chem Biol*, 5:10:539–53, 1998.
- [107] G. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Reading(Mass.), 1949.
- [108] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, 1984.
- [109] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl. Acid. Res.*, 9:133–148, 1981.

Tabellarischer Lebenslauf

Name	Ulrike Göbel
Geburtstag, Geburtsort	07.03.1964 in Ludwigshafen/Rhein
Familienstand	ledig
Schulbildung	4 Jahre Grundschule, 9 Jahre Gymnasium, Abitur
Hochschulbildung:	
1983 - 1990	Universität Heidelberg, Biologie. Diplom Januar 1990 (Gesamtnote "sehr gut"). Thema der Diplomarbeit: Entwicklung von Programmen zur Erkennung und Repräsentation von Mustern in Protein-Primärsequenzen NF1-artiger Transkriptionsfaktoren
1990 - 1995	Fernuniversität Hagen, Mathematische Systemanalyse mit Nebenfach Informatik.
Winter 1994/95	Vordiplomprüfungen im Studiengang Mathematische Systemanalyse. Beendigung des Zweitstudiums.
Arbeitsverhältnisse:	
15.02.1990 - 15.05.1990	Werkvertrag am DKFZ Heidelberg mit dem Thema "Zeitoptimierung eines Algorith- mus' für das multiple Alignment von Sequenzen"
15.09.1990 - 31.09.1994	Werkverträge als Programmiererin in der Gruppe von Dr. Chris Sander, EMBL Heidelberg. Beteiligung an einem wissenschaftlichen Projekt der Gruppe.
Dissertation:	
01.07.1995 - 31.12.1998	IMB Jena, Abteilung "Molekulare Evolutionsbiologie" (Prof. Peter Schuster).
01.01.1999 - 31.01.2000	Selbständige Fortführung der Arbeit als Gast in den Arbeitsgruppen "Genomana- lyse" (IMB Jena, Prof. André Rosenthal) und "Genetik und Evolution" (Max- Planck-Institut für Chemische Ökologie Jena, Prof. Thomas Mitchell-Olds).
Titel	Neutral Networks of Minimum Free Energy RNA Secondary Structures

List of Publications

Chicken NFI/TGGCA proteins are encoded by at least three independent genes: NFI-A, NFI-B and NFI-C with homologues in mammalian genomes.

Rupp RA, Kruse U, Multhaup G, Göbel U, Beyreuther K, Sippel AE
Nucleic Acids Res 1990 May 11 18:9 2607-16

Correlated mutations and residue contacts in proteins.

Göbel U, Sander C, Schneider R, Valencia A
Proteins 1994 Apr 18:4 309-17

Structural constraints and neutrality in RNA

Göbel U, Forst CV, Schuster P

In : Ralf Hofestädt (ed) LNCS/LNAI Proceedings of GCB96, Lecture Notes in Computer Science, Berlin, Heidelberg, New York, 1997. Springer Verlag.