



universität
wien

DISSERTATION

Titel der Dissertation

A PHYLOGENETIC DEFINITION OF STRUCTURE

angestrebter akademischer Grad

Doktor/in der Naturwissenschaften (Dr. rer.nat.)

Verfasserin / Verfasser: Tanja Gesell
Matrikel-Nummer: 0647738
Dissertationsgebiet: Molekulare Biologie
Betreuerin / Betreuer: Prof. Arndt von Haeseler

Wien, am 21. August 2009

TO THE OTHER



... a rose is a rose is a rose is a can ... is rosa luxemburg ... is rosa parks ... is rosa v. praunheim (contra rosa and other perverse angels) ... is axel rose ... is guns'n roses ... is rosalind franklin ... a rose is miss stein's automatically handgun ... and rosa parks is the woman, who refused to stand up at 1.12.55 in montgomery, alabama, in a full bus as a black rose to give her place to a white ... the rose at the beginning of the civil right movement ... a rose is a rose is a rose is a can

this thesis would not have been with and without t. it's dedicated to the hypothetical alter ego. "copy me"

Abstract: A Phylogenetic Definition of Structure 0647738

What is a structure ?.....

The present thesis poses this question in the field of RNA molecular biology. While doing so, the aim is to contribute to the understanding of the intertwined relationship between structure, substitution process and evolutionary history. The thesis starts with an introduction into two fields: *RNA & phylogeny*, followed by the research chapters.

SISSI's Simulacrum, a framework for **SI**mulating **S**ite-**S**pecific **I**nteractions along phylogenetic trees, mimics sequence evolution under structural constraints in a unifying framework including arbitrary complex models of sequence evolution. This feeds into:

A Phylogenetic Definition of Structure, which consists of three aspects: The substitution matrix, a neighbourhood system and the phylogenetic tree. The substitution matrix specifies the evolutionary process of nucleotide evolution. However, the matrix is influenced by the neighbourhood system that defines the interactions among sites in a sequence. The phylogenetic tree introduces an additional dependency pattern in the observed sequences. In this chapter the general ideas of a **Phylogenetic Structure (PS)** are illustrated with examples. Consequently, this thesis focusses on particular approaches, devoting one chapter to each of the three aspects of a PS.

MATA's Neighbourhood System Aspect is considered in the context of so-called consensus structure from an alignment. Using the parametric bootstrap, **MATA**, **M**easurement of **A**ccurate **T**hresholds of **A**lignments, enables the detection of functionally associated correlations from a sequence alignment incorporating the phylogeny of the sequences combined with an automatic threshold procedure.

SISSIZ's Substitution Model Aspect is illustrated in the field of non-coding RNAs. We build up the SISSI framework to directly combine a new null model, based on a complex substitution model, with a consensus folding algorithm resulting in a new variant of a thermodynamic structure-based RNA gene finding program that is not biased by the dinucleotide content.

OSM's Phylogenetic Tree Aspect introduces another view on sequence evolution. The **One Step Mutation Matrix** encodes the phylogenetic tree directly and leads to analytical formulae for the posterior probability distribution of the number of substitutions for an alignment column. So far, our phylogenetic definition of structure has specified the evolutionary process of nucleotide evolution with site-specific interactions. Here, the definition is discussed as a description in pattern space.

The outlook discuss the (in)completeness of a phylogenetic definition of structure. However, the three approaches to each aspect including SISSI provide a very promising possibility to unite all three aspects of a PS together. Finally, the thesis concludes with a description of combining these methods towards structure evolution, which revers back to the original question:What is a structure ?

Abstract: Eine phylogenetische Definition von Struktur

Was ist eine Struktur?

Die vorliegende Doktorarbeit stellt diese Frage im Feld der RNA-Molekularbiologie. Ziel ist es, einen Beitrag zum Verständnis der eng miteinander verwobenen Beziehungen zwischen Struktur, Substitutionsprozess und Evolutionsgeschichte zu leisten. Die Darlegung beginnt mit einer Einführung in zwei Felder: *RNA & Phylogenie*, anschließend folgt die wissenschaftliche Analyse.

SISSI's Simulacrum ist ein Konzept für **SI**mulating **S**ite-**S**pecific **I**nteractions, in dem erstmalig Sequenzevolution sowohl entlang der Äste phylogenetischer Evolutionsbäume als auch in ihren strukturellen Abhängigkeiten bei Verwendung arbiträrer, komplexer Modelle der Sequenzevolution imitiert wird. Daraus resultiert:

Eine phylogenetischen Definition von Struktur, bestehend aus den drei folgenden Aspekten: Substitutionsmatrix, Nachbarschaftssystem und phylogenetischer Baum. Die Substitutionsmatrix spezifiziert den evolutionären Prozess der Nukleotidevolution und wird durch das Nachbarschaftssystem beeinflusst, das die Interaktion zwischen den Seiten innerhalb einer Sequenz bestimmt. Der phylogenetische Baum führt ein weiteres Abhängigkeitsmuster ein. In diesem Kapitel werden die grundlegenden Paradigmen einer **Phylogenetischen Struktur (PS)** mittels Beispielen illustriert. Daran anschließend werden einzelne Ansätze zur Erfassung einer PS aufgezeigt, wobei jedem der drei Aspekte der PS jeweils ein Kapitel gewidmet wird.

MATA's Nachbarschaftssystem wird der Konsensusstrukturvorhersage eines Alignments gegenüber gestellt. **MATA (Measurement of Accurat Thresholds of Alignments)** ermöglicht die Vorhersage funktional verknüpfter Korrelationen eines Sequenz Alignments mittels einer parametrischen Bootstrap Methode, sowohl unter Berücksichtigung phylogenetischer Aspekte als auch in Bezug auf einen automatisch ermittelten Schwellenwert.

SISSIZ's Substitutionsmodell wird im Feld nicht-kodierender RNAs veranschaulicht. **SISSIZ** kombiniert ein neues Null-Modell, ein komplexes Substitutionsmodell unter **SISSI**, mit dem konsensusbasierten Faltungsalgorithmus und produziert mittels thermodynamischer Strukturvorhersage und unter Berücksichtigung der Dinukleotidzusammensetzung eine neue Variante eines *RNA Gene finders*.

OSM's Phylogenetischer Baum führt einen weiteren Gesichtspunkt der Sequenzevolution ein. Die **One Step Mutation Matrix** kodiert den phylogenetischen Baum direkt und führt zur analytischen Formel für die Posterior-Wahrscheinlichkeitsverteilung bezüglich der Anzahl der Substitutionen pro Alignmentsspalte. Wurde bisher der evolutionäre Prozeß der Nukleotidevolution anhand der phylogenetischen Strukturdefinition mit seitenspezifischen Interaktionen spezifiziert, wird nun die phylogenetische Strukturdefinition als eine Beschreibungsform im Raum der Muster diskutiert.

Im Ausblick wird die (Un)vollständigkeit der phylogenetischen Strukturdefinition dargelegt. Dennoch sind die drei aufgezeigten Methoden, einschließlich **SISSI**, eine viel versprechende Möglichkeit, die drei Aspekte einer PS zu verbinden. Die vorliegende Arbeit reflektiert abschließend die Kombination aller drei Aspekte zu einer möglichen Beschreibung von Strukturevolution und führt zur Ausgangsfrage zurück: Was ist eine Struktur?

Preface

Parts of this thesis have been published in the following articles:

- (i) T. Gesell and A. von Haeseler (2006) In silico sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics*, 22, 716-722.
- (ii) T. Gesell and S. Washietl (2008) Dinucleotide Controlled Null Models for Comparative RNA Gene Prediction. *BMC Bioinformatics*, 9, 248.
- (iii) M. Dehmer, F. Emmert-Streib, and T. Gesell (2008) A comparative analysis of multidimensional features of objects resembling sets of graphs. *Appl. Math. Comput.*, 196, 221-235.
- (iv) S. Klaere, T. Gesell, and A. von Haeseler (2008) The impact of single substitutions on multiple sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 363, 4041-4047.
- (v) S. Washietl, T. Gesell, Chapter: Graph Representations and Algorithms in Computational Biology of RNA Secondary Structure, In: *Structural Analysis of Networks*, M. Dehmer (Editor), Birkhäuser Publishing, 2009, in press.

In preparation:

- (i) B. Zimmermann, T. Gesell, D. Chen, C. Lorenz and R. Schroeder
Evolution of RNA Sequences Under SELEX Constraints.
- (ii) C. Lorenz *et. al.* Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts.
- (iii) T. Gesell, T. Schlegel, M. Machatti and A. von Haeseler
Reconstructing site-specific interactions with phylogenetic trees.

The notation of each chapter is selfconsistent.

Each chapter is devoted to a friend and an inspiration.

Contents

1	Intention	1
1.1	STRUCTURE	1
1.2	Objectives of this Work	1
2	The Two Fields of This Study	5
2.1	RNA Structure	6
2.2	Phylogeny	20
2.3	The Two Fields Crossing	33
3	SISSI's Simulacrum	41
3.1	<i>In Silico</i> Sequence Evolution with Site-Specific Interactions	42
3.2	Extension of The Framework	53
3.3	SISSI with Energy	55
3.4	SISSI with Indels	59
3.5	The Sublime by SISSI	64
4	A Phylogenetic Definition of Structure	67
4.1	A Phylogenetic Definition of Structure	68
4.2	Realisations of Phylogenetic Structure	69
4.3	Structural Constraints	72
4.4	(Un)structured RNAs	82
4.5	The Definition Paradox	84
5	MATA's Neighbourhood Aspect	91
5.1	Measurements of Accurate Thresholds of Alignments	92
5.2	MATA's First-Test: Estimating Pairwise Correlations	94
5.3	MATA's Second-Test: Positions without Ancestry	95
5.4	Reconstructing Site-Specific Interactions with Phylogenetic Trees	96
5.5	Double Life	103

6	SISSIZ's Model Aspect	105
6.1	Background Models for RNA Gene Prediction	106
6.2	Randomising Genomic Alignments	108
6.3	SISSIZ: First Dinucleotide Based RNA Gene Finder	122
6.4	The Beauty of Elephants	126
7	OSM's Tree Aspects	129
7.1	OSM: One Step Mutation Matrices	130
7.2	A Phylogenetic Definition in Pattern Space	142
7.3	Awesome Times	144
8	Outlook	149
8.1	(In)complete Phylogenetic Definition of Structure	150
8.2	Towards Structure Evolution ('Happy Together')	153
8.3	WHAT IS A STRUCTURE	158
	Appendix	163
	PS	163
	Material and Methods	169
	Bibliography	190
	List of Chapter Figures and Quotes	193
	CV	194
	Acknowledgment	200

Chapter 1



I. Wiener & D. Roth

Intention

“A structure of a string is a Turing machine (TM) that, with or without input, is capable of printing the string (a string, then, viewed as a trivial TM, is a structure of itself); alternatively: a (part of a) TM that “accepts” the string (that is, the function it computes is defined on the string); alternatively: a TM computing a Boolean function and affirming the string.” (Wiener, 1994).

TO OSWALD

1.1 STRUCTURE

Structure is a much-used word. But what is it? Although, for example poststructuralist theory (cf. Deleuze, 1973) has focussed on structure, it seems that there is no possibility in forming, or there is a lack of a clear general definition of structure. Indeed, the lack of consensus within this literature raises a number of questions regarding the possibility of providing a single general definition of structure. Is there a single concept of structure or just different concepts of structure found in different disciplines? Are structures discovered or are they invented? Are they simply patterns, or do they offer a more profound understanding of the properties of different entities? Is there a single, correct structure with which each entity is associated, or is there a range of different potential structures, with pragmatic considerations determining the assignment of particular structures to particular entities? Is there a general challenge in thinking about structure? How do scientists apply structure today? What is a structure?

1.2 Objectives of this Work

A fundamental part of life science is the search for structure definitions. This thesis focuses these questions in the field of molecular biology. More precisely:

“What is an RNA structure ?” At first sight, it seems obvious what an RNA structure is, but going deeper we end up with similar questions like the above. However, RNA structure seems to be an appropriate, illustrative and fascinating example to focus on

definitions of structure in an individual scientific discipline. Especially, RNA secondary structure is considered as one of the best compromises between theoretical tractability and empirical accessibility on a large scale (Fontana, 2002). However, after a first glance, we note that so far the phylogenetic viewpoint has not been noted much, e.g. especially in the field of structure prediction from a thermodynamic viewpoint. A lot of work has been done with context-free grammars, but without taking thermodynamic into account. One reason is that the fields of structure prediction and phylogenies apparently still differ and are disjoint. A general aim of this thesis is to close this gap. Therefore, we start with an introduction into two fields:

RNA & phylogeny. First we consider the commonly used definitions of RNA structure and some background information on biochemical properties of RNA. To contribute to the understanding of the intertwined relationship between structure, the substitution process and evolutionary history, substitution models are explained. Chapter three starts with our first research chapter:

SISSI's Simulacrum, a framework for *SI*mulating *SI*te-*SI*pecific *SI*nteractions, was developed. Based on a concept of a well defined neighbourhood system, SISSI simulates the evolution of a nucleotide sequence along a phylogenetic tree taking into account user defined site-specific interactions. However, due to the stochastic nature of the substitution process it is likely to observe sequences in the course of evolution that exhibit – at least temporarily – a “structure” that deviates substantially from the “user” defined “structure”. Thus, it is possible to mimic sequence evolution under structural constraints in a unifying framework including arbitrary complex models of sequence evolution. The SISSI framework feeds directly into chapter four.

A Phylogenetic Definition of Structure, which consists of three aspects: The substitution matrix, a neighbourhood system and the phylogenetic tree. The substitution matrix specifies the evolutionary process of nucleotide evolution. However, the matrix is influenced by the neighbourhood system that defines the interactions among sites in a sequence. The phylogenetic tree introduces an additional dependency pattern in the observed sequences. In this chapter the general ideas of a *Phylogenetic Structure* (PS) are illustrated with examples. Using SISSI's simulations we consider the diversity of realisations under different prediction methods of realistic, as well as unrealistic PSs. Measuring structural conservation is important for *structure* prediction of e.g. non-coding RNAs. We show that a PS can generate sequences, which although they show no structure conservation on one realisation level (e.g. with respect to thermodynamic structure conservation), they might show on another realisation level very high structure conservation. Therefore, we discuss commonly used structure definition in correlation to a PS and the basic necessity of a *phylogenetic structure conservation index*.

Then, this thesis focusses on particular examples, respectively approaches, devoting one chapter to each of the three aspects of the phylogenetic definition.

MATA's Neighbourhood System Aspect is considered vis-à-vis so-called consensus structure from an alignment. Following a PS a structural constraint is clearly divided into ancestral and neighbourhood constraints versus ancestral and functional correlations. In the so-called consensus structure prediction methods, ancestral correlations represent false positive predictions, while functional correlations represent the true positives. This leads to a bias in comparative methods due to the fact that biological sequences are generally related by a phylogenetic tree. A method, MATA (*Measurement of Accurate Thresholds of Alignments*), is proposed using the parametric bootstrap that enables the detection of functional associated correlations from a sequence alignment incorporating the phylogeny of the sequences combined with an automatic threshold procedure. MATA's principle appears to be an useful complement to other existing tools and flexible enough to include other programs. For the combination with thermodynamic methods a model for linked chains in covariations, called overlapping dependencies, is necessary. Such a model is presented in the next chapter.

SISSIZ's Substitution Model Aspect is illustrated in chapter six. With an example we want to show that a phylogenetic definition is not an academic gimmick. Thus, we illustrate the usefulness of a complex substitution null model. Structure prediction programs, in particular those using a thermodynamic folding model including stacks, can be influenced by the genomic dinucleotide content. As a consequence, it is difficult to accurately estimate the false discovery rate of such genomic screens. There is need for a null model that considers this dinucleotide content, or more practically speaking, an algorithm to randomize genomic alignments that preserve the dinucleotide content while removing any correlations arising from RNA structures. While there have been algorithms to randomize single sequences preserving dinucleotide content for more than twenty years the problem was not solved for multiple alignments. We extended the SISSI framework from chapter three to directly combine a new null model with a consensus folding algorithm resulting in a new variant of a thermodynamic structure based RNA gene finding program called SISSIZ (*SISSI & ALIfoldz*) that is not influenced by the dinucleotide content. Finally, we discuss our algorithm for other applications.

OSM's Phylogenetic Tree Aspect introduces another view on sequence evolution. While so far the matrix is influenced by the neighbourhood system, here the phylogenetic tree defines the substitution matrix. This *One Step Mutation* matrix leads to analytical formulae for the posterior probability distribution of the number of substitutions for an alignment column and offers a variety of potential applications in molecular systematics. The OSM matrix encodes the phylogenetic tree directly and leads to an evaluation of paths through pattern space. While in chapter four the phylogenetic definition of structure specifies the

evolutionary process of nucleotide evolution with site-specific interactions, here the phylogenetic definition is discussed as a description in pattern space.

The **Outlook**, chapter nine, starts with a discussion about the (in)completeness of a phylogenetic definition of structure. A main goal for future research is to include all aspects directly into one matrix to describe all three aspects simultaneously, though it is neither clear if that is possible nor if it would explain a PS adequately. However, **SISSI** and the three approaches, each devoted to one aspect of a PS, provide very promising possibilities. We give a description of the next important steps, combining these approaches. The outlook ends with general comments about the unifying framework for other character states than RNA and comes back to the original question:What is a structure ?

Chapter 2

The Two Fields of This Study



Max, Cologne (2008)

*As late as the Middle Ages, the witch was still the Hagazussa, a being that sat on the Hag, the fence, which passed behind the gardens and separated the village from the wildness. She was a being who participated in both worlds. As we might say today, she was semi-demonic. In time, however, she lost her double features and evolved more and more, into a representative of what was being expelled from culture, only to return, distorted, in the night (Duerr, 1985). **TO THOMAS***

James Watson, one of the researcher that discovered the structure of the Double Helix, begins his text book *Molecular Biology of the Gene* in 1965 with an euphoric introduction into evolution to underline its general importance. Today, 200 years after Darwin (1809-1882), the evolutionary viewpoint is still emphasised: *Nothing in Biology Makes Sense Except in the Light of Evolution* by the geneticist Theodosius Dobzhansky (Dobzhansky, 1973) is one of frequently cited sentences at conferences today. However, in particular, Griffiths (2008) argued that an evolutionary perspective is indeed necessary, but that it must be a *forward-looking perspective* formed by a general understanding of the *evolutionary process*, not a *backward-looking perspective* formed by the specific evolutionary history of the species being studied. In practice, this viewpoint is often still missing from other communities which use phylogeny. For example, we notice a lack between the communities of *RNA & Phylogeny* in Bioinformatics at the starting point of this thesis. Taking this into account, this chapter introduce both fields. First, on RNA structure and its terminology. We will present a timeline of RNA research and an overview of existing structure definitions. In more detail, we will start with some background information on biochemical properties of RNAs, as well as the commonly used classification of structural elements in RNAs as formalized by a definition of RNA secondary structures. After introducing RNA sequences and describing their biological functions, the second part of this chapter is about *evolutionary relationships*. Here, the function implies a structuring of the sequences, which can be described by sequence evolution models. Finally, some concepts are discussed as a potential link between phylogeny and RNA from a thermodynamic viewpoint.

2.1 RNA Structure

Traditionally, **structural biology** has a clear focus on protein structures. This does not come as a surprise, given the high complexity of protein structures and their diverse functional spectrum. However, in recent years it has become evident that the biological importance of RNAs has been vastly underestimated. The first description of “ribozymes” showed that RNAs can catalyze biochemical processes, an activity which before was only known for protein enzymes (Guerrier-Takada *et al.*, 1983; Kruger *et al.*, 1982). The discovery of micro RNAs led to a paradigm shift in our understanding of gene regulation (Bartel, 2004). New high-throughput experimental techniques, as well as computational predictions, suggest that there are tens of thousands of so far unrecognized RNAs in mammalian cells (Washietl *et al.*, 2005a, 2007a; Carninci *et al.*, 2005). Although it is still unclear whether it is justified to proclaim a modern “RNA world”, there is no doubt that RNAs have to be considered as important key players in the cell and that structural biology of RNAs will be of particular importance in the next years.

This fascinating story is summarized in Table 2.1. However, this thesis focuses on the basics of the computational approach to RNA structure. Figure 2.1 shows the principles of RNA structure, which will be extended in Table 2.2 to the important definitions, e.g. of the minimum free energy structure (mfe) and further important RNA structure definitions given today. Their importance and their basics will be explained in more detail in the next part of this chapter.

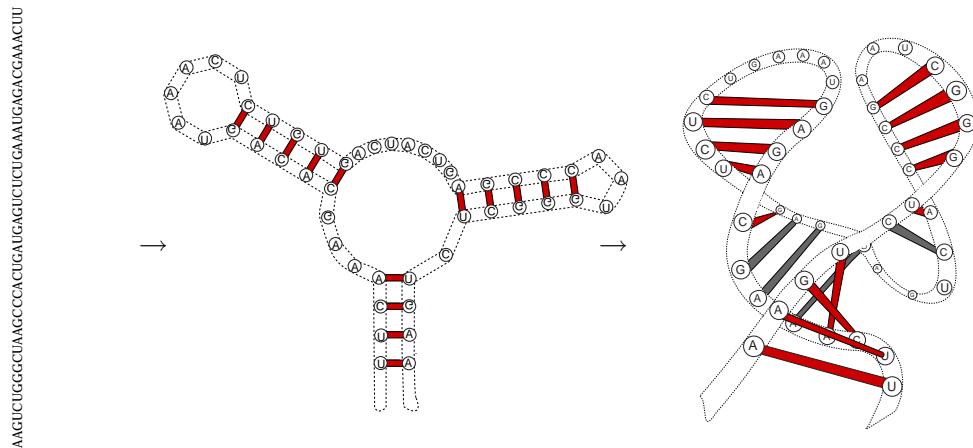


Figure 2.1: Principles of RNA structure: The primary structure (left) is defined by the succession of the three different nucleotide types A,C,G,U. Pairing patterns between AU,GC and GU form the secondary structure (middle). The secondary structure elements interact with each other in a complex three dimensional pattern, the tertiary structure (right). Note that in the tertiary structure non-standard base-pairs can occur (gray) that are usually not considered in algorithms analyzing the secondary structure. The example shows a so-called hammerhead ribozyme, a short self-cleaving RNA.

RNA-Timeline

- 1869** Nucleic acids were first described by Miescher (1969).
- 1909** Some nucleic acids contain ribose found by Phoebus Levene (see review by Choudhur (2003)).
- 1941** Cellular sites of protein synthesis are rich in RNA (Brachet, 1941)
- 1953** Description of the **DNA double helix** (Watson and Crick, 1953).
- 1954** RNA was suspected to play a role in information transfer from DNA to proteins (Rich and Watson, 1954).
- 1957** Discovery of tRNA (Hoagland *et al.*, 1958, 1957).
- 1964** Deciphering of the genetic code (Nirenberg *et al.*, 1965).
- 1970** The Central Dogma of molecular biology formulated by Francis Crick (Crick, 1970, 1958). The Central Dogma is a set of rules about the direction that information can flow in a cell. It basically states that there cannot be an information transfer originating from protein. The Central Dogma is not concerned with the role of RNA (DNA, Proteins) in the cell. However, interpretations by the scientific community at large changed the meaning of the central dogma, and RNA was reduced to a carrier of information.
- 1975** First algorithms were developed to predict RNA secondary structure (Pipas and McMahan, 1975).
- 1978** Fast algorithm to predict secondary structure (Nussinov *et al.*, 1978; Nussinov and Jacobson, 1980)
- 1981** Based on the definition of secondary structure, an algorithm was developed to predict the minimum free energy (mfe) structure (Zuker and Stiegler, 1981a).
- 1982** Discovery of catalytic activity of RNA in modern cells. For that, Sidney Altman and Thomas Cech won a Nobel prize (Guerrier-Takada *et al.*, 1983; Kruger *et al.*, 1982), but this was not entirely sufficient to remove the carrier of information label from RNA.
- 1990** The partition function by McCaskill (McCaskill, 1990b) was developed as the proper description of the RNA molecule at thermodynamic equilibrium or in the limit of infinite time.
- 1998** Discovery of RNA inference, as well as the miRNA pathways in animals and plants. This led to a paradigm shift of our understanding of gene regulation and scientist's perception of RNA, 20 years after the Central Dogma of molecular biology.
- 2002** Ribosome is essential a Ribozyme. Ribosomes are essentially ribozymes. Ever since this discovery, RNA has received more attention by molecular biologists.
- Today** New high-throughput experimental techniques, as well as computational predictions, suggest that there are tens of thousands of so far unrecognized RNA (Washietl *et al.*, 2007a; Carninci *et al.*, 2005).
- Future** Emergence of a "Modern RNA world"
-

Table 2.1: Today RNA is in the focus of attention of many scientists and **structures of RNA molecules** are also of functional interest.

RNA Structure Definitions

Primary Structure is defined as the succession of the four different bases: adenine (A), cytosine (C), guanine (G), and uracil (U). RNA is a polymer made of **nucleotide units** consisting of a ribose group, a phosphate and one of the four different bases above.

Secondary Structure a list of base pairs that can be visualized by a planar graph. Please refer to the subsection 2.1.2.

Tertiary Structure is the three dimensional structure, which is formed through arrangements of the secondary structure elements in space.

Quaternary Structure is formed by inter-molecular base pairing and other interactions between two different molecules. The function of an RNA molecule often depends on its interaction with other RNAs.

Minimum Free Energy Structure (mfe) Based on the definition of secondary structure that can be calculated since 1981 (Zuker and Stiegler, 1981a), using a dynamic programming algorithm. Zuker’s **k-loop decomposition** identifies and classifies basic building blocks of a structure, e.g. hairpin and interior loops. The mfe is the structure with the most favourable folding energies. Usually, ΔG , the free energy of a folding relative to the unfolded sequence, is optimized.

Suboptimal Structures accompany the mfe structure and contribute to the molecular properties in the sense of a Boltzmann ensemble (see subsection 5.7). The partition function by McCaskill (McCaskill, 1990b) is the proper description of the RNA molecule at thermodynamic equilibrium or in the limit of infinite time.

Consensus Structure infers a common structure for two or more different RNA sequences. For example for rRNAs, tRNAs, and many other small non-coding RNA it is known “a priori” that the aligned sequences should fold into a common secondary structure. Various measures exist to measure the covariation between two positions in an alignment, ranking from simple mutual information to more advanced covariation measures (Lindgreen *et al.*, 2006). Comparative methods are in principle not limited to secondary structure. Furthermore, it is possible to consider the coevolution of sites intramolecular (within a single molecule) or intermolecular by taking each site in a distinct data set.

Coarse grained structures are abstracted from this main definitions: for example *abstract shapes* (Giegerich *et al.*, 2004; Steffen *et al.*, 2006; Voss *et al.*, 2006) and *coarse grained structure* defined by Ancel and Fontana (2000); Meyers *et al.* (2004).

Table 2.2: RNA-Structure Definitions: Please, refer to the text in this chapter for details. *Structure kinetic* is different to the other types of structure definitions taking the process of the kinetic Folding of RNA into account (e.g Flamm and Hofacker, 2008).

2.1.1 Structural Properties of RNA Molecules

RNA is a polymer made of individual units called *nucleotides* (cf. Nelson and Cox, 2004). Nucleotides consist of a ribose group, a phosphate group and one of four different bases adenine (A), cytosine (C), guanine (G), and uracil (U). The succession of the four different bases of the nucleotides defines the **primary structure** or **sequence** of the molecule (Fig. 2.1). Adjacent nucleotides in the primary sequence are connected by covalent bonds, i.e. strong chemical bonds that do not open under normal conditions. These bonds build the “backbone” of the molecule. RNA is generally single stranded, but complementary regions in the molecule can fold back onto itself and form double helices similar to the well-known DNA helix. In RNA, we usually find the so-called Watson-Crick pairs CG and AU, as well as GU “wobble pairs”. The intra-molecular base-pairing results in a pattern of double helical stretches interspersed with unpaired regions, which is called the **secondary structure**. Unlike the covalent bonds of the backbone, this base-pairing is made by weaker hydrogen bonds that can be opened and closed under physiological conditions. The arrangement of secondary structure elements in space finally forms the three-dimensional **tertiary structure**.

RNA folding is a hierarchical process. The secondary structure usually forms before and independently of the tertiary structure and contributes most of the stabilizing energy. The formation of tertiary structure usually does not induce changes in the secondary structure. The function of the molecule is ultimately dependent on the tertiary structure. However, secondary structure can serve as coarse-grained approximation and is an extremely useful level on which to understand RNA function.

2.1.2 Secondary Structure

An early graph-based definition of RNA secondary structure is due to Waterman (Waterman, 1978).

Definition 2.1.1. *A secondary structure is a vertex labeled graph on n vertices with an adjacency matrix $A = (a_{i,j})$ fulfilling:*

- (i) $a_{i,i+1} = 1$ for $1 \leq i \leq n - 1$
- (ii) For each $i \in \{1, \dots, n\}$ there is at most one $a_{i,j} = 1$ where $j \neq i \pm 1$.
- (iii) If $a_{i,j} = a_{k,l} = 1$ and $i < k < j$ then $i < l < j$

The first part defines the continuous “backbone” of the primary structure of the molecule. The second part defines the secondary structure interactions and allows each nucleotide (vertex) to be paired with at most one other nucleotide not immediately adjacent in the backbone. An edge of this type (i, j) , $j \neq i \pm 1$, is called a *base pair*. A vertex i connected only to $i \pm 1$ is called *unpaired*. The third part of the definition excludes interactions that

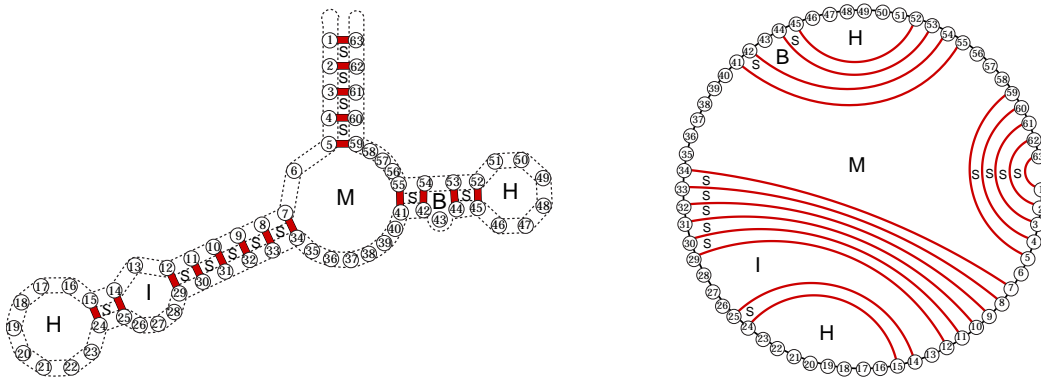


Figure 2.2: Graph representation and structural elements of RNA secondary structures. **H**: hairpin **I**: interior loop, **S**: stacked pair, **B**: bulge loop, **M**: multi loop. Left: conventional drawing of the structures as used by biochemists and molecular biologists. Right: circle representation emphasizing the graph-like nature of the secondary structure. The circle represents the backbone of the RNA. Each nucleotide is connected to its immediate neighbours within the backbone. In addition, each nucleotide can form one (and only one) base pair to another nucleotide (red arcs) or stay unpaired. The definition of RNA secondary structures excludes pseudoknotted structures, i.e. the arcs are not allowed to cross. The faces of the graph correspond to the different substructure elements.

are (somewhat arbitrarily) classified as tertiary structure interactions. In particular, this rule excludes structures known as “pseudoknots”.

The molecule geometry in RNA does not allow sharp bends with unpaired regions shorter than three. In practical applications, one usually adds additional biological reality to this definition by requiring $a_{i,j} = 0$ if $1 < j-i \leq 3$.

Secondary structure graphs following this definition are *outerplanar*, i.e. they have an embedding in the plane such that all vertices lie on the boundary of their exterior region (Fig. 2.2). The edges representing the base-pairs lie inside and do not cross.

Classification of Structural Elements

To describe and understand complex RNA secondary structures, biologists distinguish different structural elements. Thus, for formal treatment it is helpful to identify and classify the basic building blocks in a secondary structure. For Zuker’s structure prediction algorithm ((Zuker and Stiegler, 1981b), subsection 2.1.3) the so called *k-loop decomposition* (Zuker and Sankoff, 1984) is used.

Definition 2.1.2. A base k is called *immediately interior to the base-pair* (i, j) if $i < k < j$ and there is no other base pair (p, q) such that $i < p < k < q < j$.

Definition 2.1.3. The base pair (i, j) and all bases immediately interior to (i, j) are called a *loop closed by* (i, j) . The number of base-pairs contained in the loop (including the closing base-pair) is called *degree of the loop*.

Loops correspond to the faces of the outerplanar secondary structure graph (Fig. 2.2). Commonly used structural elements of RNA secondary structures can be defined using this formalism.

Definition 2.1.4. *Classification of structural elements:*

- A loop of degree 1 is called **hairpin**.
- A loop of degree 2 is called **interior loop**. Let (i, j) be the closing base-pair and (p, q) the base-pair immediately interior. There are two special cases of interior loops:
 - **stacked pair** if $p - i = 1$ and $j - q = 1$
 - **bulge** if $p - i > 1$ or $j - q > 1$ but not both.
- A loop of degree ≥ 3 is called **multi loop**.

Counting Secondary Structures

The combinatorial problem of counting secondary structures that can be formed by a sequence of a given length is particularly interesting. Its recursive solution was first noticed by Waterman (Waterman, 1978; Waterman and Smith, 1978) thirty years ago and it is the basis for many of the folding algorithms or for application to the landscape concept we describe later in this chapter.

Let x be a sequence of n nucleotides $x_i \in \{\text{A,C,G,U}\}$, $1 < i \leq n$. If we assume a specific sequence not all positions can pair, but only those following the base pairing rules for RNA structures (subsection 2.1.1). We use the base-pairing matrix Π with the entries $\Pi_{i,j} = 1$, $1 \leq i, j \leq n$, if sequence positions i and j can form a base pair, i.e., if (x_i, x_j) is in the set of allowed base-pairs $B = \{\text{GC, CG, AU, UA, GU, UG}\}$, and $\Pi_{i,j} = 0$ otherwise. Further, let $x_{i,j}$ be the sub-sequence from i to j , and $N_{i,j}$ the number of secondary structures that can be formed by $x_{i,j}$.

To calculate $N_{i,j}$, we assume that we already know $N_{i+1,j}$, i.e. the number of structures of a sub-sequence shorter by one base. A newly added base can either be unpaired or form a base-pair with some other base k . In the first case, the unpaired base is followed by any possible structure in sub-sequence $x_{i+1,j}$. In the latter case, the new base-pair divides the sequence in two sub-sequences $x_{i+1,k-1}$ and $x_{k+1,j}$. Since base-pairs do not cross, both sub-sequences can be treated independently and their numbers can be simply multiplied. These considerations lead to the following recursion:

$$N_{i,j} = N_{i+1,j} + \sum_{\substack{i+1 \leq k \leq j \\ \Pi_{ik} = 1}} N_{i+1,k-1} N_{k+1,j} \quad (2.1)$$

with $N_{ii} = 1$.

RNA secondary structure graphs lead to many other interesting combinatorial questions (e.g. Hofacker *et al.* (1998) and references therein) which are, however, not of immediate relevance for most practical applications in Bioinformatics.

Structure Prediction Using Simple Base-Pairing Rules

The “RNA folding” problem, i.e. the prediction of the secondary structure for a given primary sequence, is without doubt the most relevant problem for practical applications. Experimental determination of structures can be laborious and is not feasible on a large scale. Computational predictions are, therefore, widely used in the everyday analysis of RNAs.

Thermodynamic methods for RNA folding are the most established and most frequently used methods today. Put in simple terms, the goal is to find the structure with the most favourable folding energy. Usually, the free energy ΔG of folding relative to the unfolded sequence is used. Paired regions add stabilizing (by convention negative) energy contributions to ΔG , while unpaired regions add destabilizing (positive) energy terms.

The first attempts to calculate optimal secondary structures for simplified energy models were provided by Nussinov and co-workers (Nussinov *et al.*, 1978; Nussinov and Jacobson, 1980). In the simplest case, one assigns each type of base-pair a negative and fixed energy contribution. Then the problem reduces to finding the structure with the maximum number of base-pairs. In a more sophisticated (but still largely unrealistic scenario), one assigns each type of base-pair (i, j) a specific energy contribution $\beta_{i,j}$. The overall energy of a fold is the sum of all base-pair energies. In this model, we find the minimal energy $F_{i,j}$ of a sequence $x_{i,j}$ using a very strategy as used for enumerating all structures in Equ. (2.1). Adding one base at a time, either the new base is unpaired or it forms a pair with some base k . The overall minimum is the minimum of these two cases. To obtain the minimum of the latter case in which i forms a base-pair, all possible base-pairs (i, k) are evaluated. Each base-pair (i, k) separates the sub-sequence in two intervals and due to their independence, the minimum free energy can be obtained using the following recursion:

$$F_{i,j} = \min \left\{ F_{i+1,j}, \min_{\substack{i+1 \leq k \leq j \\ \Pi_{ik}=1}} \{ F_{i+1,k-1} + F_{k+1,j} + \beta_{ik} \} \right\} \quad (2.2)$$

This is an example of a dynamic programming algorithm frequently encountered in bioinformatics. A matrix containing the optimal solution for all possible subsequences is filled and the entry $F_{1,n}$ finally contains the optimal solution for the whole sequence of length n . The algorithmic complexity of this procedure is $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ in memory and CPU, respectively.

However, evaluating Equ. (2.2) gives only the minimum free energy and not the structure itself. A so-called *backtracking* or *backtracing* procedure is used to get the list of base-pairs

corresponding to the optimal energy. A helper matrix K is filled during the recursion. We set $K_{i,j} = k$, where k is the base which gives the optimal secondary structure when paired with i for a sub-sequence from i to j . If i is unpaired in the optimal structure, we set $K_{i,j} = 0$. We can then retrieve the list of base-pairs of the optimal structure using a simple recursive procedure as shown in Algorithm 2.1.1. We start with input $(i, j) = (1, n)$, i.e. we consider $K_{1,n}$ which holds the pairing partner k of position 1 in the optimal structure of the whole sequence of length n . $K_{1,n} = k$ divides the sequence in two independent sub-sequences which are evaluated by recursively calling the same function again.

Algorithm 2.1.1: Recursive backtracking procedure to retrieve the list of base-pairs in the optimal structure.

```

function Backtrack( $i, j$ )
  begin
    if  $i > j$  then return
    if  $K_{i,j} = 0$  then Backtrack( $i + 1, j$ ) else
      output: ( $i, K_{i,j}$ )
      Backtrack( $i + 1, K_{i,j} - 1$ )
      Backtrack( $K_{i,j} + 1, j$ )
    end
  end

```

2.1.3 MFE: Minimum Free Energy Structure

While structure prediction using simple base-pairing rules clearly gives the optimal structure in an algorithmic sense, structure predictions obtained this way are generally not biological realistic. The energy model based on simple base-pairing rules only poorly reflects the biophysical properties of real RNA molecules. Most of the stabilizing energy in RNA secondary structures comes from stacking interactions of neighbouring base-pairs. A realistic energy model, thus, needs to consider the *loops* in a structure (subsection 2.1.2). The so-called loop-based energy model or nearest-neighbour model assigns each loop l in a structure \mathcal{S} a free energy ΔG . The total free energy of the structure is the sum of all loops:

$$\Delta G(\mathcal{S}) = \sum_{l \in \mathcal{S}} \Delta G(l) \tag{2.3}$$

The energy rules used in current state-of-the art prediction programs are quite complex (Xia *et al.*, 1998; Mathews *et al.*, 1999) and it would be out of the scope of this thesis to present them in detail here. Generally, the energy depends on the type of the loop (see Def. 2.1.2), the size-of the loop, the closing base-pairs and the bases immediately interior to the closing base-pair. The energy values have been determined empirically using melting

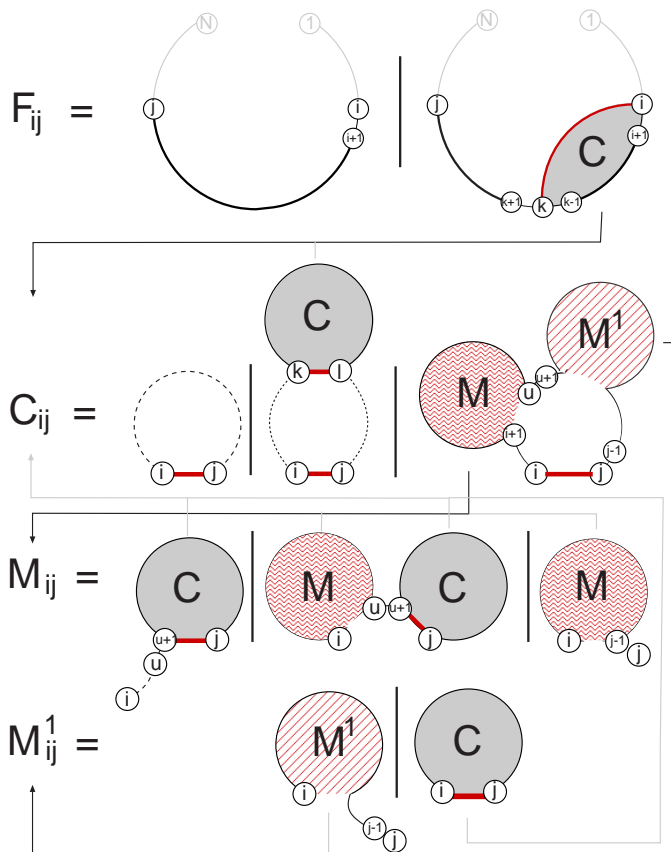


Figure 2.3: Illustration of the recursive structure decomposition steps in Zuker's folding algorithm. The property of a sequence with chain length N is built up recursively from the properties of smaller segments under the assumption that the contributions are additive. The procedure requires four matrices: $F_{i,j}$, $C_{i,j}$, $M_{i,j}$ and $M_{i,j}^1$ (cf. Equ. 2.4). The red lines indicate base-pairs, the dotted lines indicate unpaired substructures and the solid black lines indicate arbitrary structures. Please refer to the text for a detailed description of the procedure.

experiments that measure the energy which is required to open specific structural elements. Only stacks and some other small loops are tabulated exhaustively. The energy rules for other types of loops usually contain extrapolations and other approximations (Turner and Sugimoto, 1988).

In principle, one can find the minimum free energy model using a similar strategy as shown before. However, it is not sufficient to distinguish only two cases in the recursion. Instead, all possible loop types have to be considered in a systematic decomposition procedure. Recursions for this problem were first been proposed by Zuker and Stiegler (1981b). Here, we show a version following references (Hofacker and Stadler, 2007) that decomposes structures in such a way that each substructure is considered exactly once. Fig. 2.3 shows a graphical outline of the decomposition steps. The procedure requires four matrices. $F_{i,j}$ contains the free energy of the overall optimal structure of the subsequence $x_{i,j}$. The newly added base can be unpaired or it can form a pair. For the latter case, we introduce the helper matrix $C_{i,j}$ that contains the free energy of the optimal substructure of $x_{i,j}$ under the constraint that i and j are paired. This structure closed by a base-pair can either be a hairpin, an interior loop or a multi-loop. The hairpin case is trivial because no further decomposition is necessary. The interior loop case is also simple because it reduces again to the same decomposition step. The multi-loop step is more complicated. The energy of

a multi-loop depends on the number of components, i.e. substructures that emanate from the loop. To implicitly keep track of this number there is need for additional two helper matrices. $M_{i,j}$ holds the free energy of the optimal structure of $x_{i,j}$ under the constraint that $x_{i,j}$ is part of a multi-loop with at *least one* component. $M_{i,j}^1$ holds the free energy of the optimal structure of $x_{i,j}$ under the constraint that $x_{i,j}$ is part of a multi-loop and has *exactly one* component closed by pair (i, k) with $i \leq k < j$. The idea is to decompose a multi-loop in two arbitrary parts of which the first is a multi-loop with at least one component and the second a multi-loop with exactly one component and starting with a base-pair. These two parts corresponding to M and M^1 can further be decomposed into substructures that we already know, i.e. unpaired intervals, substructures closed by a base-pair, or (shorter) multi-loops. We can summarize the recursion as follows:

$$\begin{aligned}
F_{i,j} &= \min \left\{ F_{i+1,j}, \min_{i < k \leq j} (C_{ik} + F_{k+1,j}) \right\} \\
C_{i,j} &= \min \left\{ \mathcal{H}(i, j), \min_{i < k < l < j} (C_{kl} + \mathcal{J}(i, j; k, l)), \right. \\
&\quad \left. \min_{i < u < j} (M_{i+1,u} + M_{u+1,j-1}^1 + a) \right\} \\
M_{i,j} &= \min \left\{ \min_{i < u < j} ((u - i + 1)c + C_{u+1,j} + b), \right. \\
&\quad \left. \min_{i < u < j} (M_{i,u} + C_{u+1,j} + b), M_{i,j-1} + c \right\} \\
M_{i,j}^1 &= \min \{ M_{i,j-1}^1 + c, C_{i,j} + b \},
\end{aligned} \tag{2.4}$$

$\mathcal{H}(i, j)$ is the energy for a hairpin closed by base-pair (i, j) and $\mathcal{J}(i, j; k, l)$ the energy for an interior loop closed by the two base-pairs (i, j) and (k, l) (Zuker and Stiegler, 1981a). Multi-loop energies are approximated by a simple linear relationship: $E_{\text{ML}} = a + b \cdot \text{degree} + c \cdot \text{size}$. The constant a is used to penalize opening a multi-loop in the first place. The constant b and c penalize the number of components ("degree") and the size of unpaired intervals, respectively. Multi-loops are generally considered destabilizing. The constant a is used to penalize opening a multi-loop in the first place. The constant b and c penalize the number of components and the size of unpaired intervals, respectively. Using these recursions, the minimum free energy and — using an appropriate backtracking procedure — the optimal structure under the full loop-based energy model can be found. This approach is currently the most widely used method to predict RNA secondary structures. The most popular implementations are `mfold` (Zuker, 2003) and `RNAfold` from the Vienna RNA package (Hofacker *et al.*, 1994a).

Suboptimal Secondary Structure

At room temperature, the energy contributions from the base-pairing in a molecule is in the same order of magnitude as the thermal energy. As a consequence, base-pairs can

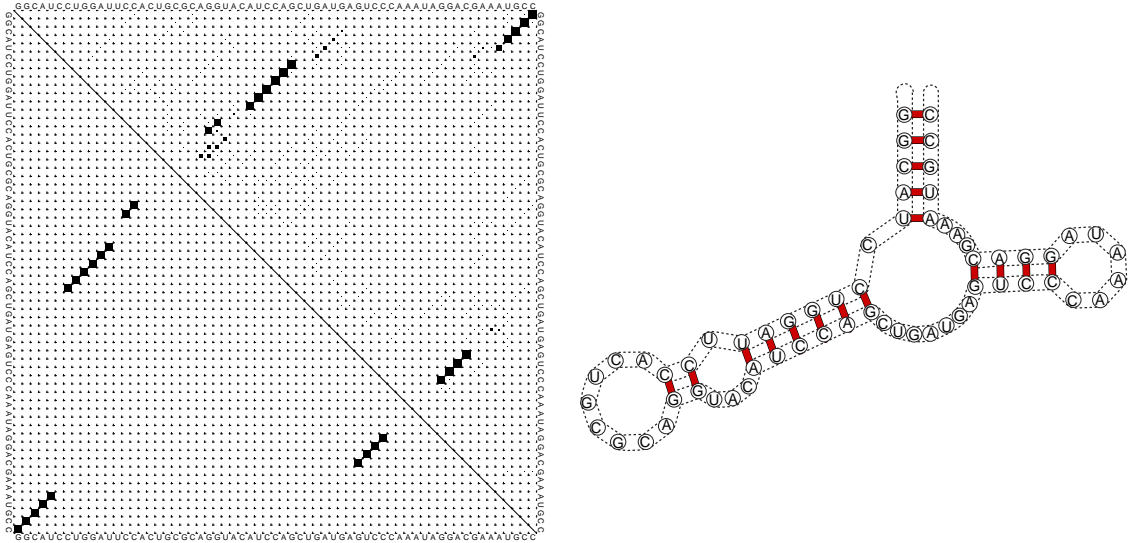


Figure 2.4: Base-pairing probability matrix. The area of the dots in the upper right triangle of the matrix is proportional to probability that a specific base-pair forms in the thermodynamic equilibrium. The lower left triangle shows the pairing pattern in the optimal structure of minimum free energy. Again, a hammerhead RNA is shown as example (conventional drawing on the right hand side). The structure was calculated using the program `RNAfold`

open and close and an RNA molecule does not only fold into a single structure, but forms an *ensemble* of different structures. Following basic principles of thermodynamics, the probability of a given structure \mathcal{S} is proportional to its Boltzmann factor:

$$\text{Prob}(\mathcal{S}) = \frac{\exp(-\Delta G(\mathcal{S})/RT)}{Z} \quad (2.5)$$

where T is the absolute temperature and R the universal gas constant. The normalization factor Z is a particularly important quantity. It is the Boltzmann weighted sum over all possible structures and called *partition function*:

$$Z = \sum_{\mathcal{S}} \exp(-\Delta G(\mathcal{S})/RT) \quad (2.6)$$

As shown by McCaskill (McCaskill, 1990a), the partition function can be calculated using similar recursions and dynamic programming algorithms as used for calculating minimum free energy. For the simple base-pair energy model, the recursion to calculate the partition function can be formulated as follows

$$Z_{i,j} = Z_{i+1,j} + \sum_{\substack{i+1 \leq k \leq j \\ \prod_{ik}=1}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\beta_{ik}/RT) \quad (2.7)$$

Please note the analogy to Equ. (2.2). We can simply replace the minimum by the sum, the sums with multiplications and the energy contribution by its Boltzmann factor. The value of the partition function by itself is usually not of immediate interest. In practice, the most interesting information is the probability of a specific base-pair within the equilibrium ensemble, or more precisely the probability $p_{i,j} = \sum_{(i,j) \in \mathcal{S}} \text{Prob}(\mathcal{S})$ of observing a structure \mathcal{S} that contains the base pair (i, j) . To calculate $p_{i,j}$ we need to know the partition function over all structures forming (i, j) and the total partition function Z :

$$p_{i,j} = \widehat{Z}_{i,j} Z_{i+1,j-1} \exp(-\beta_{i,j}/RT) / Z \quad (2.8)$$

The helper quantity $\widehat{Z}_{i,j}$ is the partition function over all structures *outside* the subsequence $x_{i,j}$. Using similar considerations as for the “forward” recursion, one arrives at

$$\begin{aligned} \widehat{Z}_{i,j} = \widehat{Z}_{i,j+1} &+ \sum_{\substack{1 \leq k < i \\ \Pi_{k,j+1}=1}} \widehat{Z}_{k,j+1} \exp(-\beta_{k,j+1}/RT) Z_{k+1,i-1} \\ &+ \sum_{\substack{j+2 \leq k \leq n \\ \Pi_{k,j+1}=1}} \widehat{Z}_{i,k} \exp(-\beta_{k,j+1}/RT) Z_{j+2,k-1} \end{aligned} \quad (2.9)$$

A common way to summarize the structural properties of an RNA molecule in the thermodynamic ensemble is to calculate the probability matrix of all possible base-pairs. This matrix can be conveniently visualized as “dot-plot”. Fig. 2.4 shows an example of a pairing matrix for a short hammerhead RNA calculated with the program `RNAfold` of the Vienna RNA package (Hofacker *et al.*, 1994a) that implements the partition function calculations described here for the full loop based energy model.

Sequences of Desired Structures

The design of RNA sequences that fold into a desired structure is the so-called inverse RNA folding problem. So far, this has been done for a predefined mfe structure with an inversion of the conventional folding procedure (Subsec. 2.1.3) implemented as `RNAinverse` from the Vienna RNA package (Hofacker *et al.*, 1994a). `INFO-RNA` has been developed recently with a different initialization and a stochastic local search (Busch and Backofen, 2007). Further programs exist, e.g. of Flamm *et al.* (2001). However, all these programs have different focuses than phylogenetic relationships.

2.1.4 Visualisation/Representation of RNA Structures

There are many different ways to visualise secondary structure, which are often called representations in the literature (Moulton *et al.*, 2000; Hofacker and Stadler, 2007). So far we have illustrated RNA secondary structure as conventional secondary structure graphs

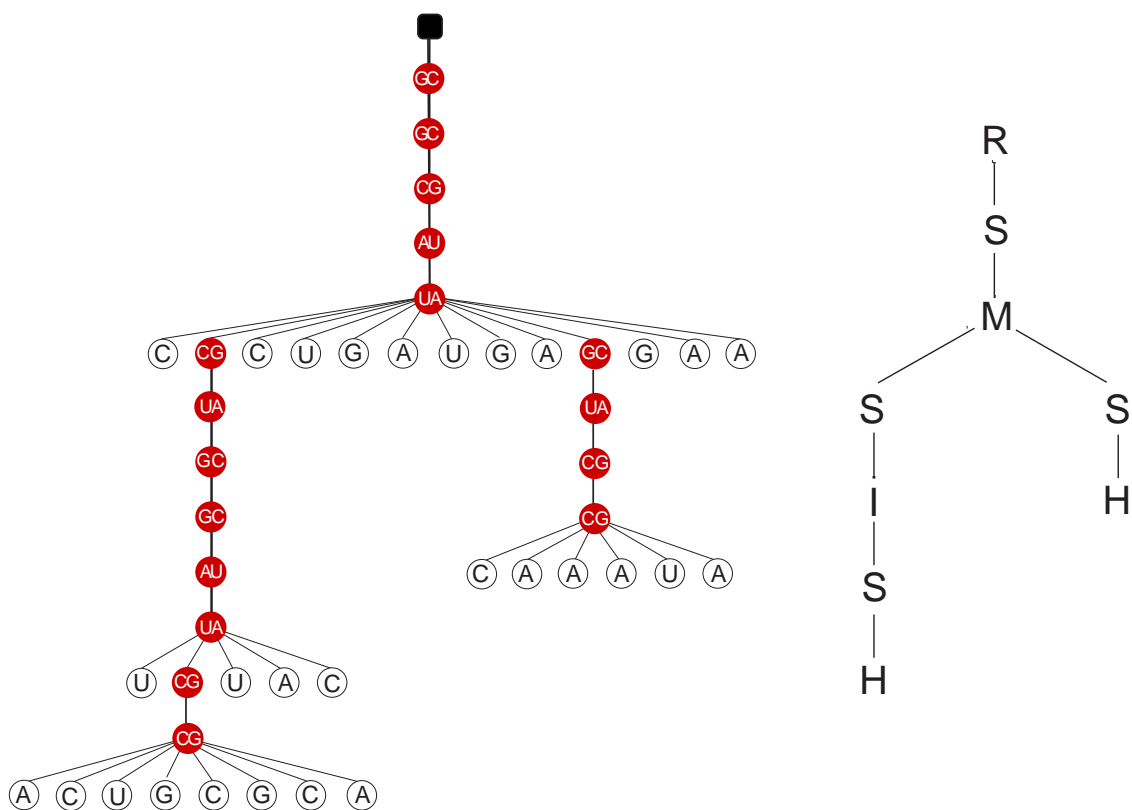


Figure 2.5: Tree representations of RNA secondary structures. Left: the “full tree” representation (Fontana *et al.*, 1993a) Right: Shapiro-style tree (Shapiro and Zhang, 1990a; Shapiro, 1988). R,S,M,I denote root, stem, multi-loop, interior loop and hairpin loop nodes, respectively.

in Fig. 2.4 on the right side, circleplots or dot plots (Fig. 2.2). However, there are other possibilities for visualisations. Here, we summarised the commonly used ones.

Conventional Drawing of structures are used often by biochemist or molecular biologists. In the case of secondary structure, they are called conventional secondary structure graph (Fig. 2.4 at the right side).

Circle Plot can represent secondary structure as well as tertiary structure. The vertices (sites) are arranged in a circle and the edges connect two vertices inside the circle, represent the base pairs all lines crossing the circle. Tertiary structure is easily recognized by crossing of lines in the circular representation in Fig. 2.2 on the left side.

Dot Plot representation, usually a base-pair probability matrix (Fig. 2.4), where the area of the dots in the upper right triangle of the matrix is proportional to the probability that a specific base-pair forms in the thermodynamic equilibrium and the lower left triangle shows the pairing pattern in the optimal structure of minimum free energy.

Dot-Bracket Representation is a compact representation in one line consisting of

parentheses and dots, through replacing unpaired position by a dot and each base pair by a open and closed bracket in the i th and j th positions, respectively .

Linked diagram representation arranged the sequence along the x -axis and the base pairs are drawn as arcs confined to the upper half-plane.

Mountain representation works well for long sequences. Each base pair is represented by a horizontal line over the primary sequence at a height by its position in the sequence (Hogeweg and Hesper, 1984).

Tree Representation encodes RNA secondary structures as as rooted, ordered, labeled trees (Fig. 2.5). In the full tree representation (Fontana *et al.*, 1993a) each internal node represents a base-pair, while leaves represent unpaired bases. The root vertex does not correspond to a physical part of the RNA. Shapiro *et al.* used a more abstract encoding (Shapiro and Zhang, 1990a; Shapiro, 1988) in which internal nodes correspond to the different loop types (stack, interior loop, bulge, multi-loop, hair-pin). Depending on the type of representation the labels have different meaning.

2.1.5 Tertiary Structure

RNA secondary structures usually form before and independently of the tertiary structure and contributes most of the stabilising energy in contrast to protein secondary structure. Thus, as folding intermediates RNA secondary structure is a useful tool for the interpretation and prediction of RNA function. However, the function of the molecule ultimately dependent on the tertiary structure. The folding problem for tertiary structure is enormously complex at least as demanding as in the case of proteins, and it is far from being solved. Recently, recurrent structural RNA motifs and isostericity matrices for tertiary structure, which provide basic knowledge for new algorithm for tertiary structure prediction in future research, have been inferred (Lescoute *et al.*, 2005; Leontis *et al.*, 2006).

2.1.6 Quaternary Structure

Quaternary structure in protein describes the arrangement of multiple folded molecules in a multi-subunit complex. This term is much less used for RNA. However, RNA base pairs may be formed within or between molecules. The function of an RNA molecule often depends on its interaction with other RNAs. Straightforwardly, recent extensions of standard thermodynamic and kinetic folding algorithms was developed to predict structures formed by two RNA molecules upon hybridization (Mückstein *et al.*, 2006; Bernhart *et al.*, 2006). From a thermodynamical viewpoint, quaternary structures are called biologically structure that underlie the folding problem by other biological constraints on the sequence.

2.2 Phylogeny

Darwin realised in *The Origin of Species* that all of life on earth is related, and the pattern of relatedness is shaped like a tree (Fig. 2.6). Around 150 years later, given the molecular sequences observed today, a general goal in phylogeny is to reconstruct history, typically a phylogenetic tree. A further important aim is to understand the processes that govern evolution. The sequences may be either DNA, protein, RNA or other character-based sequences. Although it seems clear that our focus is purely on RNA, there are also other types of biological sequence data like genome, gene or protein structures, and course, these all could be subject to structure-definition research work as well.

The first part of this chapter was about the RNA structural elements which determine the function of a molecule and imply the structuring of the sequence, the second part describes sequence evolution in general.

2.2.1 Sequence Evolution Models

Sequence structuring can be described through the evolution of sequences on a tree in terms of substitutional changes of single positions during some time span t (Fig. 2.8).

Explicit formal sequence evolution models, are most prominently expressed in maximum likelihood approaches (e.g. Felsenstein, 1981; Jukes and Cantor, 1969; Kimura, 1980; Tavaré, 1986). The definition of a model usually requires some assumptions about the evolutionary process. During evolution, mutation and natural selection can only act upon the molecules present in an organism that have no knowledge of their previously history. Such a *lack of memory* is one of the primary assumptions: known from a statistical process the *Markov process*. It means that future evolution is only dependent on its current state and not on its ancestral (previous) state. With a *stationary, time homogeneous and reversible Markovian substitution process* one can compute the probability of nucleotide j given nucleotide i for every positive t , where i is the initial state, which evolves into

Figure 2.6: An evolutionary tree by Charles Darwin (*First notebook on transmutation of species, 1837*). The ancestral species is at position '1'. Extant species are denoted by endpoint and letters and the remaining pendant edges represent extinctions. During drawings Darwin wrote "I think" and in the *The Origin of Species* (Darwin, 1859): "The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species..."

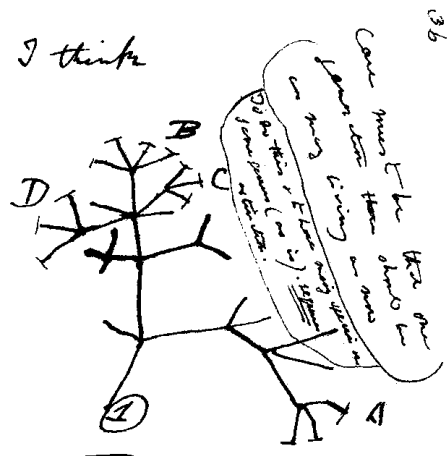
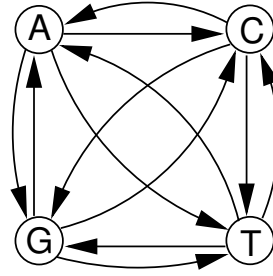


Figure 2.7: Standard models assume, independent evolution of sites and a continuous time Markow process is defined by its instantaneous rate matrix $\mathbf{Q} = (Q_{ij})_{i,j,\dots,|\mathcal{A}|}$, an $|\mathcal{A}| \times |\mathcal{A}|$ matrix, where $|\mathcal{A}|$ is the number of character states. This thesis will be mainly consider RNA evolution with $\mathcal{A} = \{A, C, G, U\}$, hence $|\mathcal{A}| = 4$. For nucleotides, amino acids and codons, the alphabet is 4, 20 and 64, respectively.



another state j after time t . Thereby, the most common models use independence along the sites, as well as other assumptions of Tab. 2.3.

To define such a process one only has to specify a instantaneous rate matrix \mathbf{Q} , in which each entry $q_{ij} > 0$ stands for the rate of change from state i to state j during an infinitesimal period of time, illustrated in Fig. 2.7. The rate at which some change from that state occurs is the sum of all the rates of changes from the state, and this determines the waiting time that particular state before moving to another. Thus the diagonal elements of \mathbf{Q} are given by

$$q_{ii} = - \sum_j q_{ij} \quad (2.10)$$

such that rows sum up to zero.

In molecular sequence data we observe actual characters at some given time and not the rate at which they are evolving. The probability $P_{ij}(t)$ to be in state j after time t given that the initial state was i can be compute using the instantaneous rate matrix \mathbf{Q} , which is related to the probability matrix $P(t)$, since the process is time-homogeneous, via

$$\mathbf{P}(t) = e^{(\mathbf{Q}t)}, \quad (2.11)$$

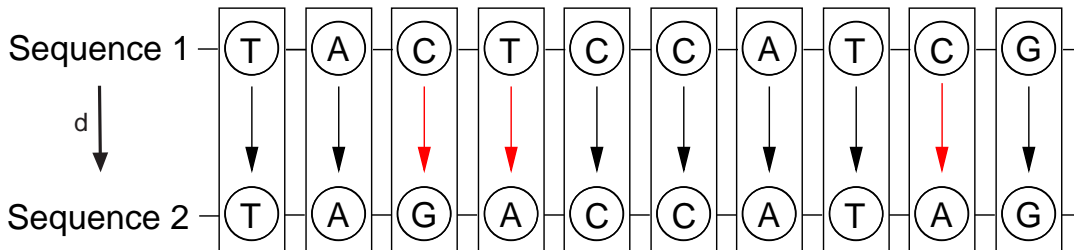


Figure 2.8: The evolution of sequences can be described in terms of substitutional changes of single positions during some timespan t , measured in number of substitution per site d . Thereby, we assume for a independent-sites model the positions evolve independently and according to the same process.

with the identity matrix I , the matrix exponential is defined by the following series (cf. Norris, 1997)

$$\sum_{n=0}^{\infty} \frac{(\mathbf{Q}t)^n}{n!} = I + t\mathbf{Q} + \frac{(t\mathbf{Q})^2}{2!} + \frac{(t\mathbf{Q})^3}{3!} + \dots \quad (2.12)$$

This series is calculated numerically using standard linear algebra techniques. The most popular method in molecular phylogeny uses eigendecomposition, which can be found through diagonalization of \mathbf{Q} ,

$$\mathbf{P}(t) = \mathbf{U} \cdot \text{diag}\{e^{\lambda_1 t}, \dots, e^{\lambda_{|\mathcal{A}|} t}\} \cdot \mathbf{U}^{-1} \quad (2.13)$$

where $\text{diag}\{\dots\}$ denotes a diagonal matrix containing the eigenvalues, and \mathbf{U} is the matrix of the (right) corresponding eigenvectors (Karlin and Taylor, 1975).

The distribution, to which a given initial distribution π^a converges, independent of the starting state after a long time, is the stationary distribution $\pi = (\pi_1, \dots, \pi_{|\mathcal{A}|})$. Reversibility is given by the \mathbf{Q} or \mathbf{P} matrix with:

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \text{or equivalently} \quad (2.14)$$

$$\pi_i q_{ij} = \pi_j q_{ji}. \quad (2.15)$$

for all $i, j \in \{1, \dots, |\mathcal{A}|\}$. This is just a mathematical convenience, but not a biological requirement. It follows from

$$\sum_i \pi_i p_{ij} = \pi_j \quad (2.16)$$

that a stationary distribution π exists and can be found by solving

$$\pi \mathbf{P}(t) = \pi \quad (2.17)$$

for any time t , or equivalent (cf. Norris, 1997) by solving

$$\pi \mathbf{Q} = 0 \quad (2.18)$$

-
- 1 Every site of the sequence evolves independently.
 - 2 The substitution process has no memory of past events (Markov property)
 - 3 The process remains constant through time (homogeneity)
 - 4 The process starts at equilibrium (stationarity)
 - 6 Substitutions occur in continuous time.

Table 2.3: Assumptions of the commonly used nucleotide substitutions models, of which some are collected in Tab. 2.4. The further assumption that the rate of substitution is the same for all nucleotides, can be relaxed including rate heterogeneity.

	A	C	G	T	A	C	G	T	A	C	G	T
	JC69				K80				HKY			
A	*	α	α	α	*	β	α	β	*	$\beta\pi_C$	$\alpha\pi_G$	$\beta\pi_T$
C	α	*	α	α	β	*	β	α	$\beta\pi_A$	*	$\beta\pi_G$	$\alpha\pi_T$
G	α	α	*	α	α	β	*	β	$\alpha\pi_A$	$\beta\pi_C$	*	$\beta\pi_T$
T	α	α	α	*	β	α	β	*	$\beta\pi_A$	$\alpha\pi_C$	$\beta\pi_G$	*
	TN93				F81				GTR			
A	*	$\beta\pi_C$	$\alpha_1\pi_G$	$\beta\pi_T$	*	π_C	π_G	π_T	*	$a\pi_C$	$b\pi_G$	$c\pi_T$
C	$\beta\pi_A$	*	$\beta\pi_G$	$\alpha_2\pi_T$	π_A	*	π_G	π_T	$a\pi_A$	*	$d\pi_G$	$e\pi_T$
G	$\alpha_1\pi_A$	$\beta\pi_C$	*	$\beta\pi_T$	π_A	π_C	*	π_T	$b\pi_A$	$d\pi_C$	*	$f\pi_T$
T	$\beta\pi_A$	$\alpha_2\pi_C$	$\beta\pi_G$	*	π_A	π_C	π_G	*	$c\pi_A$	$e\pi_C$	$f\pi_G$	*

Table 2.4: A collection of different instantaneous rate matrices, based on a set of convenient assumptions of Tab. 2.3. The first model was developed by Jukes and Cantor (1969), and is specified by a single free parameter. Followed by models including more and more parameters. For example, K80: Kimura two parameter model, which distinguishes between transitions and transversions (Kimura, 1980), incorporating more general single substitution models incorporate different base compositions, like: HKY: Hasegawa-Kishino-Yano model (Hasegawa *et al.*, 1985), TN93: Tamura-Nei model (Tamura and Nei, 1993), F81: Felsenstein 81 model (Felsenstein, 1981), GTR: general time reversible model.(Rodriguez *et al.*, 1990). The main diagonal elements (*) are given by Equ. 2.10.

Time is measured in expected numbers of substitutions. We can normalise the instantaneous rate matrix with any factor, since time and rate are confounded and only their product can be inferred without extrinsic information (Felsenstein, 1981). Typically, we scale time such that the expected rate of substitutions per site is

$$-\sum_i \pi_i q_{ii} = 1. \quad (2.19)$$

Finally, we calculate the number of substitution d per site as

$$d = -\sum_{i \in A} \pi_i q_{ii}. \quad (2.20)$$

Due the multiple substitutions we never observe d . We rather observe the number of differences h , that is computed as

$$h = 1 - \sum_{i \in A} \pi_i p_{ii}(t). \quad (2.21)$$

Rate Heterogeneity

The assumption that regions of sequences evolve under the same mutational process and selective constraints does not hold true. It is therefore common to relax this, assuming that many of the complexities of molecular evolution are primarily manifested as a difference in the relative rate for sites changes, whilst maintaining all other aspects of the evolutionary process. In other words, each site has a defined probability of evolving at a given rate, independent of its neighbours. Often, the distribution of relative rates r is assumed to follow a gamma distribution (Uzzell and Corbin, 1971; Yang, 1993),

$$g(r) = \frac{\alpha^\alpha r^{\alpha-1}}{e^{\alpha r} \Gamma(\alpha)} \quad (2.22)$$

whereby α specifies the shape of the distribution with expectation 1 and variance $1/\alpha$. Rates are homogenous, if α tends to the infinity. If α is smaller than one, then strong rate heterogeneity is obtained, α around one indicates a weak rate heterogeneity and a bell-shaped distribution. Further development, breaking the distribution into a pre-specified number of categories makes the model computationally more efficient (Yang, 1994). Van de Peer *et al.* (1993) used empirical pair-wise methods to infer site-specific rates of alignment positions. Based on this idea, Meyer and von Haeseler (2003) introduced a maximum likelihood framework for estimating site-specific rates from pairs of sequences with an iterative extension to compute site-specific rates and the phylogenetic tree simultaneously. The bias introduced in sequence analysis by ignoring heterogeneous rates among sites has been studied in population genetics (cf. Aris-Brosou and Excoffier, 1996) and phylogenetic reconstruction (Yang, 1996), where it is shown that the inclusion of Γ -distribution usually improves the estimation of other evolutionary parameters, including the tree topology. In the sense of reflecting different selective constraints at different sites hidden Markov models (HMM) are used to assign rates of change to each site, according to a Markov process that depends on the rate of change at the neighboring site. These approaches model site dependencies through shared rate parameters, but still assume independent changes at the different sites (e.g. Felsenstein and Churchill, 1996).

2.2.2 Evolutionary Dependence Across Sites

Even though the assumption of independence across sites makes computation feasible it is not biologically realistic. For many years, various authors have attempted to overcome this assumption. The simplest cases include the dependence structure with jointly modelling substitution events. One example is a codon model with three nucleotides or a second example is due to the impact of conserved RNA structure with two nucleotides. Relaxing the assumption of independently evolving sequence fragments is difficult and a challenge. This is necessary for modelling CpGs or including overlapping dependencies in general, e.g for the nearest neighbour energy model (see Sec. 2.1.3).

Codon Substitution Models & a Parameter of Selection

Goldman and Yang (1994) considered dependence of neighbouring sites instead of mononucleotide models within a codon in protein coding regions. When a triplet codon is taken as a unit of evolution, the assumption of independence among sites is naturally relaxed (Fig. 2.10). This leads to a bigger rate matrix and makes computations more demanding, but it is a more realistic model for protein coding region compared with nucleotide models. Similarly, Schöniger and von Haeseler (1994) suggested jointly modelling substitution events in RNA helical regions (see Sec. 2.2.2). Muse and Gaut (1994) also proposed a codon approach, as well as modelling equilibrium frequencies of nucleotides instead of codons. Based on these developments of codon models the selection pressure on the protein coding regions could be studied. The next important step was done by Yang (2000) by simplifying the original model of Goldman and Yang (1994) and introducing a selection parameter ω for the ratio of nonsynonymous to synonymous rate. Then, the substitution rate from codon i to j ($i \neq j$) is given by the following model,

$$q_{ij} = \begin{cases} \pi_j & \text{for synonymous transversion} \\ \kappa\pi_j & \text{for synonymous transition} \\ \omega\pi_j & \text{for nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{for nonsynonymous transition} \\ 0 & i \text{ and } j \text{ differ at more than one site,} \end{cases} \quad (2.23)$$

where κ is the transition-transversion rate ratio, π_j is the equilibrium frequency of codon j and ω is the nonsynonymous-synonymous rate ratio, where $\omega > 1$ suggests that the rate of non-synonymous substitutions are higher and the substitution is beneficial that is more likely to be fixed (positive selection). The case $\omega < 1$, on the other hand, indicates higher rate of synonymous substitutions (negative selection). The approach has become a standard procedure to detect natural selection in protein coding regions. We refer to Anisimova and Kosiol (2009) for a recent review of the development on further mechanistic as well as new empirical and mechanistic substitution models for protein coding sequence evolution.

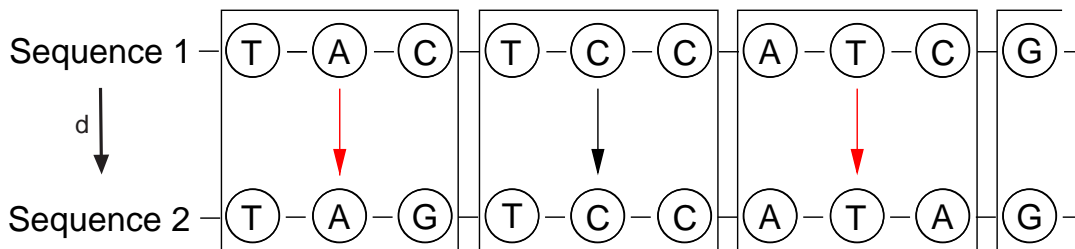


Figure 2.9: Jointly modelled substitution event: for example, a triplet codon is taken as a whole unit of evolution for protein coding regions. Another example is illustrated in Fig. 2.10.

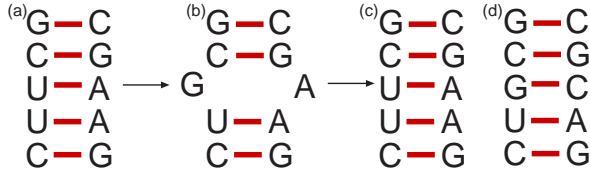


Figure 2.10: Compensatory mutation in a helix: it is quite likely that a substitution at a non-base-pairing doublet (b) will lead to a base-paired doublet within a relatively short time interval. This can be modeled with joint substitution events of two nucleotides (base-pairs).

RNA Base-Pair Substitution Models & Compensatory Substitutions

Nucleotides in a stem region of RNA molecules obviously do not evolve independently of their base-pairing counterparts. Given the frequencies of the admissible base-paired doublets, it is quite likely that a substitution at a non-base-pairing doublet (Fig. 2.10b) will lead to a base-paired doublet within a relatively short time interval, representing a so-called compensatory mutation, see Fig. 2.10. All sites can be classified into two categories: helical regions and loop regions. However, with the assumption of a fixed RNA secondary structure. While the units of loop regions are mononucleotides like in the conventional independent models (Tab. 2.4), the units of helical regions are doublets (base-pairs). In the helical regions, the state space is extended to all possible 16 pair combinations, $i, j \in \mathcal{A} \times \mathcal{A} = \{AA, AC, AG, \dots, GU, UU\}$. Schöniger and von Haeseler (1994) have extended the F81 model (2.4) with the stationary frequencies $\pi_\mu = \{\pi_{AA}, \pi_{AC}, \pi_{AG}, \dots, \pi_{GU}, \pi_{UU}\}$ to the following 16×16 instantaneous rate matrix:

$$q_{ij} = \begin{cases} \pi_j & \text{for } i \neq j \text{ und } h(i, j) = 1 \\ - \sum_{\substack{k \in \mathcal{A} \times \mathcal{A}: \\ k \neq i}} q_{ik} & \text{for } i = j \\ 0 & \text{sonst} \end{cases} \quad (2.24)$$

with $i, j \in \mathcal{A}^2$ and with the Hamming distance h for the usual restriction that only one substitution per unit time is admissible (see Fig. 3.3 in Chap. 3). Finally, the expected number of substitutions can be calculated using

$$d = -\frac{1}{2} \sum_{i=1}^{16} \pi_i q_{ii} \quad (2.25)$$

and the number of observed differences with

$$h = \frac{1}{2} \sum_{i=1}^{16} \sum_{\substack{j=1 \\ H(i,j)=1}}^{16} \pi_i p_{ij}(t) + \sum_{i=1}^{16} \sum_{\substack{j=1 \\ H(i,j)=2}}^{16} \pi_i p_{ij}(t). \quad (2.26)$$

Thus, it is possible to describe the whole process analytically. Several further attempts model dependencies by base-pairing into a Markov model of sequence evolution (Tillier,

1994; Muse, 1995; Rzhetsky, 1995; Tillier and Collins, 1995, 1998). Muse (1995) introduced a new pairing parameter λ to present the effect of forming or destroying a base-pair. Thus, if a stem structure is favored, the relative probability of a change from an unpaired state to a paired state should be greater than the corresponding probability when sites are not paired. On the other hand, instead changes from a paired state to an unpaired state should occur with lower frequencies. Doing that, it is not necessary to choose doublet frequencies. The most general one of these 16×16 matrices is given by

$$q_{ij} = \begin{cases} \pi_t & \text{for transversion, pairing unchanged} \\ \kappa\pi_t & \text{for transition, pairing unchanged} \\ \pi_t\lambda & \text{for transversion, unpaired} \rightarrow \text{paired} \\ \kappa\pi_t\lambda & \text{for transition, unpaired} \rightarrow \text{paired} \\ \pi_t\frac{1}{\lambda} & \text{for transversion, paired} \rightarrow \text{unpaired} \\ \kappa\pi_t\frac{1}{\lambda} & \text{for transition, paired} \rightarrow \text{unpaired} \\ 0 & i \text{ and } j \text{ differ at more than one site,} \end{cases} \quad (2.27)$$

where, π_t is the frequency of the nucleotides that differ in i and j .

Other models consider only the six possible pairs in a 6×6 instantaneous rate matrix (Tillier, 1994), or in a 7×7 matrix with one state for all mismatch pairs (Tillier and Collins, 1998). Usually, the instantaneous substitution rate of two nucleotides simultaneously in one doublet are assumed to be zero. However, there are also models which allow doublet substitutions. In a comparative study Savill *et al.* (2000) have shown that such models performs best using statistical tests (e.g. likelihood-ratio test in Sec. 2.3.3), which is not a surprising result. Also models that permit a nonzero rate of double substitutions performed better than those that assume the usual restriction that only one substitution per unit time is admissible. Furthermore, the idea of empirical rate matrix was adopted for RNA sequences from a large number of sequences by Smith *et al.* (2004).

Since Kimura (1980), the mechanism of compensatory mutations is accepted as an explanation of the conserved structure. Further studies about the rate of compensatory mutations were done by (Stephan, 1996; Innan and Stephan, 2001). Recently, Smit *et al.* (2007) have shown that different secondary structure categories evolve at different rates. Jointly modelling substitution event with extending the state space to $|\mathcal{A}|^k$ to independently evolving sequence fragments up to length $k = 3$ is given by Siepel and Haussler (2004). A general description of the Markov process for any fragment length k is described by von Haeseler and Schöniger (1998). Further relaxed assumptions of independently evolving sequence fragments is more complicated and a big challenge towards more realistical models.

Context-dependent Substitutions

The assumption of independence among sites is naturally relaxed through independent units of evolution, e.g. codons or base-pairs. However, to relax the assumptions of independently evolving sequence fragments, e.g. overlapping dependencies is more difficult. There is clear evidence that such phenomena exist and it seems worthwhile to incorporate these into probabilistic models.

For instance, the stacking interactions between residue pairs that are adjacent in an RNA helix (see Sec. 2.1.3) should be taken into account. Furthermore, a number of empirical studies have found indications that the assumptions of independent evolution of sites is too restrictive, including the dependency of the identity of flanking sites (e.g. Bulmer, 1986; Morton, 1995).

Pioneering work was done by Jensen and Pedersen (2000) using a Markov model of nucleotide sequences evolution in which the instantaneous substitution rates at a site are allowed to depend on the states of a neighbouring site at the instant of the substitution. However, they could only consider pairs of sequences. Their model consists of a first component that depends on the type of change, while the second component models the CpG-deamination process. This is extended towards an approach including arbitrary reversible codon substitution models with more CpG-deamination flexibility by Christensen *et al.* (2005). In these approaches, inference is obtained using MCMC or EM-based pseudo-likelihood estimation and phylogenetic inference is still a problem (Baele, 2009).

Therefore, context-reducing models were developed, where context-dependent substitutions can be handled in an approximate way with some simple extensions, which themselves are direct extensions of Felsenstein's original framework. These models require certain limitations on the independence between sites, but they allow for exact inference without too much additional cost in computation, e.g. the approach of Siepel and Haussler (2004). However, recently, Bérard *et al.* (2008) have presented analytical results of a special sub-case of a Tamura+CpG model, which was already suggested informally by Duret and Galtier (2000), later formally by Arndt *et al.* (2003). They have shown that these models are solvable.

Global context dependency models first correlated to considering protein structure were developed in Jeffrey Thorne's group using a Bayesian MCMC. Robinson *et al.* (2003) defined an instantaneous rate matrix that specifies rates of change from each possible sequence to each other possible sequence with the usual restriction that no more than one position is allowed to change in a particular instant. Yu and Thorne (2006) modify the approach of Robinson *et al.* (2003) to formulate and explore a possibility where the relative rate of sequence evolution is affected by approximate free energy of RNA secondary structure. They define an instantaneous rate matrix that specifies rates of change from each possible sequence to each other possible sequence. The rate matrix entries are

$$q_{ij} = \begin{cases} \mu\pi_t\kappa e^{s(E(i)-E(j))} & \text{for transition} \\ \mu\pi_t e^{s(E(i)-E(j))} & \text{for transversion} \\ 0 & i \text{ and } j \text{ differ at more than one site,} \end{cases} \quad (2.28)$$

where π_t is the frequency of the nucleotides that differ in i and j , κ is the transition-transversion rate ratio, μ is a rate-scaling factor, and $E(i)$ and $E(j)$ denote the approximate free energy of the whole sequences, which differ in i and j . When the parameter s is zero, the structure does not affect the substitution rates, and the model reduces to the HKY independent model. With this dependence structure the rate matrix is $4^l \times 4^l$ with sequence length l , and it is not feasible to exponentiate the instantaneous rate matrix unless l is extremely small. To overcome this high dimensionality, they use a sequence path approach (Jensen and Pedersen, 2000; Pedersen and Jensen, 2001).

Another approach to modelling is to compare context-dependent rates across clades. Hwang and Green (2004) estimate separate context-dependent rate matrices, where the context is described by two adjacent ancestral nucleotides. However, they do not use well-known evolutionary models. In this context, Baele *et al.* (2008) study the influence of the composition of the neighboring bases on the substitution probabilities for a given site, using the well-known general time-reversible model. It is known that too many additional parameters to model site dependence will only add noise and imply risk of overfitting the data. Careful model-building strategies in combination with additional parameters are necessary. These models call for extensive studies of their mathematical behaviour, e.g. the convergence to equilibrium, and of their ability to reproduce statistical properties observed in biological sequences. Therefore, simulations seems to be a very appropriate tool to find out which parameters are needed to model site dependence.

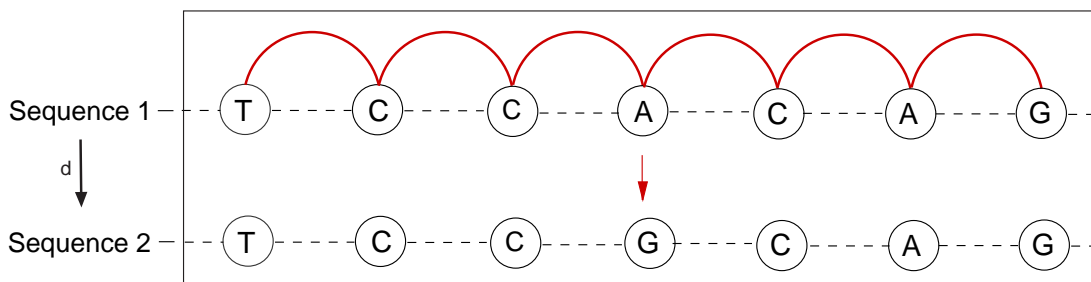


Figure 2.11: Modelling global context dependency specifies rates of change from each possible sequence to each other possible sequence with the usual restriction that no more than one position is allowed to change in a particular instant. This dependency structure leads to high dimensionality rate matrices. Different approaches have been developed to overcome this problem. Please refer to the text for details.

2.2.3 Phylogenetic Tree, Inference & Alignments

According to Darwin's theory (Fig. 2.6), we assume that n sequences S_n are related according to a rooted tree T where the leaves represent the sequences in the alignment and the branch length of T reflects the amount of evolution. Especially, in this thesis we are interested in nucleotide patterns of RNAs along a phylogenetic tree $T = (V, E)$ with node set V and branch set $E \in V \times V$ (Semple and Steel, 2003). The node set V contains the taxon set S that forms the leaf set.

Definition 2.2.1. A phylogenetic tree T is a connected, undirected, acyclic graph with leaves labelled bijectively by the taxon set S that forms the leaf set.

- (i) An unrooted phylogenetic tree T has no vertices of degree two.
- (ii) A rooted phylogenetic tree has an internal vertex, which may have degree two and a so-called root.
- (iii) A star tree is a phylogenetic tree with one internal vertex, which may have degree of the cardinality of the taxon set S .

The *tree-length* Λ_T is the sum of the branch lengths.

$$\Lambda_T = \sum_{e \in E} \lambda(e) \tag{2.29}$$

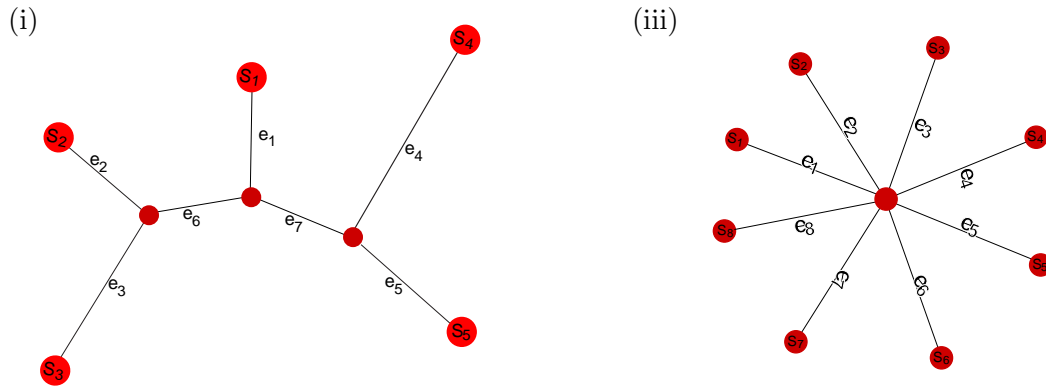


Figure 2.12: Phylogenetic Trees which shows the relatedness of n sequences (S_n). The e_i symbolize the branch length. The distance between any two sequences can be computed by summing up the lengths of the branches which connect them. (i) Unrooted tree of five sequences (S_1 to S_5): this tree does not contain a node (plain red node) which corresponds to the ancestor of the five sequences. (iii) Star tree (of eight sequences, S_1 to S_8), that is all external nodes of the tree have one common ancestor (plain red point in the middle). See Fig. 2.13 for an example of an rooted tree of five sequences (S_1 to S_5).

where $\lambda(e) > 0$ represents the *length* of a branch $e \in E$. In commonly used methods the branch length is measured in numbers of substitutions per site. The distance d between two vertices is called genetic distance.

Thus, sequences are related or homologous if they share one common ancestor. Since, we have only information about contemporary sequences, the evolutionary history needs to be reconstructed. Normally, the commonly used assumption of Tab. 2.3 are used. Of course, not only the assumptions of independence between the sites very simplified. For example, the assumption that the evolutionary rate at a given homologous position varies across time, so-called *heterotachy*, has been investigated, and also time-heterogenous models have been developed recently (e.g Lopez *et al.*, 2002; Philippe *et al.*, 2003; Lockhart *et al.*, 2006). So far, there have been two different approaches. One called the *covarion approach*, where sites switch from variable to invariable states and vice versa and the other known as the *mixtures of branch length model*, which suggests that alignment patterns arise from one of several sets of branch lengths under a given phylogeny. Fig. 2.13 illustrates how variable sites can evolve in a lineage-specific manner due to changes in evolutionary constraints.

For phylogenetic inference, there are currently four main methods: methods based on the parsimonious principle, i.e. maximum parsimony (Fitch, 1971), statistical methods such as maximum likelihood (Felsenstein, 1981) or Bayesian inference (Rannala and Yang, 1996), as well as distance-based methods like neighbour-joining (Saitou and Nei, 1987). In this thesis, we are use maximum likelihood approach as well as neighbour-joining with own distance methods to reconstruct the phylogeny of an alignment. For further tree reconstruction methods and descriptions in detail we refer for example to Felsenstein (2004).

We will ignore throughout the whole thesis genomic rearrangements such as recombinations, inversions or transpositions, which complicate sequence comparison by destroying the original order by the sequences. Instead, we only consider sequence changes that occur during substitutions, however taking the influence of site-specific interactions into account.

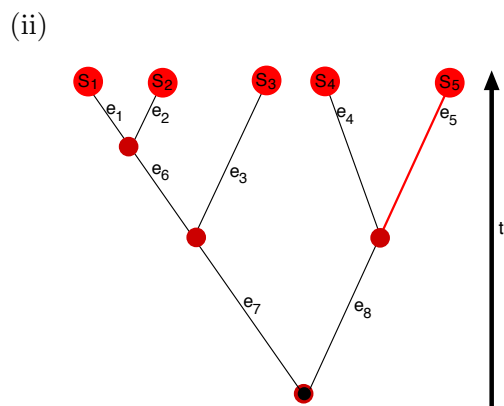


Figure 2.13: (ii) Rooted tree of five sequences (S_1 to S_5), where the internal black node is labeled as a root. Red and black branches illustrates a time-heterogenous process: For example due to selection or drift, some properties of the evolutionary process can change with time. The difference in the process is illustrated in the figure with different colours of the branches. Sequences along this tree evolve in a lineage-specific manner due to changes in evolutionary constraint.

SIMULATION / OBSERVED DATA

S1: UAUUCGGCACGUACGAGGAAUGCUGGUAAG
 S2: UAUUCGGCACGUACGAGGAAUGCUGGUAAG
 S3: UUUACCGGUAUG
 S4: UUUACCGCACGUACGAGGAAUCCUGUUUG
 S5: UUUACCGCACGUACGAGGAAUCCUGUUUG

ESTIMATION / MSA

S1: UAUUCGGCACGUACGAGGAAUGCUG -UAAG
 S2: UAUUCGGCACGUACGAGGAAUGCUG -UAAG
 S3: UUUAC_____CUGGUAUG
 S4: UUUACCGCACGUACGAGGAAUCCUGUUUG
 S5: UUUACCGCACGUACGAGGAA - CCUCGUUUG

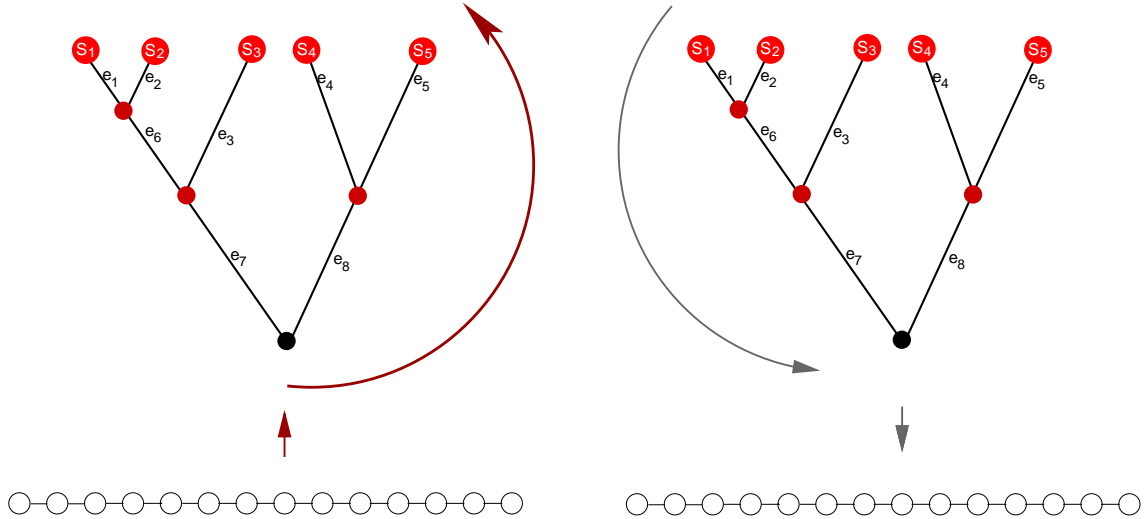


Figure 2.14: The homology relations of the individual bases that result from sequence evolution via point mutations, insertions and deletions can be displayed as a multiple sequence alignment (top right corner). Alignments are on the basis of many molecular analysis. The observable of this process are just the unaligned sequences (top left corner).

In addition, we will start to briefly consider insertion and deletions, which leads to such data sets of unaligned biological sequences of different sequence length (see Fig. 2.14). Generally, in order to apply phylogenetic inference methods, we have to know which bases of the sequences are homologous, sharing a common ancestor, and which are only present in a subset of the sequences due to insertion or deletion events. This can be represented in a form of sequence alignments. An alignment is a simply a data matrix where each row corresponds to one of the sequences and where those bases which are assumed to be homologous to each other stand in the same column (see Fig. 2.14). Usually the underscore "-" is used as a character inserted to make all of the sequence to the same length. Global sequence alignments cover the entire sequences, while the relatedness of some parts of sequences is so-called local alignments. An alignment with only two sequences is called pairwise sequence alignment. Alignments are the basis of molecular analysis such as structure prediction as well as phylogenetic inference. To discuss the different alignment methods and the state of the art is out of the scope of this thesis. We refer to the following literature. To process these methods would be one of the greatest challenges in bioinformatics and is of high importance for structure prediction as well.

2.3 The Two Fields Crossing

The last two sections have introduced the state of the art of RNA structure, especially from the thermodynamic viewpoint of RNA secondary structure, and phylogeny, particularly from the sequence evolution viewpoint. Frameworks which combine these would be of value for themselves because they would contribute to our understanding of the intertwined relationship between RNA structure and the evolutionary process. Although, work was already done in development of RNA base-pair substitution models in terms of compensatory mutations under models with non-overlapping tuples (page 26), it has not been used in practice so far. One reason could be that a priori knowledge about the structure is necessary and methods to construct the necessary structure are insufficient. The other reason could be that the improvements in phylogenetic inference with these models are not significant enough in comparison to independent models with rate heterogeneity. Schöniger and von Haeseler (1995a) have shown with an analysis of the efficiency of three reconstruction methods that the inferred tree is not very much affected in the presence of these kind of correlations. This could be different with other complex dependency models. However, the direct combination, especially a thermodynamic nearest neighbor model with a phylogenetic approach is considerable harder. Standard methods can no longer be used for likelihood computation and parameter estimation if Markov random fields arise. Therefore, a lot of work is in progress to cover the technical skills and these methods are still not practical on a wide-scale level.

2.3.1 RNA Structure as An Application to The Landscape Concept

RNA structure can be defined at several levels, see Tab. 2.2. RNA secondary structure is the best compromise between theoretical tractability and empirical accessibility on a large scale (Fontana, 2002). The application of the landscape concept promise to understand the molecular basis of structure formation, optimization, adaptation, and evolution, also with including thermodynamics viewpoints. During the last years, a lot of work has been done especially by the Vienna Group (cf. Fedoroff and Fontana, 2002; Fontana *et al.*, 1993b; Schuster and Stadler, 2007). So far, *two classes of landscapes* exist: *conformational landscapes* mapping RNA conformations into free energies of formation and *sequence-structure mapping* assigning minimum free energy structures to sequences.

However, the full power of the RNA model unfolds when sequence-structure maps and conformational landscape are merged into a more advanced mapping that signs a whole spectrum of conformations to the individual sequence (Schuster and Stadler, 2007). At present, the analysis of relations between sequences and structures is facilitated by means of three formal discrete spaces: *the sequence space* being the space of all sequences of chain length n , the *shape space* meant here as the space of all secondary structures that can be formed by sequences of chain length n and as a subset a conformation space containing all structures that can be formed by one particular sequence of chain length n .

Even more general is the work including the kinetic effects in the landscapes, e.g. Schuster and Stadler (2007). This introduces a kinetic process over time into the landscape. However, to the best of our knowledge taking phylogenetic time in number of substitutions per site, as well as the relationships, typically in a phylogenetic tree, in the landscape into account is still missing. From the phylogenetic model perspective there could be various reasons for the lack of consideration of phenotype information into an evolution model and vice versa. Probably, one reason could be the computational complexity, but furthermore the shortcomings in communication between the different communities.

2.3.2 Structure Prediction From a Set of Sequences: Consensus Structure

Using the thermodynamic energy model to predict the RNA structure of a single sequence is more accurate than protein predictions, but not accurate enough to satisfy. Moreover, not all thermodynamics parameters are known with appropriate accuracy. Therefore, further methods have been developed to improve the predictive power. For example, one approach to do that is to take a set of homologous sequences into account. Most functional RNA molecules have characteristic secondary structures that are highly conserved in evolution. For example, by compensatory mutations (Fig. 2.10) related RNAs can differ in sequence while having the same structure. If we assume no heterogeneous process in time (see page 30), sequences should fold into a common secondary structure. Different methods exist to calculate the so-called *consensus structure*. One approach aligns the sequences to a secondary structure, that is simultaneously performing sequence alignment and structure prediction. Sankoff (1985) has proposed an algorithm to do that and different simplifications exist and are implemented (Mathews and Turner, 2002; Hull-Havgaard *et al.*, 2005; Hofacker *et al.*, 2004; Holmes, 2005; Dalli *et al.*, 2006; Will *et al.*, 2007). `SimulFold` (Meyer and Miklos, 2007) approaches simultaneously inferring RNA structures and alignments, as well as inferring a phylogenetic tree using a Bayesian MCMC framework.

Another approach is to get first an alignment. After aligning it should be possible to construct a structure that can be formed simultaneously by all (or almost all) input sequences.

It is common to deduce the *consensus structure* by detecting patterns of co-evolution via measuring covariation between two positions (alignment columns). The idea is to find sites where the degree of co-occurring mutations is higher than by random expectation. Thus, the null hypothesis H_0 is that the joint probability $\mathbb{P}(X_i, X_j)$ of a observed base pair at the alignment sites i and j equals the observed pairs under independence that is $\mathbb{P}(X_i) \cdot \mathbb{P}(X_j)$. Correspondingly the alternative hypothesis H_1 is given by

$$H_1 : \mathbb{P}(X_i, X_j) \neq \mathbb{P}(X_i) \cdot \mathbb{P}(X_j) \quad \text{with } X_i, X_j \in \mathcal{A} = \{A, C, G, U\}. \quad (2.30)$$

In practice, the probabilities are calculated from the observed frequencies of nucleotide pairs $f(X_i, X_j)$ and nucleotides $f(X_i)$ and $f(X_j)$ at the sites i and j .

Various measures exist to measure the covariation between two positions in an alignment, ranking from simple mutual information to more advanced covariation measures (Lindgreen *et al.*, 2006). Moreover, we classify these approaches between using only sequence data and additionally incorporate phylogenetic information.

Comparative methods are in principle not limited to secondary structure. For example, pseudoknots and other site-specific interactions of a tertiary structure can be detected (e.g. Gutell *et al.*, 1992a; Tabaska *et al.*, 1998; Dowell and Eddy, 2004). Furthermore, it is possible to consider the coevolution of sites intramolecular (within a single molecule) or intermolecular by taking each site in a distinct data set.

Mutual Information Measure

One wants to find pairs of columns that show a higher degree of covariation than expected by chance. The classical measure for this is the Mutual Information measure (MI) (Shannon, 1948) and was introduced for RNA predictions by Chiu and Kolodziejczak (1991); Gutell *et al.* (1992b); R. Durbin and Mitchison (1998), given by the following expression:

$$MI_{i,j} = \sum_{(X_i, X_j)} f(X_i, X_j) \log_2 \frac{f(X_i, X_j)}{f(X_i) \cdot f(X_j)} \quad \text{with } X_i, X_j \in \mathcal{A}, \quad (2.31)$$

where the sum is over all possible pairs (X_i, X_j) . $f(X_i, X_j)$ are the observed frequencies of nucleotide pairs and $f(X_i)$ and $f(X_j)$ are the nucleotides at the sites i and j . MI or variants hereof are used in a number of programs, e.g. **KNetFold** (Bindewald and Shapiro, 2006), **COVE** (Eddy and Durbin, 1994), **ILM** (Ruan *et al.*, 2004), **MatrixPlot** (Gorodkin *et al.*, 1999) and **ConStruct** (Lück *et al.*, 1996; Wilm *et al.*, 2008).

χ^2 Statistics

Klingler and Brutlag (1993) have used several statistical measures, including χ^2 Monte Carlo simulations and an information measure for n sequences. For each pair of positions they construct a 4×4 contingency table including the numbers of each sequence pair seen in the two positions in the data set. A χ^2 -statistic is used to test for non-independence.

$$X^2(X_i, X_j) = n \sum_{X_i, X_j} \frac{\{f(X_i, X_j) - f(X_i) \cdot f(X_j)\}^2}{f(X_i) \cdot f(X_j)} \quad \text{with } X_i, X_j \in \mathcal{A}, \quad (2.32)$$

with nine degrees of freedom. Then, the null hypothesis is rejected if: $X^2 > \chi_{\alpha,9}^2$ with significance α .

Advanced Covariance Measures

Standard MI is widely used. However, Lindgreen *et al.* (2006) showed that this is not the best measures. For example **RNAalifold** (Hofacker *et al.*, 2002a) or **MSARI** (Coventry *et al.*, 2004) have implemented an advanced covariation measure. More advanced measures were developed and combinations thereof:

- MI summing only Watson-Crick and wobble base pairs (Gorodkin *et al.*, 1999)
- Normalized MI (Martin *et al.*, 2005)
- Covariation measure used in `RNAalifold`
- MI using gap penalties
- MI including stacking

The best measured tested by Lindgreen *et al.* (2006) is the `RNAalifold` covariation measure modified to include stacking (see Hofacker *et al.*, 2002a, for details).

Covariance and Thermodynamics

Some methods calculate the *consensus structure* from an alignment with thermodynamic methods, averaging the energy contribution over all sequences and incorporating covariance methods, e.g. `RNAalifold` (Hofacker *et al.*, 2002a) and `ConStruct` (Lück *et al.*, 1996; Wilm *et al.*, 2008). An open question is the optimal weighting between thermodynamics and covariance.

Sometimes these methods are coined phylogenetic methods (e.g Steger, 2003; Hofacker and Stadler, 2007), also if they take no phylogeny into account. From our viewpoint, we consider these to be pseudo-phylogenetic methods because the phylogenetic relationship between the sequences is not taken into account. However, a small number of current methods have started to take phylogenetic into account, but without considering thermodynamics. Instead, these methods are still based only on comparative mutations and probabilistic approach.

Incorporating Phylogeny

Measures including phylogenetic information for structure predictions have not been used in practice so as widely as the above described covariance methods, although it has been suggested that they are the most powerful (Akmaev *et al.*, 1999, 2000). For example, Lapedes *et al.* (cf. 1999) have shown in respect to structure predictions that sequences related by a phylogenetic tree do not constitute an independent sample. Thus, unless the sequences are related by a star tree, covariance methods without phylogeny are too generous in suggesting associations. Finally, Dutheil *et al.* (2005) have presented a phylogenetic method for coevolving sites, which takes the uncertainty over ancestral states and multiple substitution events into account. Using a substitution model to map the substitutions that occurred at each site onto the branches of the underlying phylogenetic tree, a substitution vector containing the posterior estimates of the number of substitutions in each branch is computed. For example, $V_i = (v_{i,1}, \dots, v_{i,b}, \dots, v_{i,m})$ be a vector of dimension m , the number of branches in the tree, where $v_{i,b}$, is the posterior estimate of the number of substitutions that occurred on branch b for site i . Then, a Pearson correlation coefficient between two corresponding substitution vectors V_i and V_j in comparison to the expectation under the null hypothesis of independence measures the amount of coevolution with

$$P_{i,j} = \frac{\text{cov}(V_i, V_j)}{\text{sd}(V_i) \times \text{sd}(V_j)}, \quad (2.33)$$

where $\text{cov}(V_i, V_j)$ is the sample covariance of V_i and V_j and $\text{sd}(V_i)$ and $\text{sd}(V_j)$ their standard deviations. To evaluate the null distribution of $P_{i,j}$ a parametric bootstrap approach was used, where 100000 simulations were performed under a HKY model with rate heterogeneity.

Stochastic Grammars

In the 1950s, Noam Chomsky began developing a formalisation of grammars to construct sentences in languages (Chomsky, 1959). His theory of generative grammar and a hierarchy of increasingly complex grammars, which can create complex sentences, has had not only influence on linguistics. For example, it has influenced the philosophy of language and mind, and computer scientists adopted it to describe programming languages. Moreover, since 1994 grammars are used to describe RNA secondary structure (Eddy and Durbin, 1994; Sakakibara *et al.*, 1994). Thus, different applications were developed, e.g. alignment algorithms, ncRNA gene finder like **QRNA** (Rivas and Eddy, 2001) and **Evo1Fold** (Pedersen *et al.*, 2006), or programs that predict pseudo knots (Rivas and Eddy, 1999). **Pfold** (Knudsen and Hein, 2003, 1999) computes the consensus structure from an alignment using context free grammar, a 16×16 sequence evolution model and an inferred tree. Andersen *et al.* (2007) have implemented an extended version of **Pfold** that identifies base pairs that have high probabilities of being conserved and of being energetically favorable. However, they have not developed a combined probabilistic model for evolution and folding.

CONTRAFold (Do *et al.*, 2006) was developed as a generalisation of stochastic context free grammar (SCFG) using conditional log-linear models (CLLMs), which can incorporate sophisticated scoring schemes of thermodynamic methods. Moreover, it can replace thermodynamic scoring schemes with training scores. Thus, **CONTRAFold** scores for all possible base pairs exist. As a result, it can predict structures that are definitely not the mfe structure.

Evolution of noncoding RNAs, Structural Stability & Structural Conservation

Little is known about the evolution of RNA at the structural level. However, accurate measures for structural conservation are essential, for instance to predict ncRNAs. The most up-to-date methods of RNA structure conservation can be subdivided on different levels: methods with comparisons of predicted minimum free energy, comparison of single structures, as well as ensembles of structures representing the whole folding space and some specialised methods. The different strategies were tested in a recent study (Gruber

et al., 2008). They showed that simple base-pair distance metric and the folding energy based, called structure conservation index (SCI), are by far the most accurate to measure structure conservation. The structure conservation index is calculated by

$$SCI = \frac{E_{cons}}{\bar{E}_{single}}, \quad (2.34)$$

where E_{cons} is the consensus mfe resulting of the consensus structure prediction programs described above and \bar{E}_{single} is the average mfe of the single sequences, such that the index is normalized and independent of nucleotide composition and length of the alignment.

Another important measurement for RNA is the stability of an RNA-structure. Typically, the normalized z-scores of the mfe $z(i)$ of a sequence i is computed (Washietl and Hofacker, 2004; Clote *et al.*, 2005), with

$$z(i) = (mfe(i) - \mu(i))/\sigma(i), \quad (2.35)$$

where the mean mfe $\mu(i)$ and the standard deviation $\sigma(i)$ of mfe are calculated from shuffled sequences i . A z-score indicates the deviation of the minimum free energy of the RNA sequences from the mean folding energies of a random data set. Negative z-scores indicates that the minimum free energy of the native sequences is more stable than those of the randomized sequences, whereas positive z-scores indicate more unstable structures than expected by chance. Of course, also other folding measures exist. Furthermore, a comparison of RNA folding measures is given in Freyhult *et al.* (2005).

In comparative genome analysis strategies to detect and annotate noncoding RNAs have been broadly developed based on such measurements, (cf. Griffiths-Jones, 2007; A. F. Bompfünewerer Consortium *et al.*, 2007). Even so, the classification of the resulting candidate sequences is still difficult. In the Rfam database (Griffiths-Jones *et al.*, 2005), RNA families are defined by homology. RNA classes are defined via functional and/or structural similarities. This is naturally correlated with the questions to the correct assignment of homologous characters, which is attached to problems as taxon coverage, sequence conservation difference between different regions as well as variation in substitution rates. It is still unclear, how important dependency models are to answer these questions.

2.3.3 Appropriateness of the Description of Sequence Evolution Models

In phylogenetic literature, several test statistics were developed by examining parameters estimated from the data or by comparing how well models explain sequence evolution, e.g. by Goldman (1993), who employed a test statistic suggested by Cox (1962) to check the adequacy of stochastic models. This requires the formation of two hypotheses, the null, H_0 , and the alternative hypothesis, H_1 , represented by models M_0 and M_1 with different constraints. Then, the log-likelihoods $S_0 = \log l(\hat{T}_{M_0} | M_0, \mathcal{D})$ and $S_1 = \log l(\hat{T}_{M_1} | M_1, \mathcal{D})$ for a sequence alignment \mathcal{D} is computed with the maximum-likelihood estimated tree \hat{T}_M

as: $l(\hat{T}_M|M, \mathcal{D}) = \max_{T \in \tau} \{l(T|M, \mathcal{D})\}$ with the likelihood $l(T|M, \mathcal{D})$ of a tree T and the space τ of all possible trees. The model with the higher likelihood fits the data better. The likelihood ratio test compares directly the likelihoods of the null and the alternative hypothesis and is computed as $\delta_{M_1-M_0} = (S_1 - S_0)$. If the models are not nested, the distribution of this statistic is not known and it is estimated by Monte Carlo simulations. This approach is frequently used (Whelan *et al.*, 2001; Yang *et al.*, 2000) and many complex biological problems about the evolutionary process have been investigated using carefully constructing nested hypotheses.

Generally, methods are abound for no site-specific interactions. Some methods are developed which determine whether the evolution of sequences on a phylogenetic tree is better described by a joint evolutionary model (Sec.2.2.2) rather than a model with independent sites (Goldman, 1993; Navidi *et al.*, 1991).

In the case of RNA secondary structure and under the assumption that structure is known, e.g. Schöniger and von Haeseler (1999) tested a correlation model as an alternative to independence models to estimate the percentage of stem positions that do not appear to

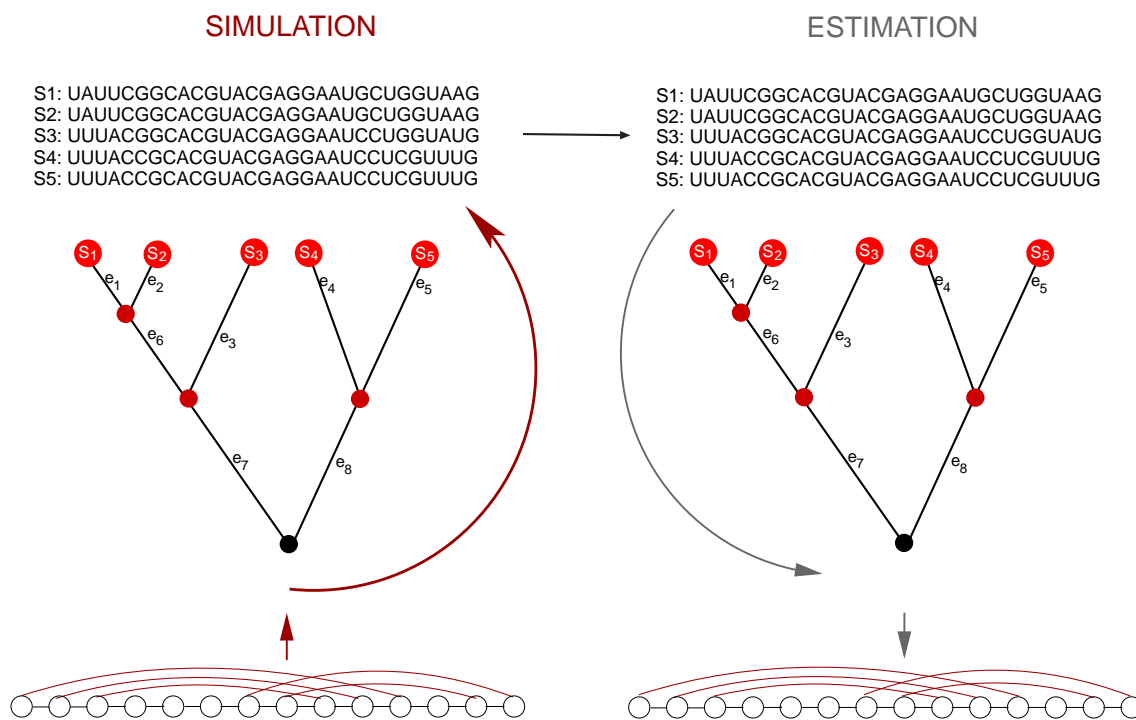


Figure 2.15: Appropriateness of the description. Whenever we analyse a set of homologous sequences, they are related by a phylogenetic tree. That is, we have to take into account the evolutionary history. More complex simulation models and frameworks have to be developed. For instance, since the distribution of statistics it is not known, it can be estimated by Monte Carlo simulations. Moreover, simulating sequence evolution, which take the structure into account, is helpful to investigate the performance of tree-building methods as well as structure prediction methods.

be correlated. Based on the outcome of this test secondary structure prediction could be improved. However, as discussed above, so far the simple RNA base-pair models have not succeeded in the RNA community. In this stage, it is still unclear how *a priori* knowledge about structure can be declared as sufficient and how complex the model has to be to take the shortcoming of current models into account.

Additional work will be required to sort out which the most important context effects are, and whether simpler parameterizations or context-dependent rate matrices can be justified. More complex simulation models and frameworks have to be developed before we can go further into this direction of research. Furthermore, new challenges like finding lineage specific structures or detecting mis-aligned sequences can be considered via complex simulations and several test statistics. Therefore, we start with a simulation framework for arbitrary complex models in the first research chapter of this thesis.

Chapter 3



Romy Schneider (1938-82)

SISSI's Simulacrum

Romy Schneider wrote in her diary in 1949: "I absolutely must be an actress. I must!" In the end, the SISSI-Trilogy (1955) sticks to her like semolina pudding....

TO THE OCTOPUS

This chapter is devoted a simulacrum, since structure is only a simulacrum of the objective in reality. We developed a simulation program called SISSI (SIMulating Site-Specific Interactions) to understand the intertwined relationship between structure and the substitution process. While progress has been made in devising new models for inference with dependencies among sites, there is a lack of simulation tools which would allow the assessment of this progress. The generation of synthetic data is a nontrivial task, because one needs to generate simulated data with the same underlying parameters and statistics as the real data on which the tool will eventually be used. Furthermore, *stochastic simulations* typically consider the random formation and decay of single molecules and multi-component complexes *explicitly* (Kaern *et al.*, 2005). Especially, small numbers mean that the randomness of molecular encounters and the fluctuations in the transitions between the conformational states of a macromolecule become noticeable. *Deterministic* approaches, on the other hand, cannot capture the potentially significant effects of factors that cause *stochasticity*. Fluctuations in living systems may be much more than just a nuisance. Living systems change constantly and Fedoroff and Fontana (2002) pointed out the question of the source of this constant unfolding, adaptation, and change. If stochasticity is a fact of life, states are by definition metastable, and fluctuations can cause transitions between them. Thus, our simulation framework is a useful tool in considering stochasticity a foundation of this thesis as well as of structure definition in stochasticity. In this chapter we present a unifying framework to simulate sequence evolution with arbitrary complexity.

3.1 *In Silico* Sequence Evolution with Site-Specific Interactions

Evolutionary analysis of biological sequences typically assumes that sites are evolving independently of each other (cf. Tavaré, 1986). However, this simplifying assumption does not hold true generally. Thus, in recent years evolutionary models have been suggested to remedy this unsatisfactory situation. Markov models taking the base-pairings in stem regions of RNA molecules into account were among the first to model the process of evolution more realistically (Schöniger and von Haeseler, 1994; Tillier, 1994; Muse, 1995; Rzhetsky, 1995; Tillier and Collins, 1998; Savill *et al.*, 2000; Smith *et al.*, 2004). Models to detect protein sites with correlated patterns of evolution have also been proposed (Pollock *et al.*, 1999). Furthermore, models including selection against CpG-dinucleotides were studied as an example of overlapping context dependencies (Jensen and Pedersen, 2000; Arndt *et al.*, 2003; Siepel and Haussler, 2004). More specialised models with overlapping reading frames (Pedersen and Jensen, 2001) and with focus on protein structure, were also suggested (Robinson *et al.*, 2003). Recently, irreversible complex models with overlapping neighbouring nucleotide pairs were developed (Lunter and Hein, 2004) and Pedersen *et al.* (2004a) modeled the substitution process in protein-coding regions with embedded conserved RNA structures.

Modeling the evolution of a collection of homologous sequences is not simply an academic gimmick. A profound knowledge of how sequences evolve can help us to improve the reconstruction of phylogenetic trees based on sequences. However, the true mode of sequence evolution between homologous sequences is unknown with few exceptions. Therefore, simulating sequence evolution is helpful to investigate the performance of tree-building methods (Huelsenbeck, 1995). So far, different programs have been designed to simulate nucleotide sequences and protein sequences along a tree (Schöniger and von Haeseler, 1995b; Rambaut and Grassly, 1997; Grassly *et al.*, 1997; Yang, 1997; Stoye *et al.*, 1998; Nicholas *et al.*, 2000; Tufféry, 2002; Kosakovsky Pond *et al.*, 2005). One of the most widely used programs Seq-Gen (Rambaut and Grassly, 1997) has implemented a wide range of independent nucleotide substitution models (e.g., Jukes and Cantor, 1969; Kimura, 1980; Felsenstein, 1981; Hasegawa *et al.*, 1985). The PHASE package (Hudlot *et al.*, 2003) has implemented base-paired substitution models, but is specially designed for RNA sequences with secondary structure. To the best of our knowledge a general sequence simulation program including site-specific interactions based on a well defined neighbourhood system does not exist.

While progress has been made in devising new models for inference of sequences with dependencies among sites, there is a lack of simulation models which would allow the assessment of this progress. Especially if one wants to assess the robustness of phylogenetic inference, the models used for simulation need to be more accurate and complex descriptions of nature than those used for inference. Furthermore, simulations taking into account site-specific interactions that evolved along a phylogeny are of value in themselves because they contribute to our understanding of the intertwined relationship between

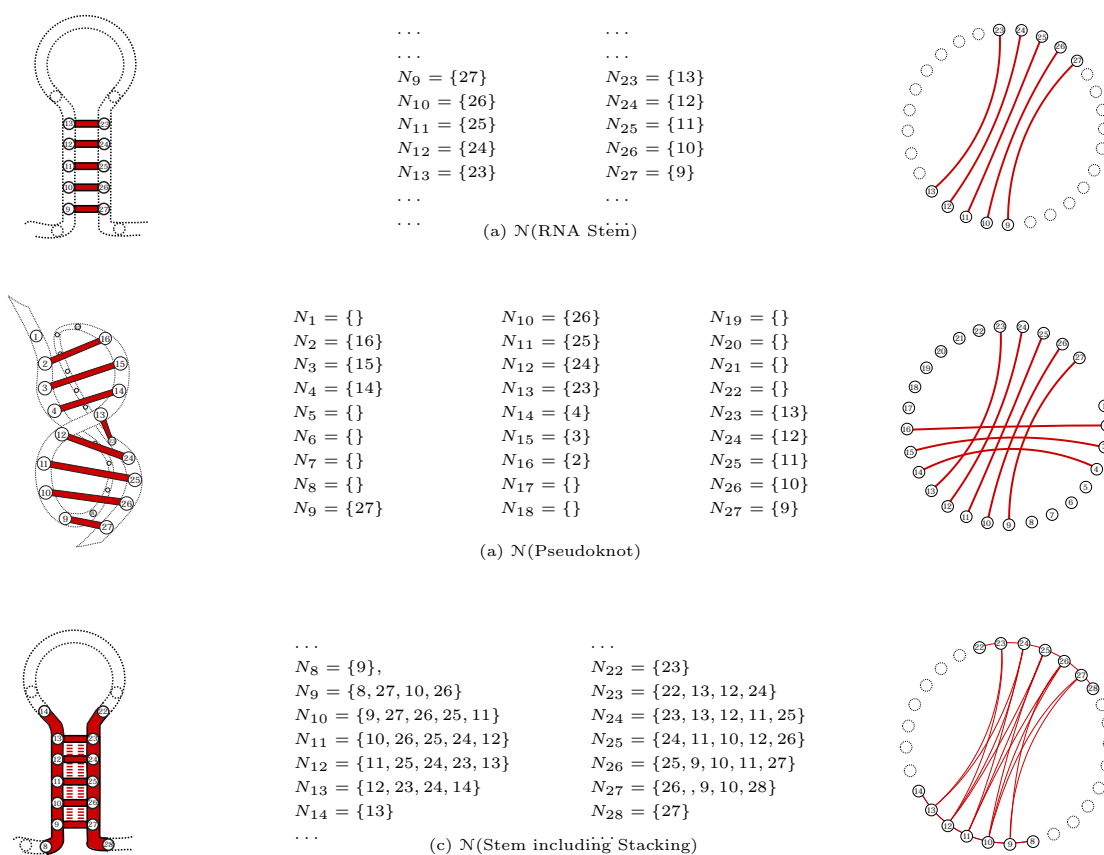


Figure 3.1: Three examples show how the neighbourhood system may be used to encode various structural elements in an RNA sequence. **Left:** schematic representations; **middle:** neighbourhood system notation; **right:** circle plots, useful to display complex features of molecules. (Sites are written in the circumference of a circle and interacting sites are connected by chords.) (a) Typical examples for interacting sites are base pairs in RNA stems. (b) Pseudoknots show intersecting edges in circle plots. (c) To take base stacking in RNA stems into account a lot of overlapping dependencies must be considered.

structure and substitution process. The use of supervised sequence evolution (i.e. simulations) allows us to control and study the extent of structural and sequence conservation. In the following section we describe the representation of site-specific interactions using a neighbourhood system. It allows for a universal description of arbitrarily complex dependencies among sites. We give some simple examples of how to apply neighbourhood systems to define various structural elements in RNA sequences. Then, we define a site and neighbourhood-dependent substitution process that permits a versatile description of the various evolutionary forces acting on single sites. We show that our method is useful to simulate the evolution e.g. of RNA sequences and structure simultaneously since it can take into account both, base-pairing counterparts as well as further interactions between nucleotides in sequences.

3.1.1 The Neighbourhood System of a Sequence

In the following, we describe for each site $k = 1, \dots, l$ in a (nucleotide) sequence $\mathbf{x} = (x_1, \dots, x_l)$ the interaction of k with other sites in \mathbf{x} . To this end, we introduce the neighbourhood system $\mathcal{N} = (N_k)_{k=1,2,\dots,l}$ such that:

1. $N_k \subset \{1, \dots, l\}, k \notin N_k$ for each k
2. If $i \in N_k$ then $k \in N_i$ for each i, k .

N_k contains all sites that interact with site k . With n_k we denote the cardinality of N_k , i.e. the number of sites that interact with k . The sites $\{1, \dots, l\}$ of the sequence together with the neighbourhood system \mathcal{N} correspond to a graph, with vertices $\mathcal{V} = \{1, \dots, l\}$ and edge set $\mathcal{E} = \{(k, i) | 1 \leq k \leq l, i \in N_k\}$. This graph can be visualised with a circle plot, where the vertices are arranged in a circle and the edges connect two vertices inside the circle. Using the notation of a neighbourhood system it is easily possible to encode various secondary and tertiary structural elements in a unifying framework.

Figure 3.1 illustrates some well-known RNA structures together with the corresponding neighbourhoods in the circle plot. A stem region of an RNA molecule is encoded by a neighbourhood system, with $n_k = 1$ for sites in a stem and $n_k = 0$ for sites in a loop, where $i \in N_k$ is the site that base-pairs with k (see Figure 3.1a). Similarly, we can encode a pseudoknot, again n_k is either zero or one. Here, the resulting circle plot shows intersecting edges (Figure 3.1b). One can proceed to model even more elaborate interactions. Figure 3.1c displays the neighbourhood system that results if interactions due to base-stacking are incorporated in our model. The corresponding circle graph displays many intersecting edges and takes into account overlapping dependencies. For example, site 11 is inter alia an element of the neighbourhoods N_{10} and N_{12} . While site 10 is not in N_{12} and vice versa. Figure 3.3 finally displays the interactions deduced from a ribozyme domain (Cate *et al.*, 1996), where site 153 interacts with sites 150, 223 and 250. The described interactions are crucial for the integrity of the molecule and should be taken into account when modeling the process of evolution.

3.1.2 An Evolutionary Model Including Neighbourhoods

In the previous section, we have introduced a tool to succinctly summarize interactions among sites in a (nucleotide) sequence. In the following, we need to superimpose an evolutionary dynamics which acts on the sites of a sequence and which takes into account these interactions.

Hence, we define a substitution process for every site k , where the substitution of a given nucleotide x_k by another one depends on the states $(x_{i_1}, \dots, x_{i_{n_k}})$ of the sites $i_1, \dots, i_{n_k} \in N_k$. To be more formal, we introduce at each site k a site-specific rate matrix Q_k . Thus $Q = \{Q_k | k = 1, \dots, l\}$ constitutes a collection of possibly different substitution models acting on the sequence and an annotation of correlations among sites. Contrary to standard models

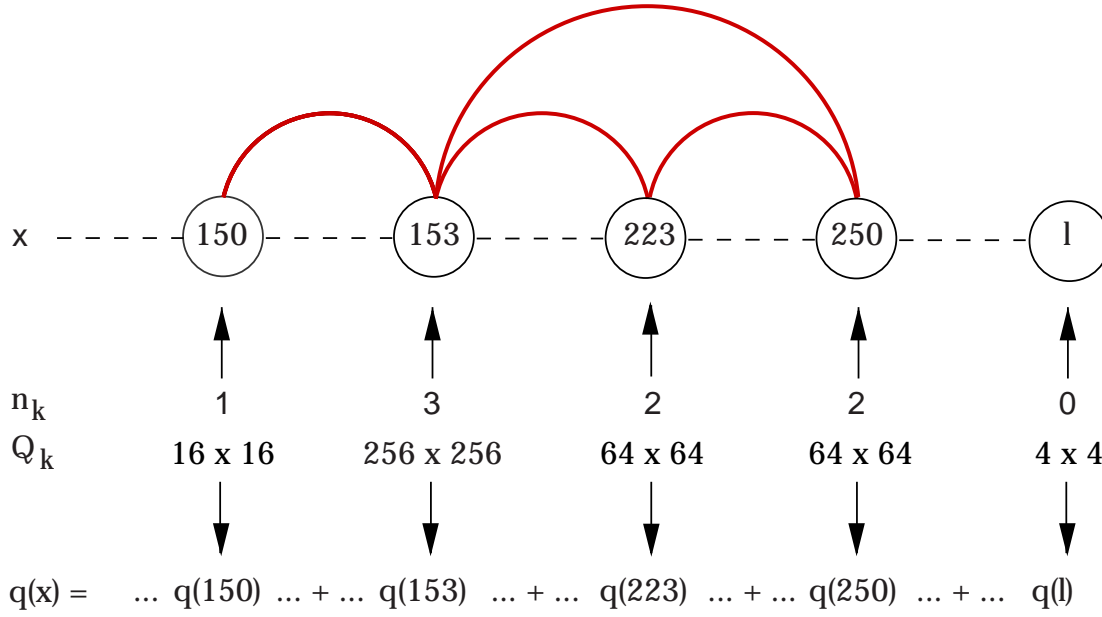


Figure 3.2: Example of a sequence \mathbf{x} with overlapping dependencies on site 153. Such dependencies occur e.g. in ribozyme domains (Cate *et al.*, 1996). The substitution rate for the whole sequence $q(\mathbf{x})$ is the sum of the rates of each site $q(k) = Q_k(\mathbf{s}_k, \mathbf{s}_k)$. The mononucleotide instantaneous substitution rate depends on the states of the neighbourhood system of this site at the instant of the substitution, described in the instantaneous rate matrix Q_k . The dimension of Q_k depends on the number of neighbours n_k at this site k .

which assume independent evolution of the sites, Q_k has dimensions $|\mathcal{A}|^{n_k+1} \times |\mathcal{A}|^{n_k+1}$, where $|\mathcal{A}|$ is the size of the alphabet. For the examples discussed here $\mathcal{A} = \{A, C, G, T\}$, or $\mathcal{A} = \{A, C, G, U\}$, hence $|\mathcal{A}| = 4$. Thus if $n_k = 0$, Q_k can be defined as one of the usual rate matrices on \mathcal{A} , i.e. we may assume a Jukes-Cantor (Jukes and Cantor, 1969) matrix, a Hasegawa-Kishino-Yano-Matrix or another independent model (Jukes and Cantor, 1969; Kimura, 1980; Felsenstein, 1981; Hasegawa *et al.*, 1985). If $n_k > 0$, then Q_k acts on subsequences of length $n_k + 1$. We impose the usual restriction that only one substitution per unit time is admissible (Schöniger and von Haeseler, 1994, 1995b). Moreover, in the model Q_k a substitution is only possible at site k . This restriction leads to sparse rate matrices. Let $\mathbf{s}_k = (x_k, x_{i_1}, \dots, x_{i_{n_k}}) \in \mathcal{A}^{n_k+1}$ represent the actual subsequence of sequence \mathbf{x} , where $\{i_1, \dots, i_{n_k}\} = N_k$ and let $\mathbf{y} = (y_0, y_1 \dots y_{n_k}) \in \mathcal{A}^{n_k+1}$ denote an arbitrary sequence of the same length. To avoid notational confusion, we assume $i_1 < i_2 < \dots < i_{n_k}$. With $(\pi_k(\mathbf{y}))$ we denote the stationary distribution of rate matrix

Q_k . Because only site k is allowed to vary, the entries of Q_k are given by

$$Q_k(\mathbf{s}_k, \mathbf{y}) = \begin{cases} \pi_k(\mathbf{y}) & \text{if } H(\mathbf{s}_k, \mathbf{y}) = 1 \text{ and } x_k \neq y_0 \\ - \sum_{\substack{\mathbf{z} \in \mathcal{A}^{n_k+1} \\ \mathbf{z} \neq \mathbf{s}_k}} Q_k(\mathbf{s}_k, \mathbf{z}) & \text{if } H(\mathbf{s}_k, \mathbf{y}) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

where the Hamming distance $H(\mathbf{s}_k, \mathbf{y})$ counts the number of differences between the sites of the subsequence \mathbf{s}_k and \mathbf{y} . In other words, an element of Q_k is greater zero if the last n_k sites in sequence \mathbf{s}_k and \mathbf{y} are pairwise identical. Thus, Q_k has $|\mathcal{A}|^{n_k+2}$ non-zero entries. We scale Q_k such that the number of substitutions d_k equals 1:

$$d_k = \sum_{\mathbf{z} \in \mathcal{A}^{n_k+1}} \pi_k(\mathbf{z}) \cdot |Q_k(\mathbf{z}, \mathbf{z})| = 1. \quad (3.2)$$

The rate matrix Q_k defined by Equation 6.2 defines the “strength” of interactions among sites in a neighbourhood by the frequencies of subsequences $\mathbf{y} \in \mathcal{A}^{n_k+1}$. Further generalizations are possible and we will discuss them later. The framework outlined here allows a rate matrix for each site in the sequence. To complete the discussion of the evolutionary process, we define the total instantaneous substitution rate for \mathbf{x} as

$$q(\mathbf{x}) = - \sum_{k=1}^l |Q_k(\mathbf{s}_k, \mathbf{s}_k)| \quad (3.3)$$

where \mathbf{s}_k is the subsequence of \mathbf{x} induced by N_k . Thus, if a nucleotide in \mathbf{x} is substituted the instantaneous rate may change. The new rate can be computed easily.

To illustrate the notation of Q_k , we continue with the examples from Figure 3.1. Fig. 3.1a displays the neighbourhood system for a stem region that mimics the doublet model for base-paired sites i and k , with $N_i = \{k\}$ and $N_k = \{i\}$. Tab. 3.2b, rewritten as a block matrix in Tab. 3.3a, displays a possible rate matrix acting on site k while taking into account the state at site i . Note that this matrix is reversible and has 15 parameters. Accordingly, we may define a similar rate matrix for site i (Table 3.3b). If $\pi_{\alpha\beta} = \pi_{\beta\alpha}$ for $\alpha, \beta \in \{A, C, G, U\}$, then both matrices have identical entries with 9 free parameters. If the matrices in Table 3.3 are applied to all sites in a stem, this gives the evolutionary process defined by Schöniger and von Haeseler (1994). The matrices can be summarized in a more condensed form. With (y_1, \dots, y_{n_k}) we denote a sequence of length n_k and $(x_k, y_1, \dots, y_{n_k})$ represents the current subsequence in \mathbf{x} as induced by N_k . The admissible substitutions for one site are written in the following submatrix (Equ. 3.4):

$$\begin{array}{cccc}
& \mathbf{A|A} & \mathbf{C|A} & \mathbf{G|A} & \mathbf{U|A} \\
\mathbf{A|A} & \left(\begin{array}{cccc} * & \pi_{CA} & \pi_{GA} & \pi_{UA} \\ \pi_{AA} & * & \pi_{GA} & \pi_{UA} \\ \pi_{AA} & \pi_{CA} & * & \pi_{UA} \\ \pi_{AA} & \pi_{CA} & \pi_{GA} & * \end{array} \right) & & & \\
\mathbf{C|A} & & & & \\
\mathbf{G|A} & & & & \\
\mathbf{U|A} & & & & \\
& \mathbf{A|C} & \mathbf{C|C} & \mathbf{G|C} & \mathbf{U|C} \\
\mathbf{A|C} & \left(\begin{array}{cccc} * & \pi_{CC} & \pi_{GC} & \pi_{UC} \\ \pi_{AC} & * & \pi_{GC} & \pi_{UC} \\ \pi_{AC} & \pi_{CC} & * & \pi_{UC} \\ \pi_{AC} & \pi_{CC} & \pi_{GC} & * \end{array} \right) & & & \\
\mathbf{C|C} & & & & \\
\mathbf{G|C} & & & & \\
\mathbf{U|C} & & & & \\
& \mathbf{A|G} & \mathbf{C|G} & \mathbf{G|G} & \mathbf{U|G} \\
\mathbf{A|G} & \left(\begin{array}{cccc} * & \pi_{CG} & \pi_{GG} & \pi_{UG} \\ \pi_{AG} & * & \pi_{GG} & \pi_{UG} \\ \pi_{AG} & \pi_{CG} & * & \pi_{UG} \\ \pi_{AG} & \pi_{CG} & \pi_{GG} & * \end{array} \right) & & & \\
\mathbf{C|G} & & & & \\
\mathbf{G|G} & & & & \\
\mathbf{U|G} & & & & \\
& \mathbf{A|U} & \mathbf{C|U} & \mathbf{G|U} & \mathbf{U|U} \\
\mathbf{A|U} & \left(\begin{array}{cccc} * & \pi_{CU} & \pi_{GU} & \pi_{UU} \\ \pi_{AU} & * & \pi_{GU} & \pi_{UU} \\ \pi_{AU} & \pi_{CU} & * & \pi_{UU} \\ \pi_{AU} & \pi_{CU} & \pi_{GU} & * \end{array} \right) & & & \\
\mathbf{C|U} & & & & \\
\mathbf{G|U} & & & & \\
\mathbf{U|U} & & & &
\end{array}$$

Table 3.1: The condensed matrix form is defined by four submatrices (A,C,G,U) of the type (3.4) as introduced in the text. Only substitutions at the current site k (bold) are admissible.

$$\begin{array}{cccc}
& \mathbf{A|}y_1, \dots, y_{n_k} & \mathbf{C|}y_1, \dots, y_{n_k} & \mathbf{G|}y_1, \dots, y_{n_k} & \mathbf{U|}y_1, \dots, y_{n_k} \\
\mathbf{A|}y_1, \dots, y_{n_k} & \left(\begin{array}{cccc} * & \pi_{C|y_1, \dots, y_{n_k}} & \pi_{G|y_1, \dots, y_{n_k}} & \pi_{U|y_1, \dots, y_{n_k}} \\ \pi_{A|y_1, \dots, y_{n_k}} & * & \pi_{G|y_1, \dots, y_{n_k}} & \pi_{U|y_1, \dots, y_{n_k}} \\ \pi_{A|y_1, \dots, y_{n_k}} & \pi_{C|y_1, \dots, y_{n_k}} & * & \pi_{U|y_1, \dots, y_{n_k}} \\ \pi_{A|y_1, \dots, y_{n_k}} & \pi_{C|y_1, \dots, y_{n_k}} & \pi_{G|y_1, \dots, y_{n_k}} & * \end{array} \right) & & & \\
\mathbf{C|}y_1, \dots, y_{n_k} & & & & \\
\mathbf{G|}y_1, \dots, y_{n_k} & & & & \\
\mathbf{U|}y_1, \dots, y_{n_k} & & & &
\end{array} \tag{3.4}$$

For example the 16×16 doublet model (Table 3.3) is defined by four matrices of this type (Table 3.1). Generally, after normalisation we can divide the $A^{n_k+1} \times A^{n_k+1}$ instantaneous rate matrix in 4^{n_k} submatrices of the type as illustrated in (3.4).

The reader may notice that our definition of a rate matrix is not limited to the F81 type of substitution matrices (Felsenstein, 1981). The submatrix (3.4) can be extended to any type of rate matrix by e.g. introducing specific substitution and irreversible rates. It is, for example, easily possible to include a transition-transversion parameter (see chapter 6, equation 6.3). In section 3.2 we introduce the extended framework of SSSI. However, for the time being, we think that the frequencies of subsequences provide a reasonably good description of interactions among sites.

(k, i)	AA AC AG AU	CA CC CG CU	GA GC GG GU	UA UC UG UU
AA	* π_{AC} π_{AG} π_{AU}	π_{CA} - - -	π_{GA} - - -	π_{UA} - - -
AC	π_{AA} * π_{AG} π_{AU}	- π_{CC} - -	- π_{GC} - -	- π_{UC} - -
AG	π_{AA} π_{AC} * π_{AU}	- - π_{CG} -	- - π_{GG} -	- - π_{UG} -
AU	π_{AA} π_{AC} π_{AG} *	- - - π_{CU}	- - - π_{GU}	- - - π_{UU}
CA	π_{AA} - - -	* π_{CC} π_{CG} π_{CU}	π_{GA} - - -	π_{UA} - - -
CC	- π_{AC} - -	π_{CA} * π_{CG} π_{CU}	- π_{GC} - -	- π_{UC} - -
CG	- - π_{AG} -	π_{CA} π_{CC} * π_{CU}	- - π_{GG} -	- - π_{UG} -
CU	- - - π_{AU}	π_{CA} π_{CC} π_{CG} *	- - - π_{GU}	- - - π_{UU}
GA	π_{AA} - - -	π_{CA} - - -	* π_{GC} π_{GG} π_{GU}	π_{UA} - - -
GC	- π_{AC} - -	- π_{CC} - -	π_{GA} * π_{GG} π_{GU}	- π_{UC} - -
GG	- - π_{AG} -	- - π_{CG} -	π_{GA} π_{GC} * π_{GU}	- - π_{UG} -
GU	- - - π_{AU}	- - - π_{CU}	π_{GA} π_{GC} π_{GG} *	- - - π_{UU}
UA	π_{AA} - - -	π_{CA} - - -	π_{GA} - - -	* π_{UC} π_{UG} π_{UU}
UC	- π_{AC} - -	- π_{CC} - -	- π_{GC} - -	π_{UA} * π_{UG} π_{UU}
UG	- - π_{AG} -	- - π_{CG} -	- - π_{GG} -	π_{UA} π_{UC} * π_{UU}
UU	- - - π_{AU}	- - - π_{CU}	- - - π_{GU}	π_{UA} π_{UC} π_{UG} *

(k, i)	AA AC AG AU	CA CC CG CU	GA GC GG GU	UA UC UG UU
AA	* - - -	π_{CA} - - -	π_{GA} - - -	π_{UA} - - -
AC	- * - -	- π_{CC} - -	- π_{GC} - -	- π_{UC} - -
AG	- - * -	- - π_{CG} -	- - π_{GG} -	- - π_{UG} -
AU	- - - *	- - - π_{CU}	- - - π_{GU}	- - - π_{UU}
CA	π_{AA} - - -	* - - -	π_{GA} - - -	π_{UA} - - -
CC	- π_{AC} - -	- * - -	- π_{GC} - -	- π_{UC} - -
CG	- - π_{AG} -	- - * -	- - π_{GG} -	- - π_{UG} -
CU	- - - π_{AU}	- - - *	- - - π_{GU}	- - - π_{UU}
GA	π_{AA} - - -	π_{CA} - - -	* - - -	π_{UA} - - -
GC	- π_{AC} - -	- π_{CC} - -	- * - -	- π_{UC} - -
GG	- - π_{AG} -	- - π_{CG} -	- - * -	- - π_{UG} -
GU	- - - π_{AU}	- - - π_{CU}	- - - *	- - - π_{UU}
UA	π_{AA} - - -	π_{CA} - - -	π_{GA} - - -	* - - -
UC	- π_{AC} - -	- π_{CC} - -	- π_{GC} - -	- * - -
UG	- - π_{AG} -	- - π_{CG} -	- - π_{GG} -	- - * -
UU	- - - π_{AU}	- - - π_{CU}	- - - π_{GU}	- - - *

Table 3.2: (a) **top matrix:** this is a description of the process of base-pairs in RNA helical regions as an extended Fel 81 model with joint substitution events and the restriction that only one substitution per unit time is admissible of one base of the doublet (Schöniger and von Haeseler, 1994), see Equ. 2.24 in Chap. 2). The instantaneous rate matrix Q has 16×16 dimension. The diagonal elements (*) are defined by the mathematical requirement that the sum of each row is zero. A '-' represents zero in the rate matrix. (b) **bottom matrix:** one example for an instantaneous rate matrix Q_k of a site k with $n_k = 1$ of our SISSI-framework for the same process acting on site k while taking into account site i . However, in contrast to the top matrix, only the current site k (bold) is allowed to mutate. The rate matrix Q_k has also 16×16 dimensions, but is very sparse.

(k, i)	AA CA GA UA	AC CC GC UC	AG CG GG UG	AU CU GU UU
AA	* π_{CA} π_{GA} π_{UA}	- - - -	- - - -	- - - -
CA	π_{AA} * π_{GA} π_{UA}	- - - -	- - - -	- - - -
GA	π_{AA} π_{CA} * π_{UA}	- - - -	- - - -	- - - -
UA	π_{AA} π_{CA} π_{GA} *	- - - -	- - - -	- - - -
AC	- - - -	* π_{CC} π_{GC} π_{UC}	- - - -	- - - -
CC	- - - -	π_{AC} * π_{GC} π_{UC}	- - - -	- - - -
GC	- - - -	π_{AC} π_{CC} * π_{UC}	- - - -	- - - -
UC	- - - -	π_{AC} π_{CC} π_{GC} *	- - - -	- - - -
AG	- - - -	- - - -	* π_{CG} π_{GG} π_{UG}	- - - -
CG	- - - -	- - - -	π_{AG} * π_{GG} π_{UG}	- - - -
GG	- - - -	- - - -	π_{AG} π_{CG} * π_{UG}	- - - -
UG	- - - -	- - - -	π_{AG} π_{CG} π_{GG} *	- - - -
AU	- - - -	- - - -	- - - -	* π_{CU} π_{GU} π_{UU}
CU	- - - -	- - - -	- - - -	π_{AU} * π_{GU} π_{UU}
GU	- - - -	- - - -	- - - -	π_{AU} π_{CU} * π_{UU}
UU	- - - -	- - - -	- - - -	π_{AU} π_{CU} π_{GU} *

(k, i)	AA AC AG AU	CA CC CG CU	GA GC GG GU	UA UC UG UU
AA	* π_{AC} π_{AG} π_{AU}	- - - -	- - - -	- - - -
AC	π_{AA} * π_{AG} π_{AU}	- - - -	- - - -	- - - -
AG	π_{AA} π_{AC} * π_{AU}	- - - -	- - - -	- - - -
AU	π_{AA} π_{AC} π_{AG} *	- - - -	- - - -	- - - -
CA	- - - -	* π_{CC} π_{CG} π_{CU}	- - - -	- - - -
CC	- - - -	π_{CA} * π_{CG} π_{CU}	- - - -	- - - -
CG	- - - -	π_{CA} π_{CC} * π_{CU}	- - - -	- - - -
CU	- - - -	π_{CA} π_{CC} π_{CG} *	- - - -	- - - -
GA	- - - -	- - - -	* π_{GC} π_{GG} π_{GU}	- - - -
GC	- - - -	- - - -	π_{GA} * π_{GG} π_{GU}	- - - -
GG	- - - -	- - - -	π_{GA} π_{GC} * π_{GU}	- - - -
GU	- - - -	- - - -	π_{GA} π_{GC} π_{GG} *	- - - -
UA	- - - -	- - - -	- - - -	* π_{UC} π_{UG} π_{UU}
UC	- - - -	- - - -	- - - -	π_{UA} * π_{UG} π_{UU}
UG	- - - -	- - - -	- - - -	π_{UA} π_{UC} * π_{UU}
UU	- - - -	- - - -	- - - -	π_{UA} π_{UC} π_{UG} *

Table 3.3: **(a) top matrix:** One example for an instantaneous rate matrix Q_k of a site k with $n_k = 1$. Typical examples are base pairs in RNA stems, illustrated in Figure 3.1a. It is the same instantaneous matrix like in Tab. 3.2 in the bottom, rewritten as a block matrix. This is a rate matrix acting on site k while taking into account site i . The diagonal elements (*) are defined by the mathematical requirement that the sum of each row is zero. The rate matrix Q_k has 16×16 dimensions, but is very sparse. Only the current site k (bold) is allowed to mutate. A '-' represents zero in the rate matrix. **(b) bottom matrix:** Corresponding rate matrix Q_k for site i . Here only substitutions on the current site i (bold) are allowed.

3.1.3 Simulations

In the following, a neighbourhood system \mathcal{N} and a collection of site-specific rate matrices $Q = \{Q_k | k = 1, \dots, l\}$ are defined. We start at mutational time 0 with sequence $\mathbf{x}(0)$ which evolves according to (\mathcal{N}, Q) and we want to generate a sequence $\mathbf{x}(d)$ after d expected substitutions. At time $d = 0$ the instantaneous substitution rate equals $q(\mathbf{x})$ (see equation 6.5). We draw a random time d_r from an exponential distribution with parameter $q(\mathbf{x})$. If $d_r < d$ then a substitution takes place in \mathbf{x} . We pick a site k with probability

$$\mathbb{P}(k) = \frac{|Q_k(\mathbf{s}_k, \mathbf{s}_k)|}{q(\mathbf{x})}, \quad (3.5)$$

the relative mutability at that site. For a chosen site k , the nucleotide x_k will be replaced by a new nucleotide y_0 with probability

$$\mathbb{P}(x_k \rightarrow y_0) = \frac{Q_k(\mathbf{s}_k, \mathbf{y})}{|Q_k(\mathbf{s}_k, \mathbf{s}_k)|}. \quad (3.6)$$

Subsequently, the actual time is updated to $d \leftarrow d - d_r$ and $q(\mathbf{x})$ is recomputed based on the new sequence and the simulation continues. This procedure is summarized in the following pseudo code:

Algorithm 3.1.1: *Computing a sequence $\mathbf{x}(d)$, d substitutions away from $\mathbf{x}(0)$.*

1. Compute $q(\mathbf{x})$ for $\mathbf{x}(0)$;
 2. Draw d_r from the exponential distribution with parameter $q(\mathbf{x})$;
 3. Init $d_c = d_r$;
- while** $d_c < d$ **do**
- a. Choose a site k with $\mathbb{P}(k)$ (Equ.3.5) ;
 - b. Replace \mathbf{x}_k with \mathbf{y}_0 with $\mathbb{P}(x_k \rightarrow y_0)$ (Equ.3.6) ;
 - c. Update $q(\mathbf{x})$ based on N_k ;
 - d. Draw new d_r from the exponential distribution with parameter $q(\mathbf{x})$;
 - e. Set $d_c = d_r + d_c$;
- end**
-

Finally, this procedure is applied recursively through a given rooted or unrooted tree topology, where the branch lengths are specified by the expected number of substitutions. This method is implemented in the program SISSI (Simulating Site-Specific Interactions).

3.1.4 Results

Simulations employing a neighbourhood system, incorporating artificial or known structural features, were run on an ordinary PC. Although the models are more complex than

	$f_{n_k=1}$				$f_{n_k=0}$
	<i>second nucleotide in doublet</i>				
<i>first nucleotide</i>	A	C	G	U	
A	0.000423	0.004228	0.012685	0.169133	0.422360
C	0.004228	0.000423	0.262156	0.000423	0.105590
G	0.012685	0.262156	0.000423	0.042283	0.236025
U	0.169133	0.000423	0.042283	0.016915	0.236025

Table 3.4: Counted doublet frequencies $f_{n_k=1}$ and single frequencies $f_{n_k=0}$ from a RNase P sequence of *Bacillus subtilis* (accessions number: M13175) taken from the RNase P database (Brown, 1999). Counts of $\alpha\beta$ and $\beta\alpha$ are symmetrized. If a nucleotide pair is not present then the count is set to 0.1.

the well-known independent-sites models the computing time for simulations is satisfactory. If $n_k = l - 1$ for all $k = 1, \dots, l$ then the run time increases quadratically with sequence length l . Run time increases linearly as a function of the number of taxa or the total branch length of the tree.

As an illustrative example, we used the neighbourhood system of RNase P from *Bacillus subtilis* with 401 sites (Figure 3.3) taken from the RNase P database (Brown, 1999). To specify the rate matrices we used the frequencies of the nucleotides $\{A, C, G, U\}$ for sites evolving independently ($n_k = 0$) and the doublet frequencies for sites with $n_k = 1$ from one corresponding sequence in the database (Table 3.4). We used only these two matrices in the simulations. In the sequence 41.15% sites evolve independently and 58.85% evolve under dependencies. Having specified (N, Q) it took on average one second to simulate a dataset of 100 sequences along a tree with mean branch length 0.3.

Figure 3.3 shows the accumulation of observed sequence differences per site as the number of substitutions per site increases. The curve shows the expected saturation behaviour as d goes to infinity. The simulated curve lies between the theoretical curves we obtain for the FEL81 model and the doublet model.

We do not know the expected number of substitutions in real data. The different speeds of accumulation of observed differences have a great impact on the estimation of the number of substitutions. With our method we can investigate the relationship between the numbers of substitutions per site and the number of observed differences simultaneously for different neighbourhood systems with varying complexity.

For non overlapping sites in the neighbourhood system and small number of neighbours n_k it is possible to calculate the number of substitutions and the number of observed differences analytically (von Haeseler and Schöniger, 1998). Our simulation results agree with the expected numbers derived from appropriately weighting the expected numbers of observed differences for independent and dependent sites of the neighbourhood system of the RNase P of *Bacillus subtilis* (Figure 3.3).

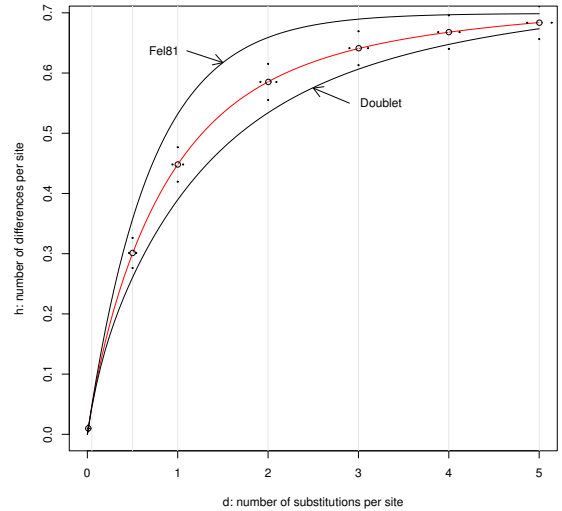


Figure 3.3: Left: this circle plot illustrates known structure features of the RNase P of *Bacillus subtilis* with 401 sites according to the RNase P database (Brown, 1999). *Right:* Relationship between number of substitutions per site d and number of observed differences per site h . **Lines:** Analytically calculated with the frequencies in Table 3.4. Upper line: Only sites with $n_k = 0$ (F81); lower line: Only sites with $n_k = 1$ (doublet model of Schöniger and von Haeseler, 1994); middle line: 41.15% sites with $n_k = 0$ and 58.85% sites with $n_k = 1$. **Circles:** Mean and standard deviation for number of substitutions and the corresponding differences under 1000 simulations with SISSI, the neighbourhood system of RNase P of *Bacillus subtilis* (Fig. left), the frequencies in Table 3.4 and the expected number of substitutions 0.01, 0.5, 1, 2, 3, 4 and 5 (x-axis).

Furthermore, in Fig. 3.4 we simulate alignments along a branch with one number of substitutions per site with a neighbourhood system with 4000 independent sites and 996 dependent sites in the middle, building a hairpin with a loop of four independent sites. The simulation parameters are given in Tab. 3.4. Then, we estimated the rates using maximum likelihoods with rate heterogeneity with the tree reconstruction programs RAxML (Stamatakis, 2006) using 2 categories of rates and IQPNNI (Vinh and von Haeseler, 2004) using site-specific rates. Although IQPNNI and RAxML estimate slightly different rates, not surprisingly, Fig. 3.4 shows that the estimated rates correlated with the annotation of the sites through the neighbourhood system. This reflects Equ. 6.5. Site-specific interactions give raise to rate variations over sites. Therefore, at first glance it makes no sense to further combine rate heterogeneity with site-specific interactions. Even so, in Chap. 6 we show how practical a site-specific scaling factor can be in our SISSI framework. Please refer for this extended algorithm to Sec 6.2.1 and 6.2.1.

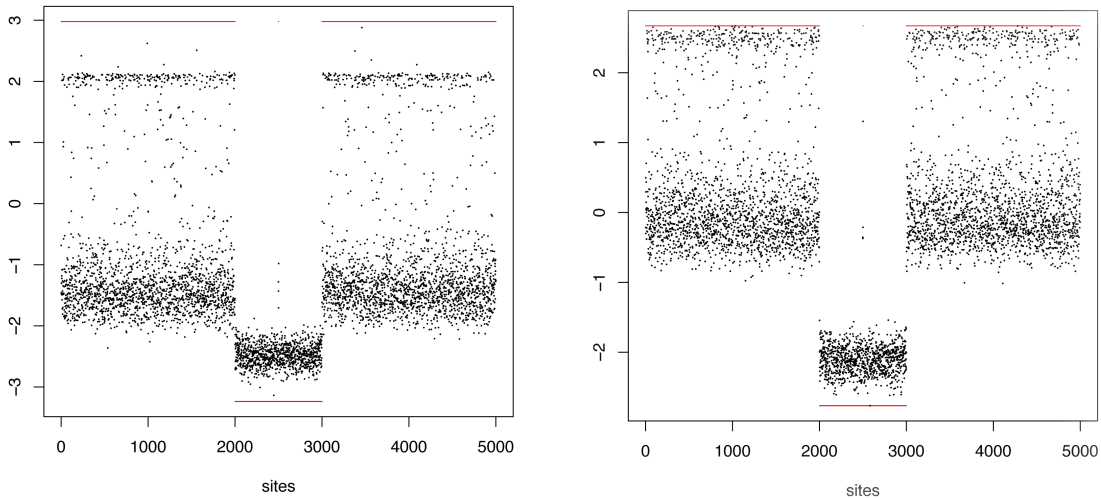


Figure 3.4: Rate heterogeneity as an indicator of structure: we simulate alignments along a star topology with 100 taxa and one number of substitutions per site using an neighbourhood system with 4000 independent sites and 996 dependent sites in the middle, building a hairpin with a loop of four independent sites. In both figures the upper red line represents the independent site, while the bottom red line represents the dependent sites. The figures show the estimated log rates on the y axis using maximum likelihoods with rate heterogeneity. **Left:** IQPNNI using site-specific rates. **Right:** RAxML using 2 categories of rates. The estimated rates correlated with the annotation of the sites (red line) through the neighbourhood system, although IQPNNI and RAxML estimate the rates slightly differently.

3.2 Extension of The Framework

So far, we have discussed types of rate matrices of the F81 (Felsenstein, 1981) model. The number of free parameters of the rate matrix Q_k increases exponentially with the number of neighbours n_k according to $|\mathcal{A}|^{n_k+1} - 1$. SISSI is in principal not limited to this type of substitution process. Fortunately, SISSI also allows other models with additional mechanistic assumptions or with fewer free parameters. The most general framework is SISSI with a parameter as a function for the neighbourhood system at site k .

Here, we give a framework for substitution models, which combine traditional phylogenetic models with site specific interactions. We introduce a parameter $\gamma_{(\mathbf{s}_k, \mathbf{y})} > 0$ as a function for the neighbourhood system at the current site k to incorporate the effect of neighbourhood constraints. Then we can define for $k = 1, \dots, l$ the instantaneous rate matrix of the composite model as

$$Q_k^\gamma(\mathbf{s}_k, \mathbf{y}) = \begin{cases} \gamma_{(\mathbf{s}_k, \mathbf{y})} \tilde{Q}_k(\mathbf{s}_k, \mathbf{y}) & \text{if } H(\mathbf{s}_k, \mathbf{y}) = 1 \text{ and } x_k \neq y_0 \\ - \sum_{\substack{\mathbf{z} \in \mathcal{A}^{n_k+1} \\ \mathbf{z} \neq \mathbf{s}_k}} Q_k^\gamma(\mathbf{s}_k, \mathbf{z}) & \text{if } H(\mathbf{s}_k, \mathbf{y}) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.7)$$

x_k	A				C				G				U			
	A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
A	*	$\beta\pi_C$	$\alpha\pi_G$	$\lambda\beta\pi_U$	*	$\beta\pi_C$	$\lambda\alpha\pi_G$	$\beta\pi_U$	*	$\lambda\beta\pi_C$	$\alpha\pi_G$	$\lambda\beta\pi_U$	*	$\frac{1}{\lambda}\beta\pi_C$	$\alpha\pi_G$	$\frac{1}{\lambda}\beta\pi_U$
C	$\beta\pi_A$	*	$\beta\pi_G$	$\lambda\alpha\pi_U$	$\beta\pi_A$	*	$\beta\pi_G$	$\alpha\pi_U$	$\beta\pi_A$	*	$\beta\pi_G$	$\alpha\pi_U$	$\beta\pi_A$	*	$\beta\pi_G$	$\alpha\pi_U$
G	$\alpha\pi_A$	$\beta\pi_C$	*	$\lambda\beta\pi_U$	$\frac{1}{\lambda}\alpha\pi_A$	$\frac{1}{\lambda}\beta\pi_C$	*	$\frac{1}{\lambda}\beta\pi_U$	$\alpha\pi_A$	$\lambda\beta\pi_C$	*	$\lambda\beta\pi_U$	$\alpha\pi_A$	$\frac{1}{\lambda}\beta\pi_C$	*	$\frac{1}{\lambda}\beta\pi_U$
U	$\beta\pi_A$	$\alpha\pi_C$	$\beta\pi_G$	*	$\beta\pi_A$	$\alpha\pi_C$	$\lambda\beta\pi_G$	*	$\frac{1}{\lambda}\beta\pi_A$	$\alpha\pi_C$	$\frac{1}{\lambda}\beta\pi_G$	*	$\lambda\beta\pi_A$	$\alpha\pi_C$	$\lambda\beta\pi_G$	*

Table 3.5: Example for sites with only one neighbour $|N_k| = 1$ for Q_k^γ (Equ. 3.7) with the HKY model (see Tab. 2.4) as a nullmodel and one pairing parameter λ (Equ. 3.8). It distinguishes between transition α and transversion β and π_j is the equilibrium frequency of nucleotide j with $\sum_{j \in \mathcal{A}} \pi_j = 1$. The diagonal elements (*) of each matrix are defined by the mathematical requirement that the sum of each row is zero. Only non-zero entries are shown.

where $\tilde{Q}_k(\mathbf{s}_k, \mathbf{y})$ denotes the rates given a chosen “original” instantaneous rate matrix $\tilde{\mathbf{Q}}$, which we call nullmodel. The parameter $\gamma_{(\mathbf{s}_k, \mathbf{y})}$ modifies the rate, only when $\gamma_{(\mathbf{s}_k, \mathbf{y})} = 1$ is the rate $Q_k^\gamma(\mathbf{s}_k, \mathbf{y})$ equal to the rate of the nullmodel $\tilde{Q}_k(\mathbf{s}_k, \mathbf{y})$. Note, that we do not need for each site a different parameter. In the next subsection, we present how to use one *pairing parameter* $\lambda > 1$ to define the parameter $\gamma_{(\mathbf{s}_k, \mathbf{y})}$. Later we use the parameter $\gamma_{(\mathbf{s}_k, \mathbf{y})}$ as a function of energy values.

3.2.1 SISSI: Simple Phylogenetic RNA Models

For a simple example like secondary structure of RNA stems, we adopt the doublet model of Muse (1995) (see Sec. 2.2.2). If a stem structure is favored, the instantaneous rate of a change from an unpaired state to a paired state should be greater than the corresponding rate, when the sites are independent. Similarly, changes from a paired state to an unpaired state should occur with lower rates. Muse introduces the *pairing parameter* $\lambda > 1$, which we use in this example for $\gamma_{(\mathbf{s}_k, \mathbf{y})}$, thus

$$\gamma_{(\mathbf{s}_k, \mathbf{y})} = \begin{cases} \lambda & \text{if pairing gained} \\ 1 & \text{if pairing unchanged} \\ \frac{1}{\lambda} & \text{if pairing lost} \end{cases} \quad (3.8)$$

Table 3.5 shows the resulting rate matrix with equation 3.7 for $|N_k| = 1$ using the HKY model Hasegawa *et al.* (1985) as a nullmodel $\tilde{Q}_k(\mathbf{s}_k, \mathbf{y})$. This model has five free parameters, three for the frequencies, one for transition and transversion ($\alpha + 2\beta = 1$) and the pairing parameter). With a refined $\gamma_{(\mathbf{s}_k, \mathbf{y})}$, with one parameter more, it is possible to treat the wobble pair GU as an intermediate state (see Muse, 1995). In the same way, we can include further 16×16 models, which are described in the introduction (Chap. 2, Subsec. 2.2.2). Moreover, all other partition models, like codon models (Sec. 2.2.2), given any cardinality n_k , using the usual restriction for one number of substitution in unit time, can be included in our framework: e.g to pre-specified dependency structure of the codons (Anisimova and Kosiol, 2009), as well as protein-coding regions that also encode the formation of conserved RNA structures (Pedersen *et al.*, 2004b) or RNA sequences, which can encode secondary structure as well as an amino-acid sequence (Pedersen *et al.*, 2004a).

Joint modelling of sites assumes that sites that are correlated evolve independently of other correlated sites. However, for simulations with our method it is possible to extend this to including overlapping dependencies.

3.3 SISSI with Energy

In the nearest neighbour model (Chap. 2, Subsec. 2.1.3) energies are not assigned to single base-pairs but rather to neighbouring base-pairs that stack on each other. Thus, including the stacking interactions between residue pairs that are adjacent in an RNA with additional energy values (Subsec. 2.1.3) is an important step to fill the gap between thermodynamic and phylogenetic research. So far, only one approach exists, where the relative rate of sequence evolution is affected by an approximated free energy of RNA. Yu and Thorne (2006) propose an evolutionary model, in which the approximate free energy of a secondary structure is used as a surrogate for fitness (see Subsec. 2.2.2). This process results in a sparse and very high dimensional instantaneous rate matrix Q (with sequence length l the dimension is $4^l \times 4^l$). However, it is not feasible to compute the matrix exponential Q unless l is small.

Instead of using an instantaneous rate matrix that specifies rates of change from each possible sequence to each other possible sequence, we are using our framework with a collection of site-specific matrices corresponding to the neighbourhood system based on the energy model. Here, we use the neighbourhood parameter $\gamma_{(\mathbf{s}_k, \mathbf{y})}$ as a function of energy values ΔG^0 ,

$$\gamma_{(\mathbf{s}_k, \mathbf{y})} := e^{\sigma(|\Delta G_{\mathbf{s}_k}^0 - \Delta G_{\mathbf{a}}^0| - |\Delta G_{\mathbf{y}}^0 - \Delta G_{\mathbf{a}}^0|)} \quad (3.9)$$

where σ is the Boltzmann factor, $\Delta G_{\mathbf{s}_k}^0$ is the energy value for the subsequence \mathbf{s}_k , $\Delta G_{\mathbf{y}}^0$ for an arbitrary sequence \mathbf{y} and $\Delta G_{\mathbf{a}}^0$ is a target-energy value, e.g. the energy value of the ancestral subsequence \mathbf{a}_k at time 0. This formulates substitution rates in terms of expected free energy.

If we use an independent-site HKY model as a null model and a site-specific scaling factor, the process is identical to Yu and Thorne (2006) in our notation corresponding to

$$\gamma_{(\mathbf{s}_k, \mathbf{y})} := e^{s(E(\mathbf{s}_k) - E(\mathbf{y}))}. \quad (3.10)$$

When s is zero, $\gamma_{(\mathbf{s}_k, \mathbf{y})}$ is one and the secondary structure does not affect the substitution rates. Thus, s links genotype and phenotype by treating the free energy as a surrogate for fitness. However, our simulation approach allow to combine with arbitrary independent and dependent model, as well as different local null models.

In the following, we are explaining the matrices for a position in an infinite helix with the neighbourhood parameter of Equ. 3.9. Including the other multiple factors of energy values described in Chap. 2, Subsec. 2.1.3 is straightforward.

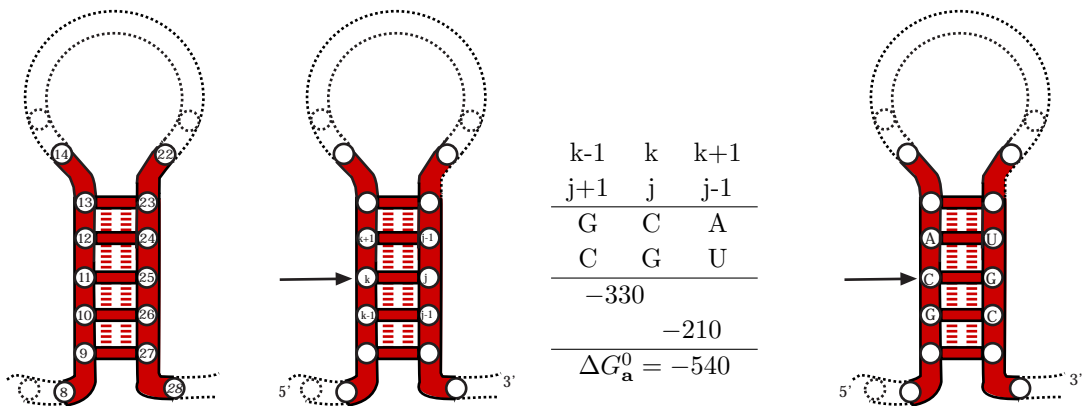


Figure 3.5: To take base stacking in RNA stems into account overlapping dependencies must be considered, see also Fig. 3.1. The cardinality of position k in the middle of a helix is $n_k = 5$ with the neighbourhood system $N_k = \{x_{k-1}, x_{k+1}, x_{j-1}, x_j, x_{j+1}\}$. Q_k^{stack} has dimensions $4^6 \times 4^6$, which can be divided into $4^5 = 1024$ submatrices of size 4×4 . Energies are given in units of 0.01 kcal/mol and details are described in the text.

3.3.0.1 Stacking Interactions

Most of the stabilizing energy in RNA secondary structures comes from stacking interactions of neighbouring base-pairs and we start with an example of an infinite helix. Fig. 3.5 shows that the cardinality of position i in the middle of a helix is $n_k = 5$ with the neighbourhood System $N_i = \{x_{k-1}, x_{k+1}, x_{j-1}, x_j, x_{j+1}\}$. Thus, Q_k^{stack} has dimensions $4^6 \times 4^6$, which can be divided in 4^5 submatrices (4×4). Then, the parameter for the neighbourhood constraints is given by

$$\gamma(x_k | x_{k-1}, x_{k+1}, x_{j-1}, x_j, x_{j+1}, y_0 | x_{k-1}, x_{k+1}, x_{j-1}, x_j, x_{j+1}). \quad (3.11)$$

As usually, we scale Q_k^{stack} (4096×4096) such that the number of substitutions d_k equals 1.

$$d_k = - \sum_{\mathbf{z} \in \mathcal{A}^6} \pi_k(\mathbf{z}) \cdot Q_k^\gamma(\mathbf{z}, \mathbf{z}) = 1, \quad (3.12)$$

Now, after the normalisation, we can divide the $4^6 \times 4^6$ matrix into the submatrices. We give an example for one of the 1024 possible submatrices for the subsequence GAUGC and compute the parameter $\gamma(x_k | GAUGC \rightarrow y_0 | GAUGC)$, as a function of energy values in a helix including the free energies for stacked pairs identical to the one described in (Mathews and Turner, 2006). The model is implemented using the C code library of the Vienna Package (Hofacker *et al.*, 1994b). All energies should be given as integers in units of 0.01 kcal/mol. The energy for the subsequence is computed by the sum of its corresponding pairs. Then, we compute the parameter for the neighbourhood constraint for each entry of Q_k^{stack} with Equ. 3.9. Below, we show some examples for the parameters.

$$\begin{aligned}
\gamma(C|GAUGC \rightarrow U|GAUGC) &= e^{|-540-(-540)|-|(-210)-(-540)|} = e^{-330} < 1 \\
\gamma(U|GAUGC \rightarrow C|GAUGC) &= e^{|-210-(-540)|-|(-540)-(-540)|} = e^{330} > 1 \\
\gamma(U|GAUGC \rightarrow A|GAUGC) &= e^{|-210-(-540)|-|0-(-540)|} = e^{-210} < 1 \\
\gamma(A|GAUGC \rightarrow U|GAUGC) &= e^{|0-(-540)|-|(-210)-(-540)|} = e^{210} > 1
\end{aligned}$$

The first equation line represents the example of Figure 3.5 and our example submatrix is given by

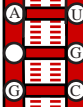
$$\begin{array}{c}
A|GAUGC \\
C|GAUGC \\
G|GAUGC \\
U|GAUGC
\end{array}
\begin{array}{c}
C|GAUGC \\
G|GAUGC \\
U|GAUGC
\end{array}
\begin{array}{c}
G|GAUGC \\
U|GAUGC
\end{array}
\begin{array}{c}
U|GAUGC
\end{array}
\begin{array}{c}
A|GAUGC \\
C|GAUGC \\
G|GAUGC \\
U|GAUGC
\end{array}
\begin{pmatrix}
* & e^{540} \cdot \tilde{Q}_k & 1 \cdot \tilde{Q}_k & e^{210} \cdot \tilde{Q}_k \\
e^{-540} \cdot \tilde{Q}_k & * & e^{-540} \cdot \tilde{Q}_k & e^{-330} \cdot \tilde{Q}_k \\
1 \cdot \tilde{Q}_k & e^{540} \cdot \tilde{Q}_k & * & e^{210} \cdot \tilde{Q}_k \\
e^{-210} \cdot \tilde{Q}_k & e^{330} \cdot \tilde{Q}_k & e^{-210} \cdot \tilde{Q}_k & *
\end{pmatrix} \quad (3.13)$$

In the simplest case the nullmodel \tilde{Q}_k is a Jukes-Cantor Model. After the normalisation of Q_k^{stack} , the submatrix is given by:

$$\begin{array}{c}
A|GAUGC \\
C|GAUGC \\
G|GAUGC \\
U|GAUGC
\end{array}
\begin{array}{c}
C|GAUGC \\
G|GAUGC \\
U|GAUGC
\end{array}
\begin{array}{c}
G|GAUGC \\
U|GAUGC
\end{array}
\begin{array}{c}
U|GAUGC
\end{array}
\begin{array}{c}
A|GAUGC \\
C|GAUGC \\
G|GAUGC \\
U|GAUGC
\end{array}
\begin{pmatrix}
-16.889 & 16.807 & 0.003 & 0.079 \\
0.000 & -0.000 & 0.000 & 0.000 \\
0.003 & 16.807 & -16.889 & 0.079 \\
0.000 & 0.557 & 0.000 & -0.557
\end{pmatrix} \quad (3.14)$$

Although it seems that we have 0's in our matrix, it is based on the fact that the numbers are rounded. However, we have a uncoupled Markov chain, which is a discrete chain whose matrix is almost block diagonal (Deufjhard *et al.*, 2000) . Thus, the eigendecomposition of the whole instantaneous rate matrix Q_k^{stack} is not feasible. However, it is possible to analyse different components of the matrix individually. We calculate the eigendecomposition of each submatrix using the Equ. 2.13. Thus, the eigenvector corresponding to the eigenvalue 0 of our submatrix (Equ. 3.14) is,

A GAUGC	C GAUGC	G GAUGC	U GAUGC
$-2.4531e - 08$	-1	$-2.4531e - 08$	$-2.4531e - 08$



For example, this eigenvector shows that with the neighbourhood system of $s_k = \{x_k|GAUGC\}$ the current site is mostly a cytosine with a C-G Watson-Crick pair, although U-G can bind as a wobble pair. Tab. 3.6 shows further examples of submatrices. In the first submatrix

we have changed the first base-pair from A-U (example above) to G-C resulting in a almost entirely absorbing state for cytosine at the current position. Likewise, the following

	A GGCGC	C GGCGC	G GGCGC	U GGCGC	
A GGCGC	-118.127	117.782	0.003	0.342	
C GGCGC	0.000	-0.000	0.000	0.000	
G GGCGC	0.003	117.782	-118.127	0.342	
U GGCGC	0.000	0.906	0.000	-0.906	
	0	1	0	0	

	A GGUGU	C GGUGU	G GGUGU	U GGUGU	
A GGUGU	-2.414	2.398	0.003	0.013	
C GGUGU	0.000	-0.000	0.000	0.000	
G GGUGU	0.003	2.398	-2.414	0.013	
U GGUGU	0.001	0.473	0.000	-0.474	
	0	1	0	0	

	A GAAGC	C GAAGC	G GAAGC	U GAAGC	
A GAAGC	-0.589	0.556	0.003	0.030	
C GAAGC	0.000	-0.000	0.000	0.000	
G GAAGC	0.003	0.556	-0.589	0.030	
U GAAGC	0.000	0.050	0.000	-0.050	
	$2.2284e - 05$	0.9971	$2.2284e - 05$	$2.2284e - 05$	

	A AAUAU	C AAUAU	G AAUAU	U AAUAU	
A AAUAU	-0.054	0.003	0.003	0.048	
C AAUAU	0.003	-0.054	0.003	0.048	
G AAUAU	0.003	0.003	-0.054	0.048	
U AAUAU	0.000	0.000	0.000	-0.000	
	0.003	0.003	0.003	0.991	

	A AAAAA	C AAAAA	G AAAAA	U AAAAA	
A AAAAA	-0.009	0.003	0.003	0.003	
C AAAAA	0.003	-0.009	0.003	0.003	
G AAAAA	0.003	0.003	-0.009	0.003	
U AAAAA	0.003	0.003	0.003	-0.009	
	0.25	0.25	0.25	0.25	

Table 3.6: Further examples of submatrices of Q_k^{stack} with the eigenvector corresponding to the eigenvalue 0. The eigendecomposition of the whole instantaneous rate matrix Q_k^{stack} is not feasible. However, it is possible to analyse different components of the matrix individually.

submatrix, where two wobble pairs are adjacent to the current position. However, with an adjacent G-C pair and one not Watson-Crick pair the probability for cytosine is high, but that state is not absorbing. It is not surprising that two A-U adjacent base-pairs and an adenosine at the opposite of the current position are resulting in a high rate value for uracil, however, not so high as in the examples before. If no base-pair is adjacent to the current position the numbers of the eigenvector for all four possible nucleotides is the same.

Summing up, the different equilibrium distributions in Tab. 3.6 indicate that conventional joint modeling by simply changing partially paired positions (Subsec. 3.2.1) is not capturing some important biological information. We have suggested a model which takes the energy into account and can be combined to different independent and dependent null models for further analysis. Furthermore, a realistic phylogenetic energy model needs to consider the *loops* in a structure (subsection 2.1.2).

3.4 SISSI with Indels

Long-term evolution often includes dynamic changes such as insertion and deletion. ROSE was the first simulation program with Indels (Stoye *et al.*, 1998). However, the process with which the indels are created is not strictly model based. Today, several other sequence simulators including indels exist: EvolveAGene (Hall, 2005), SIMPROT (Pang *et al.*, 2005), MySSP (Rosenberg, 2005), SIMULATOR (Fleißner, 2004), DAWG (Cartwright, 2005) and indel-Seq-Gen (Strope *et al.*, 2007). None of these programs take site-specific interactions into account. For simulating sequence evolution with a well-defined insertion-deletion dynamics with site-specific interactions, the following points must be considered:

- (i) A more general insertion/deletion model is necessary.
- (ii) To take the states of the neighbourhood system into account, each indel step on each branch must be known.

So far, we have worked on such an extension of our algorithm based on indel-process described by (Metzler, 2003).

Insertion-deletion Models

A number of models has been described which try to provide reasonable insertion-deletion dynamics with computational tractability (Bishop and Thompson, 1986; Thorne *et al.*, 1991, 1992; Miklós and Toroczka, 2001; Metzler, 2003), where the TKF1 model by Thorne *et al.* (1991) has received greater attention. The TKF1 model allows insertions and deletions of one single nucleotide at a time. Since this seems unrealistic, Thorne *et al.* (1992) have extended the TKF1 model to the TKF2 model, which describes the insertion and deletion process of longer fragments. But a fragment that has once been inserted can

Algorithm 3.4.1: Evolving a sequence \mathbf{x} with a indel process, recursively through a given rooted or unrooted tree topology, where the branch lengths are specified by the expected number of substitution d . Please, refer to Tab. 3.7 for footnotes (\dagger, \ddagger).

I. Compute a sequence $\tilde{\mathbf{x}}(start)$ of a desired length $l + 2m$ with sufficiently large m ;

while $Tree(*)$ **do**

 Computing a sequence $\tilde{\mathbf{x}}(d)$, d substitutions per site away from $\tilde{\mathbf{x}}(0)$;

Indel-process;

 1. Init $d_{indel_0} = 0$;

 2. Draw d_{ind} from the exponential distribution with parameter 2μ . (\dagger) ;

 3. Init $d_{indel} = d_{ind}$;

 4. **while** $d_{indel} < d$ **do**

Substitution-process;

 a. Compute $q(\tilde{\mathbf{x}})$ for $\tilde{\mathbf{x}}(d_{indel_0})$;

 b. Draw d_r from the exponential distribution with parameter $q(\tilde{\mathbf{x}})$;

 c. Init $d_c = d_r$;

while $d_c < d_{indel}$ **do**

 1. Choose a site \tilde{k} with $\mathbb{P}(\tilde{k})$ (see Eq. 3.5) ;

 2. Replace $\tilde{\mathbf{x}}_{\tilde{k}}$ with \mathbf{y}_0 with $\mathbb{P}(\tilde{x}_{\tilde{k}} \rightarrow y_0)$ (see Eq. 3.6) ;

 3. Update $q(\tilde{\mathbf{x}})$ based on N_k ;

 4. Draw new d_r from the exponential distribution with parameter $q(\tilde{\mathbf{x}})$;

 5. Set $d_c = d_r + d_c$;

end

 d. Choose a site \tilde{k} with probability $1/(l + 2m)$. (\ddagger_1) ;

 e. Choose insertion or deletion with probability(deletion)= $1/2$. (\dagger) ;

 f. Draw a number of inserted or deleted sites from the geometrically distribution with parameter ζ . (\ddagger_2) ;

 g. If Deletion, ending with selected one, else inserted according the stationary distribution Q_k to the right of the selected.;

 h. Set $d_{indel_0} = d_{indel}$;

 i. Draw new d_{ind} from the exponential distribution with parameter 2μ .;

 l. Set $d_{indel} = d_{ind} + d_{indel}$

end

end

II. From site $k + 1$ of sequence $\tilde{\mathbf{x}}(start)$ we move left and from position $m + l + 1$ we move right until we find the positions that are all homologous to positions in a column.;

III. With cutting out the left and right corresponding positions we got an alignment with the approximate length l of an evolved sequence \mathbf{x} .

-
- (†) If the insertion rate is not equal to the deletion rate, we have to draw d_{ind} from the exponential distribution with the parameter $\lambda + \mu$ and choose an insertion with the probability $\lambda/(\lambda + \mu)$.
 - (‡) 1. With $n_{\tilde{k}} > 0$ the probability to choose a site \tilde{k} for the deletion - insertion process should depends on the neighbourhood system.
 - 2. Furthermore, the number of sites, which will inserted or deleted depends on the neighbourhood system. In an easy dependency case with no overlapping, we could draw the number of inserted or deleted sites from the geometrically distribution with parameter $(n_k + 1)\zeta$. Note, that we can choose any distribution.
-

Table 3.7: Alg. 3.4.1 presents the basic algorithm for simulations with a well-defined insertion-deletion dynamics as well as site-specific interactions. However, in general their addition of pose further problems correlated to the indel creation itself like the distribution of indel lengths or where to place the indels.

only be deleted as a whole, and no other fragments can be inserted in between it. This is necessary to obtain a pair-HMM (Hidden Markov Model) structure on the alignment. A program such as ALIFRITZ for simultaneous estimation of statistical multiple alignment and phylogeny reconstruction (Fleißner *et al.*, 2005), is based on HMM, but this is not necessary for our simulation program. The TKF2 model has three parameters, the fragment length and the insertion and deletion parameters with deletion greater than insertion rate. Metzler (2003) gives a model of fragment insertions and deletion (FID) with two parameters: greater or equal to 1, the expected fragment length, and λ greater or equal to zero, the indel rate per site with deletion equal to insertion. Here, each site is a 'fragment end' independently with probability $1/\zeta$ and each is selected at rate 2λ . With probability $1/2$ the fragment is deleted, otherwise a new fragment is inserted to its right. The length of the new fragment is geometrically distributed with expectation ζ .

A more general insertion deletion (GID) model (Metzler, 2003) is like FID, but without fixed fragmentation. Set $\mu := \lambda/\zeta$. Thus, each site is selected with rate 2μ . This insertion-deletion dynamics can be combined with the Markov model that describes the substitution process with arbitrary site-specific interactions.

SISSI-Algorithm Including an Indel Process

We follow our notations so far with two additional parameters, the indel rate λ and the fragmentation length ζ .

Without insertion-deletion dynamics, we have to simulate sequences $(\mathbf{x}(1), \mathbf{x}(2))$ with length l and sites $k = 1, \dots, l$ recursively through a given rooted or unrooted tree topology. Now, for the insertion-deletion dynamics, we generate at the beginning a sequence $\tilde{\mathbf{x}}(\text{start})$ of length $l + 2m$ with sufficiently large m . E.g. the probability that the fragment

length equals a is $(1 - \zeta^{-1})^{a-1}\zeta^{-1}$, such that m must be greater than a . Accordingly, we call the sites of the sequence $\tilde{\mathbf{x}}_k$. The states of the sites are picked according to the equilibrium distribution, respectively an ancestral sequence of length l plus $2m$ sites in the equilibrium distribution left and right. The pair $(\tilde{\mathbf{x}}(1), \tilde{\mathbf{x}}(2))$ has to be simulated according to the indel process. With time the sequences end in different sequence lengths. After the evolution of sequence $\tilde{\mathbf{x}}(start)$ through the whole tree topology we move from site $k + 1$ left and from position $m + l + 1$ right until we find the positions that are all homologous in one column. With cutting out the left and right corresponding positions we get an alignment with the approximate length l of an evolved sequence \mathbf{x} .

Thus, we have presented the basic algorithm for simulations with a well-defined insertion-deletion dynamics as well as site-specific interactions. However, in general the addition of insertion-deletion dynamics pose further problems, especially the indel creation itself, for example, the distribution of indel lengths or where to place the indels, see Tab. 3.7. To solve these problems the expertise of structural biologists is necessary in future research.

Discussion

We have introduced a general framework taking site-specific interactions into account, and we can mimic sequence evolution with various complex dependencies among sites. The basic idea is the application of different substitution matrices for each site defined by the interactions with other sites in the sequence. Our implementation, **SISSI** (**S**imulating **S**ite-**S**pecific **I**nteractions), allows the evolution of nucleotide sequences along a tree for user defined systems of neighbourhoods and instantaneous rate matrices.

Simulations have shown that **SISSI** produces sequences under constrained evolution in reasonable time. While for simulations with independent sites Seq-Gen (Rambaut and Grassly, 1997) should be used because it is more time-efficient, we have shown that running time is not really an issue for our general approach. Thus, it should be possible to generate large simulated datasets that may be used to analyse the reliability of tree reconstruction methods under deviations from the independent site assumptions.

Recently, models with site-specific rate matrices have also been studied more frequently, as well as various mixture models (e.g. Koshi and Goldstein, 1995; Bruno, 1996; Thorne *et al.*, 1996; Koshi and Goldstein, 1997; Halpern and Bruno, 1998; Goldman *et al.*, 1998; Lartillot and Philippe, 2004; Pagel and Meade, 2004). Furthermore, a number of models of protein evolution have been developed to account for protein structure by accepting randomly generated mutations if they do not affect the structure too much (e.g. Parisi and Echave, 2001, 2005). With our method, allowing the specification of site-specific rate matrices or different rate matrices for different regions of the simulated sequence is straightforward.

Moreover, our framework allows the introduction of mechanistic parameters, thus making all model assumptions explicit. **SISSI**'s framework is flexible enough to introduce simpler

models that capture the major features of interactions between sites and need fewer parameters. Therefore, it is possibly more insightful to confine simulations to the relevant parameters. The extended **SISSI** framework combines traditional phylogenetic models including site-specific interactions with an arbitrary complexity.

The example given here illustrates the basic principle. However, introducing more and more realistic features to model the evolutionary process requires the specification of a large number of parameters. This does not pose a problem for simulations, the user simply has to define everything. It could be argued that these simulations are irrelevant for the reconstruction process. However, sometimes, it is useful to simulate with more complex models than those used for estimates, e.g. of phylogenetic trees. For example, simulation under a complex site-specific model and estimation with a covarion model could be very interesting for future structure evolution methods with phylogeny.

For RNA, we can readily map existing doublet models into our framework, as well as complex overlapping dependencies, which allow to take energy values into account. In addition, it is possible to combine a neighbourhood parameter which is a function of energy values with chosen traditional independent or dependency models. So far, we have solved the problem algorithmically and implemented the first important steps. In the near future the extended implementation will provide a necessary framework. The accuracy of the nearest neighbour model to compute secondary structure is very high for short RNAs (length smaller than one hundred nucleotides), but insufficient for large RNAs (e.g. SSU or LSU ribosomal RNAs). One reason is the inaccuracy of the experimentally measured thermodynamic parameters, another one is that the fold mutates through the tightly packed components of RNA and proteins. Thus, this framework has the potential to analyse this in further research throughout simulations as well as further analytical results of the model.

In general, the inclusion of energy values is a challenging extension of our model in the process of RNA evolution. This would add another realistic feature and therefore the evolutionary path through sequence space guided by the tree is more easily comparable to results produced by *RNAinverse* (Hofacker *et al.*, 1994b). *RNAinverse* searches for all sequences folding into a predefined structure, but takes no phylogenetic relationship into account. In contrast, **SISSI** mimics sequence evolution under structural constraints and it is likely to obtain sequences in the course of evolution that deviates temporarily from the neighbourhood system.

A further important issue will be modelling tertiary interactions. Finally, **SISSI** will be an appropriate tool because it is flexible enough to model energy values with other constraints together, e.g. by interactions with other molecules.

Besides applications in phylogenetic inference, simulated data sets with dependencies can be used to test structure analysis methods, e.g. RNA structure prediction. Another application of **SISSI** is the systematic study of the influence of phylogenetic relationships among

the sequences that are subject to structure prediction. Thus, **SISSI** illustrates the evolutionary path with compensatory mutation along the tree, e.g. with programs that detect nucleotide interactions. As a consequence this may result in intermediate structure that may show a large deviation from the structure defined by the neighbourhood system. If a huge fraction of closely related sequences happen to deviate by chance from the underlying structure, this will mislead structure prediction programs, which do not account for the phylogenetic relationship in a proper way. **SISSI** may help to address this particular issue, to distinguish structural (functional) from phylogenetic (ancestral) correlations. This is discussed in more detail in the next two chapters.

Furthermore, it is not necessary to restrict the simulation on one neighbourhood system. It is very well possible to define different neighbourhood systems for different regions of the tree. Such simulations may be the basis for studies about structure evolution. One of the most interesting points: "How does a *structure* evolve?"

In a forward looking perspective, a **SISSI** approach could be used to detect structure evolution and to distinguish this from ambiguously aligned regions in MSA. The current work on including the process of deletion and insertion in the **SISSI** simulation framework is one of the first steps towards the understanding of divergence mechanisms of families through evolution.

3.5 The Sublime by **SISSI**

SISSI presents a universal description for arbitrary complex neighbourhood systems in a unifying framework. The most important point from our viewpoint is to be flexible in terms of simplicity and complexity and considering the laws of simplicity, for example, in the sense of all environment perceptions or methods, comparable to Maeda (2006) in the field of design.

The **SISSI** framework is applicable to other inter- and intra site-specific interactions among nucleotides and other character-based sequences, like amino acids, codons or discrete character states from the biological viewpoint as well as other fields. However, as Strope *et al.* (2007) recently mentioned from the viewpoints of proteins, we need information on residue interaction in greater detail than we currently have. This also holds true for other potential possibilities of our arbitrary complex framework like RNA-RNA interactions, as well as RNA-proteins interactions; other character based interactions, like gene interactions and interaction networks or structure motifs in general.

To address these points, we need a real combination between the apparent different and disjoint fields of biomolecular structure prediction and phylogeny. Thus, simulations employing a neighbourhood system, incorporating artificial or known structural features needs collaborations with biological experts from the lab, as well as a clear theoretical definition of structure in general.

Chapter 4



A. Artaud (1896-1948)

A Phylogenetic Definition of Structure

Imagination, to Artaud, was reality; he considered dreams, thoughts and delusions as no less real than the "outside" world ... a theatre to see a play and, for a time, pretend that what they are seeing is real ...

TO BERND FEDOR

The necessity for a phylogenetic definition of structure can be illustrated by the ambiguous definition of RNA families. Indeed, the actual state of the Rfam data base indicates that different researchers seem to use different concepts for RNA families (see Chap. 2). Discussing this in full depth here would be out of the scope of this thesis, but we want to contribute to the basis of structure definitions and phylogeny. In Chap. 2, we have given the commonly used structure definitions, like primary, secondary, tertiary structure, minimum free energy, suboptimal structures, consensus structure, and others.

Although there are already methods that use phylogeny and structure an explicit definition from a phylogenetic viewpoint is missing. In the previous chapter, we introduced SISSI (Simulating Site-Specific Interactions). Based on the concept of a neighbourhood system, SISSI simulates the evolution of a nucleotide sequence along a phylogenetic tree taking into account site-specific interactions. Thus, in the SISSI-framework our definition of structure from a phylogenetic viewpoint is already implicit.

In this chapter, we give an explicit phylogenetic definition of structure. This could be directly fed into additions or improvements of existing structure prediction programs from alignments, as will be considered in Chap.5. Here, we refrain from most technicalities but rather outline the general ideas by discussing illustrative example. Finally, structure definitions at different abstract levels are discussed.

4.1 A Phylogenetic Definition of Structure

Our phylogenetic definition of structure consists of three aspects: The substitution model, a neighbourhood system and the phylogenetic tree. The substitution model specifies the evolutionary process of nucleotide evolution. However, the model is influenced by the neighbourhood system that defines the interactions among sites in a sequence. The phylogenetic tree introduces an additional dependency pattern in the observed sequences.

Def.: A *phylogenetic structure (PS)* is an abstract object which is defined by a neighbourhood system, a substitution model and a phylogenetic tree.

In a first step the three aspects are defined as:

I A neighbourhood system \mathcal{N} as in 3.1.1

II A substitution model constitutes a collection of possibly different substitutions processes acting on the sequence and an annotation of site-specific interactions among sites as described in 3.1.2.

III A phylogenetic tree T as defined in 2.2.3.

A PS appears in a set of sequences at different time points. Those can be transformed into *objects*, which can be described, e.g. as planar graphs (2.1.2). Transformation rules are already given, e.g. with the so-called minimum free energy structure or other concepts of commonly used definitions of chapter 2.

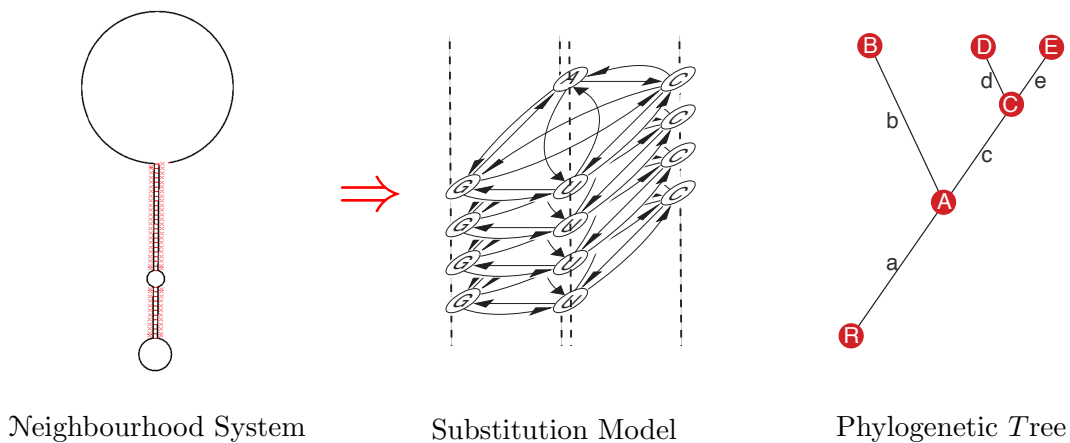


Figure 4.1: An example of a phylogenetic structure. **Left:** an example of a neighbourhood system, which is thermodynamically improbable, was chosen for didactical reasons. **Middle:** the model Q constitutes a collection of possibly different substitution models with the parameter set of Tab. 4.1. **Right:** example of a phylogenetic tree with three extant taxa.

4.2 Realisations of Phylogenetic Structure

A realisation of a PS is a relational object at time point t . The following is an illustration of what we have just described, taking the example of a phylogenetic structure of Fig. 4.1 in correlation to a minimum free energy (mfe) structure. For didactical reasons, we have used a neighbourhood system, which is thermodynamically improbable. However, this thermodynamically artificial neighbourhood system influences the collection of different substitution models acting on the sequence. This collection is defined in one model (see SISSI framework in subsection 3.1.2). As part of this concept, it is possible to mimic sequence evolution under the structural constraint of the PS. However, due to the stochastic nature of the substitution process it is likely to observe sequences in the course of evolution that exhibit - at least temporarily - predicted structures, that deviate to some extent from the neighbourhood system. Fig. 4.2 shows a predicted mfe-structure of a generated sequence as one realisation of the phylogenetic structure (Fig. 4.1) at timepoint 0.42 on branch a . In comparison to the neighbourhood system of the PS, the long helix of the neighbourhood system maps well into the helix of the realisations. In contrast the upper independent part is folded due to the thermodynamical impossibility of such a long loop. *In a nutshell:*

although the neighbourhood system is thermodynamically artificial, it is transformed in possible thermodynamic realisations given an evolutionary history.

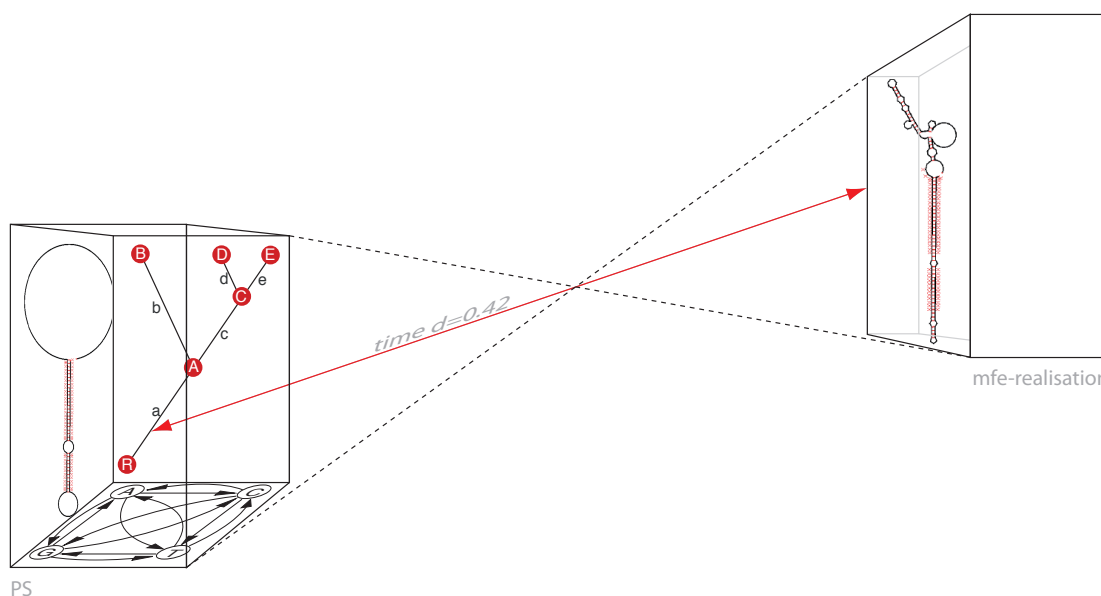


Figure 4.2: The Phylogenetic Structure (PS) of Fig. 4.1 is summarized on the left and is transformed at timepoint $d = 0.42$ in a mfe realisation and suboptimal mfe realisations on the right. By comparing the realisation with the neighbourhood system most base pairs are the same. However, the upper independent part is folded due to the thermodynamical improbability of such a long loop.

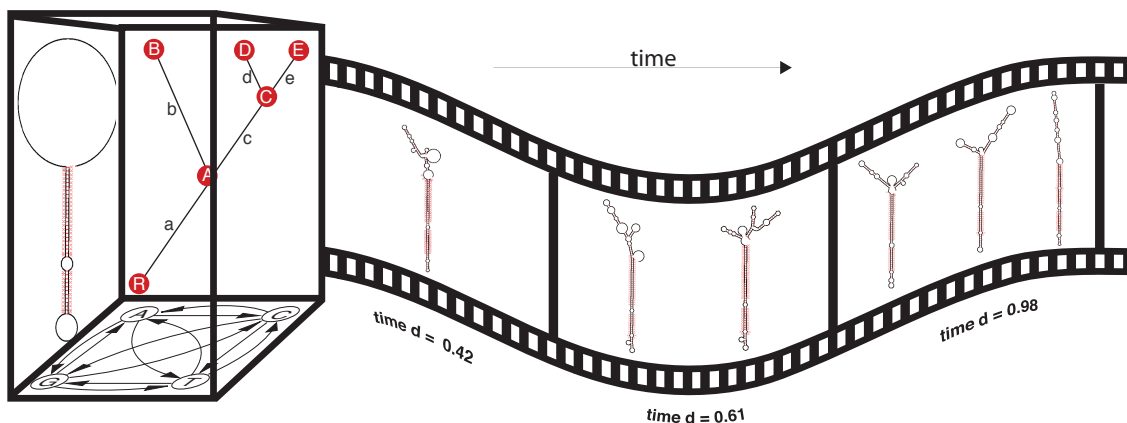


Figure 4.3: A PS film about the diversity of thermodynamic realisations, inferred as minimum free energy structures using RNAfold of the PS given in Fig. 4.1. Three frames of the diversity of thermodynamic realisations at three different time points are shown, scaled in the number of substitutions. More details of the film are described in the text and Fig. 4.4.

Here, we simply set the parameters: by counting observed doublets defined by a potential base pair of the secondary structure of the stem region of 60 metazoan small subunit ribosomal RNA (SSU rRNA) sequences (Nefes *et al.*, 1993). Thus, here all dependent sites have the cardinality one ($n_k = 1$) and the same symmetric matrix of Tab. 4.1. For the independent sites ($n_k = 0$) we use the marginal distribution of Tab. 4.1. We start at mutational time $d(0)$ with a sequence $\mathbf{x}(0)$, which evolves according to (N, Q) along a phylogenetic tree. For our simple example, we have chosen a phylogenetic tree with three “extant” taxa and five nodes. To explain our idea further, we consider the *diversity of thermodynamic realisations*. In doing this, we have to take into account the relationships of the sequences included in the evolutionary history.

Fig. 4.4 is a part of a film (Fig. 4.3) to illustrate the diversity of realisations of the PS. The first picture at the bottom of Fig. 4.4 starts at time 0.80, 0 substitutions after a speciation event. We show three frames depicting different time points after the speciation event from species C. While at timepoint 0.8 the mfe realisations of species D and E are the same and differ from B, the species start to diversify and after 0.01 substitutions also D and E are different, but still show similarities. In more detail and compared to the annotation of the neighbourhood system of the PS from Fig. 4.1: the stem region of the realisation should be similar defined by the constraint through the neighbourhood system and the substitution model, while the upper loop region has no site-specific interactions and should result in different realisations through the mfe folding. However, we see in the upper part similar mfe folds between D and E at timepoint 0.81, based on the previous frame. Moving further, more time has elapsed and the differences between D and E become more apparent. In the uppermost frame, the species B and D have more structural similarities than either has with E, although D and E are closer related.

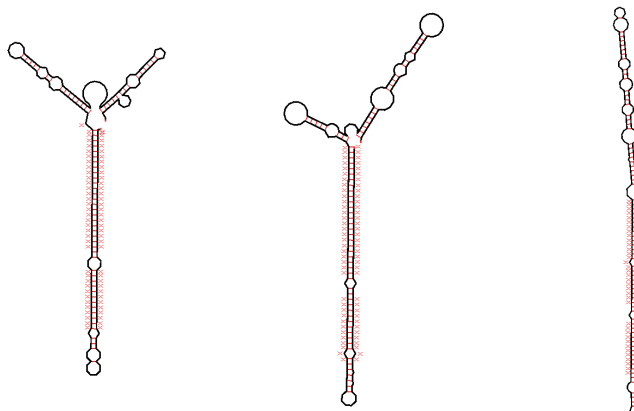
time d/taxa

ⓑ

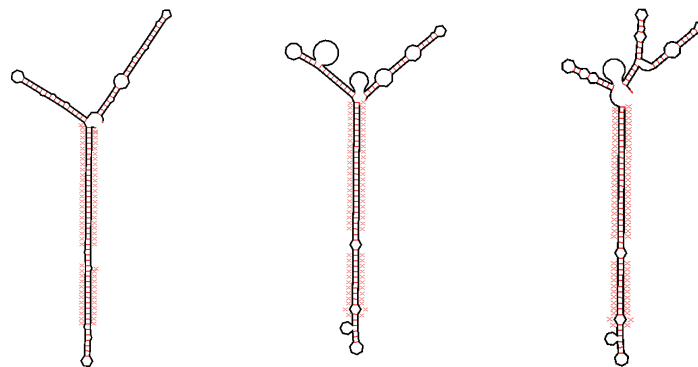
ⓓ

ⓔ

(3) time $d = 0.98$



(2) time $d = 0.81$



(1) time $d = 0.80$

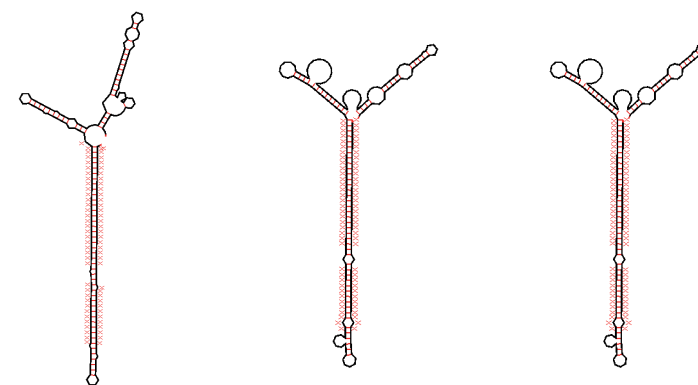


Figure 4.4: Part of a PS film (Fig. 4.3) about the diversity of thermodynamic realisations, inferred as minimum free energy structures using RNAfold of the PS given in Fig. 4.1. Here, you see the diversity of thermodynamic realisations at three different time points, scaled in the number of substitutions. More details of the film are described in the text.

Summarising Fig. 4.4 we see two important points. First: short time after a speciation event, we have similarities that do not result from the neighbourhood system. Second: after a longer period, taking the similarity between different species into considerations, it is possible to get similarities by chance, independently of the phylogenetic relationships. This could be a false signal for phylogenetic programs based on descriptions of realisations, while the first point could result in false signals for the structure prediction programs from alignments. These two aspects depend on the stability of the structure, the evolutionary constraint and the chosen realisation method. This will be discussed in the following section.

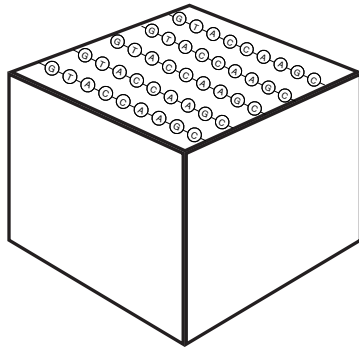
4.3 Structural Constraints

Following a PS, we have to distinguish between different constraints for observed interactions among the sites, which we summarise as structural constraints, illustrated in Fig. 4.5. A structural constraint defines the evolutionary strength of structuring sequences at different timepoints t and can be distinguished as: ancestral constraint (Fig. 4.5E) and neighbourhood constraint (Fig. 4.5F). The neighbourhood constraint are site-specific interactions acting on the sequence along the evolutionary process, in our framework defined by an annotation of a neighbourhood system. In this sense, the observable interactions between the sites through this neighbourhood constraint are called neighbourhood or functional correlation (Fig. 4.5D). Furthermore, following our phylogenetic definition of structure, we have to take into account the evolution of nucleotides. The states at the internal nodes of the phylogeny are important because of the likelihood of the state remaining unchanged after only a short period of time. This depends on the model and we referred to it as an ancestral constraint which defines the influence of ancestral nucleotide distribution at an alignment site and can be associated with observable ancestral correlations in sequences (Fig. 4.5C). Generally, we use the term associations, correlations or dependencies to represent measurements from sequences via different estimation methods. If we estimate correlations from homologous sequence data, e.g from an alignment (Fig. 4.5A), they are related through their evolutionary history and common ancestral states. Thus, ancestral as well as functional correlations can occur. However, we should ask, how can we distinguish ancestral from functional correlation, when we want to infer functional correlations (Fig. 4.5B). So far, structure prediction methods are mostly interested in predicting dependencies that result from neighbourhood constraints. In these prediction programs, ancestral correlations represent false positive predictions and should be avoided, while functional correlations represent true positives.

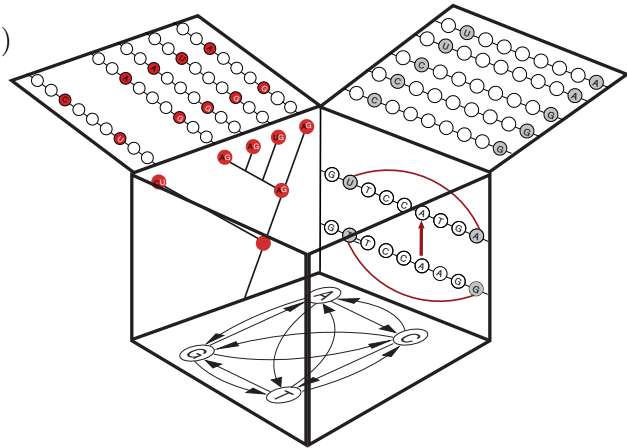
Summarising: A structural constraint defines the evolutionary strength of structuring sequences at different timepoints t and can be distinguished as:

- Neighbourhood constraint, these are site-specific interactions imposed on the sequence during evolution. The observable interactions between the sites through this

(A) Alignment



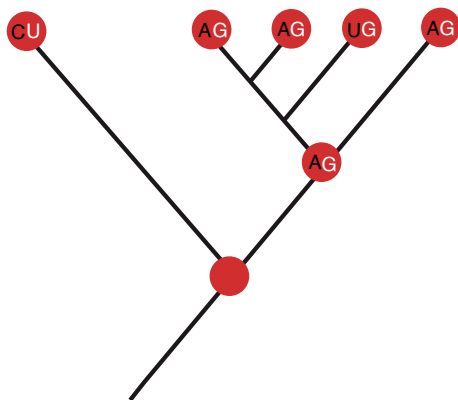
(B)



Associations

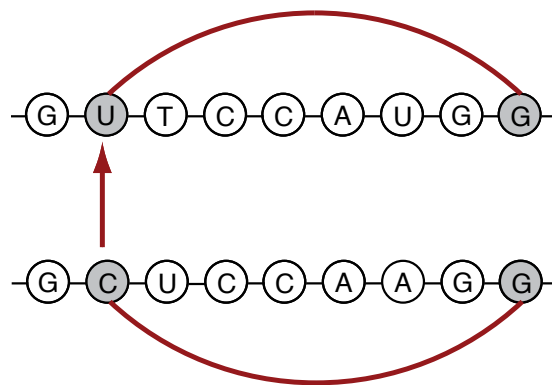
(C)

Ancestral Correlation



(D)

Functional Correlation



(E)

Ancestral Constraints

(F)

Neighbourhood Constraints

Constraints

Figure 4.5: Following our initial definition, we have to distinguish between different structural constraints: Ancestral constraints (E) and neighbourhood constraints (F). Both constraints depend on the Model Q (B), which determines the strength of evolution. The ancestral constraints can be associated with the ancestral correlation between the sites in observed sequences today (C, B). The neighbourhood constraint determines site-specific interactions in the evolutionary process associated from an alignment as so-called functional correlations or dependencies (D,B). However, these inter-site associations are difficult to distinguish, when we want to estimate correlations from homologous sequence data, e.g from an alignment (A). Note that all of these aspects are dependent on one another.

neighbourhood constraint are called neighbourhood or functional correlation.

- Ancestral constraint, this is the influence of the states at the internal nodes. It is very likely that these states remain unchanged after short evolutionary time span and can be associated with correlations in the sequence, so-called ancestral correlation.
- Both constraints are not independent of each other and depend on the model Q .

Both constraints, the ancestral and neighbourhood constraint, are not independent of each other and depend on the Model Q . How can we compare these substitution models? A simple idea is to compute the χ^2 -value of the frequencies. Tab. 4.1 and Tab. 4.2 give some examples for the following simulations to illustrate the concepts, although it is clear that the χ^2 -value is not a sufficient measurement.

Example I: Parameter

(A)

	$f_{n_k=1}$				$f_{n_k=0}$
	second nucleotide in doublet				
first	A	C	G	U	
A	0.0030	0.0049	0.0042	0.1539	0.166
C	0.0049	0.0035	0.2508	0.0032	0.262
G	0.0042	0.2508	0.0018	0.0762	0.334
U	0.1539	0.0032	0.0762	0.0052	0.239
χ^2	$\chi_r^2 = 1.96$				$\chi_e^2 = 1.79$

(B) Permutation

	$f_{n_k=1}$				$f_{n_k=0}$
	second nucleotide in doublet				
first	A	C	G	U	
A	0.0049	0.0035	0.2508	0.0032	0.262
C	0.0030	0.0049	0.0042	0.1539	0.166
G	0.0042	0.2508	0.0018	0.0762	0.334
U	0.1539	0.0032	0.0762	0.0052	0.239
χ^2	$\chi_r^2 = 1.96$				$\chi_e^2 = 1.79$

(C)

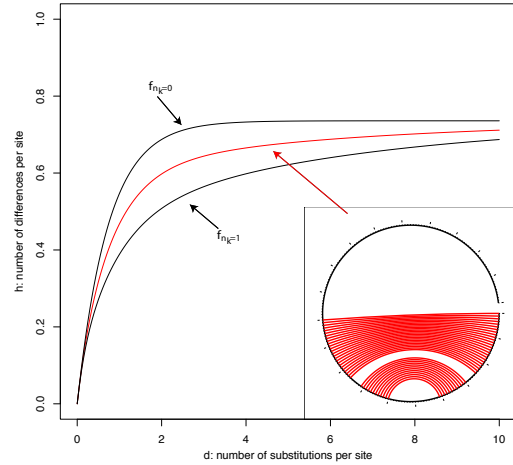


Table 4.1: Example I: **A**: parameters for the model of our example of a phylogenetic definition of structure in Fig 4.1. The frequencies are simply set by counting observed doublets (cardinality $n_k=1$, see Sec. 3.1.1) of 60 metazoan small subunit ribosomal RNA sequences (Nefes *et al.*, 1993). The marginal mononucleotide composition ($n_k = 0$) is the sum of the dinucleotide ones.

χ_r^2 value is computed assuming a uniform distribution of the doublet frequencies. χ_e^2 is the value if the expected dinucleotide composition is derived from $f_{n_k=0}$. **B**: Permutation Example: the first two rows of (A) are exchanged. Note, that A and B give the same χ^2 values. **C**: analytical calculation of the relationship between the number of substitutions d per site and number of observed differences h per site, with the corresponding frequencies of A and B using Equ. 2.25 and 2.26. The red line is the average number of substitutions per site for the sequence and the neighbourhood system (red circle plot) of our PS example in Fig 4.1.

4.3.1 Simulation Studies

To illustrate the interweavement of the three aspects of a PS we simulated with SISSI, employing neighbourhood systems, incorporating artificial and known structural features, as well as different substitution models and different trees. The focus of attention will be the so-called consensus structures (see Chap. 2). Here, we compare the mutual information context (MIC) with thermodynamic methods (TH) as an example (see Sec. 2.3.2). Thus, we use on one side the information in the form of compensatory base pair substitutions and on the other side information based on the physical properties of single sequences.

Usually, structure prediction focuses on the mean pairwise identity (MPI). However, in phylogeny the focus of attention is on the number of substitutions per site in correlation with the Hamming distance h or the observed number of differences per site. The MPI

Example II: Parameter

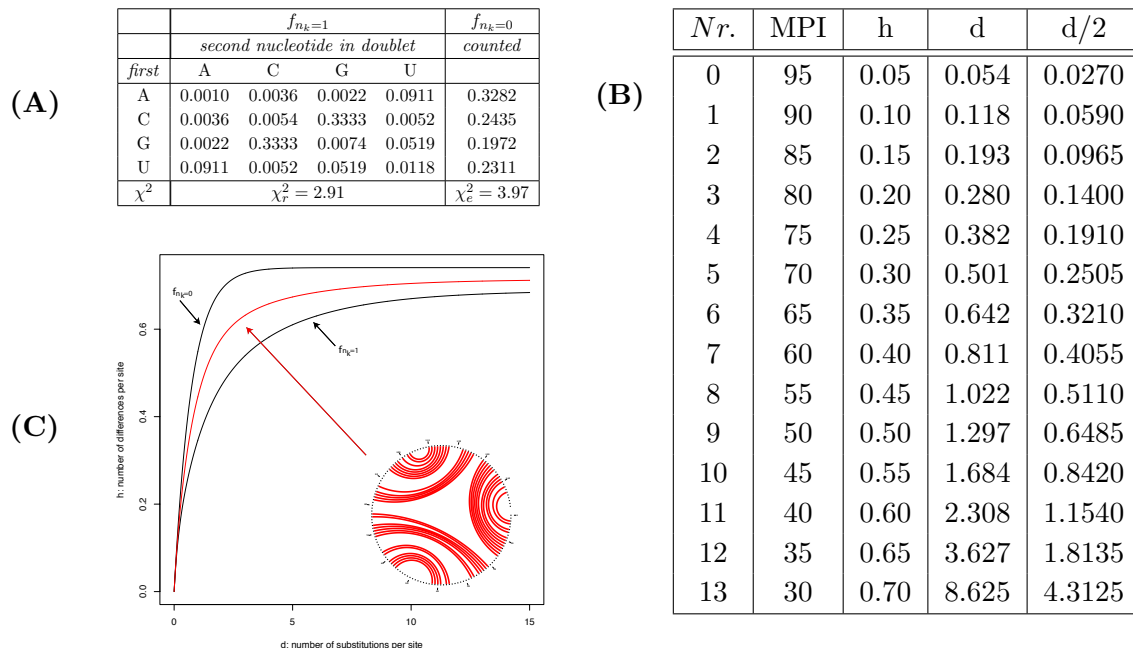


Table 4.2: Example II: parameter of the archeobacteria 5S rRNA. **A:** we have counted the observed frequencies for doublet frequencies ($f_{n_k} = 1$) and single frequencies ($f_{n_k} = 0$) along 122 nucleotides from the corresponding alignments with 85 sequence and along the corresponding structure of the 5S ribosomal RNA data bank (Szymanski *et al.*, 2000). See Tab. 4.1 for the description of the χ^2 values. **B:** computing of the corresponding numbers of substitutions for the mean branch length to get the average mean pairwise identity (MPI) using the corresponding Eq. 2.25 and 2.26. **C:** calculation of the relationship between the number of substitutions d per site and number of observed differences h per site, with the corresponding frequencies using Eq. 2.25 and 2.26. The red line is weighted with the neighbourhood system of the archeobacteria 5S rRNA (red circle plot).

Simulation Study along Star Trees

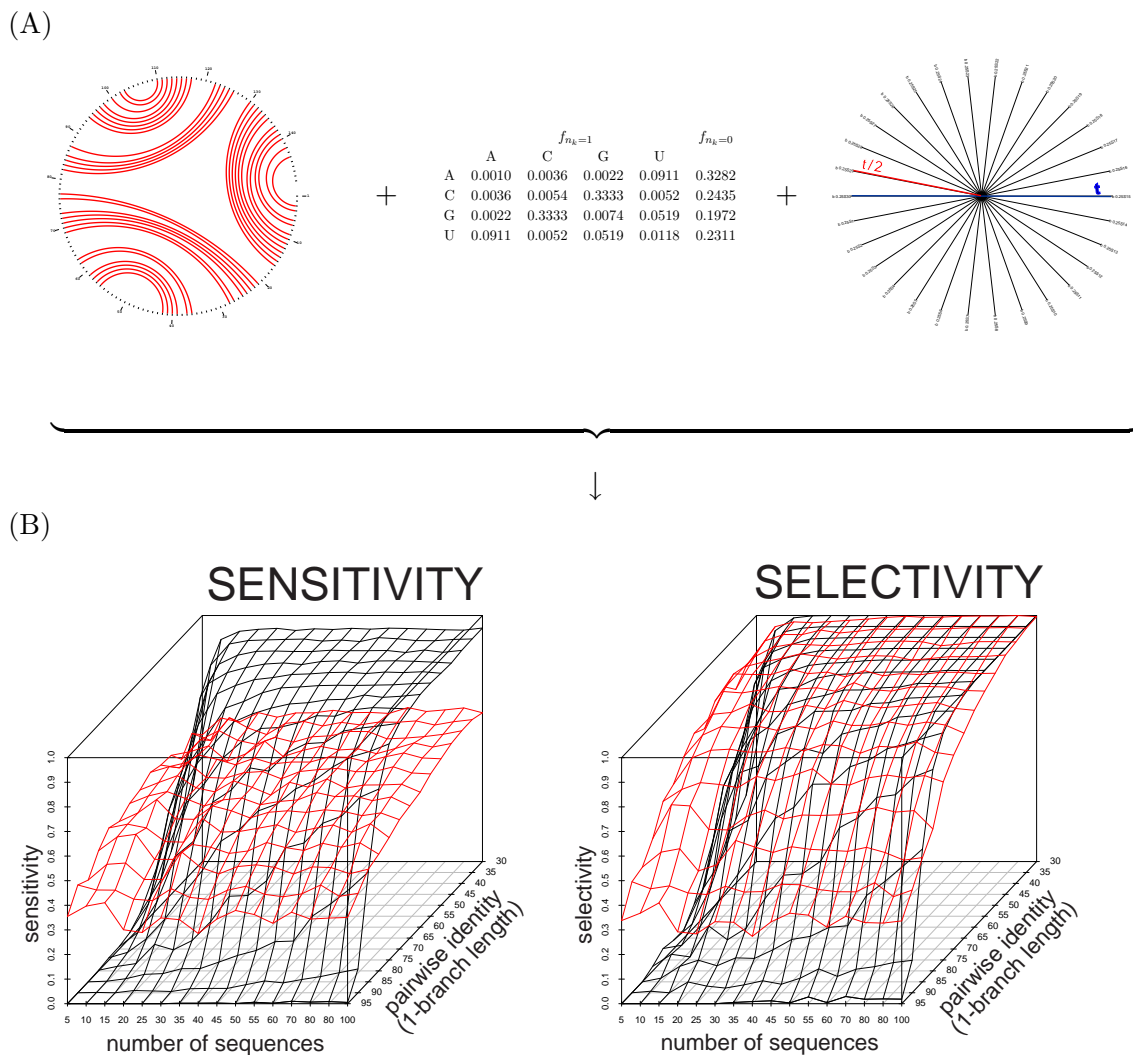


Figure 4.6: Simulation Study: **A:** Input of SISSI for the simulation runs: the neighbourhood system and the parameter for the substitution model are corresponding to Tab. 4.2A (archaeobacteria 5S rRNA). All simulations run along a star tree. Moreover, all branch lengths have the same length $d/2$ (Tab. 4.2). We did 40 runs and get alignments under different branch length for the desired main pairwise identity (MPI): 30, 40, \dots , 95 and with different numbers of sequences: 5, 10, \dots , 100 (taxa). **B:** Accuracy of the consensus structure prediction on these alignments using the common measures of sensitivity and selectivity (mean of 40 runs). **Black lines:** Mutual information context (MIC); **Red lines:** Thermodynamic consensus matrix (TH); both using the program `ConStruct` (Lück *et al.*, 1996; Wilm *et al.*, 2008).

is one minus the Hamming distance. Using the program `ConStruct` (Lück *et al.*, 1996; Wilm *et al.*, 2008), we analyse structure predictions under the mutual information context

(MIC) and the thermodynamic consensus matrix (TH). In addition, we determine the accuracy of both prediction methods (MIC, TH) using different neighbourhood systems as a reference and compare the consensus secondary structures predicted by **ConStruct** to the neighbourhood systems. We compute the common measures of sensitivity ("hit rate") and specificity (selectivity)¹. Furthermore, we calculated the structure conservation index (SCI) corresponding to Equ. 2.34 (page 38) using **RNAalifold** (Hofacker *et al.*, 2002a)². Note that **RNAalifold** covariance measure is not a mutual information score (Sec. 2.3.2).

We begin with simulations along star trees, thus ignoring the influence of the tree topology. Fig. 4.6 gives an overview of the simulations. All branches have the same length $d/2$. Thus, each taxa is d substitutions away from each other and we should get alignments with the computed corresponding MPI, see Tab. 4.2.

The prediction results of the simulation study using the data of the archebacteria 5S ribosomal RNA databank (Szymanski *et al.*, 2000) are shown in Fig. 4.6 after the overview of the simulation runs. The calculation with MIC results (black lines) in higher levels of sensitivity and selectivity with an increase in the number of sequences and an decrease of pairwise identity. We analysed the same alignments under TH (red lines). There is little difference shown in TH in terms of selectivity, however, in comparison to MIC there are fewer false positive predictions at the beginning. In contrast to MIC with TH the amount of sensitivity is low and constant. This is not such a big surprise because our simulations run so far only with compensatory mutations. For short branch lengths (high MPI) and few numbers of sequences both methods have a very low sensitivity and selectivity.

We illustrated realisations with MIC (Fig. 4.7A) and TH (Fig. 4.8C), represented in circleplots from simulated alignments with 100 taxa. In the middle of Fig. 4.7, we see the neighbourhood system defined by the PS. On the x axis there is the number of substitutions and the y axis as the Hamming distance. Under MIC (Fig. 4.7A), starting initially with very little information, the realisations become more and more similar to the neighbourhood system as $d \rightarrow \infty$. At the point of the saturated Hamming Distance, we observe the greatest sensitivity and the structure conservation index is correlated to the Hamming distance.

With TH, at the beginning, a short time after the ancestral node, false positives, as whole helices, appear with high significance. In the saturated part of the Hamming distance, no false positives appear, but the sensitivity is lower than with MIC.

The simulation study of Fig. 4.6 will continue after the final version of **SISSI** with energy

¹Sensitivity is defined by $TP/(TP + FN)$ and selectivity with $TP/(TP + FP)$, where TP is the number of "true positives" (correctly predicted base-pairs), FN is the number of "false negatives" (base-pairs in the reference structure that were not predicted) and FP is the number of "false positives" (in-correctly predicted base-pairs). We have used the script **comparect.pl** by Gardner and Giegerich (2004), which used a slightly different version to classify FP base-pairs as either inconsistent, contradicting or compatible (for details refer to Gardner and Giegerich (2004)).

²We set the covariance term for **RNAalifold** to one and respectively for the thermodynamic consensus to zero corresponding to the analysis.

(A) MIC Realisations

(B)

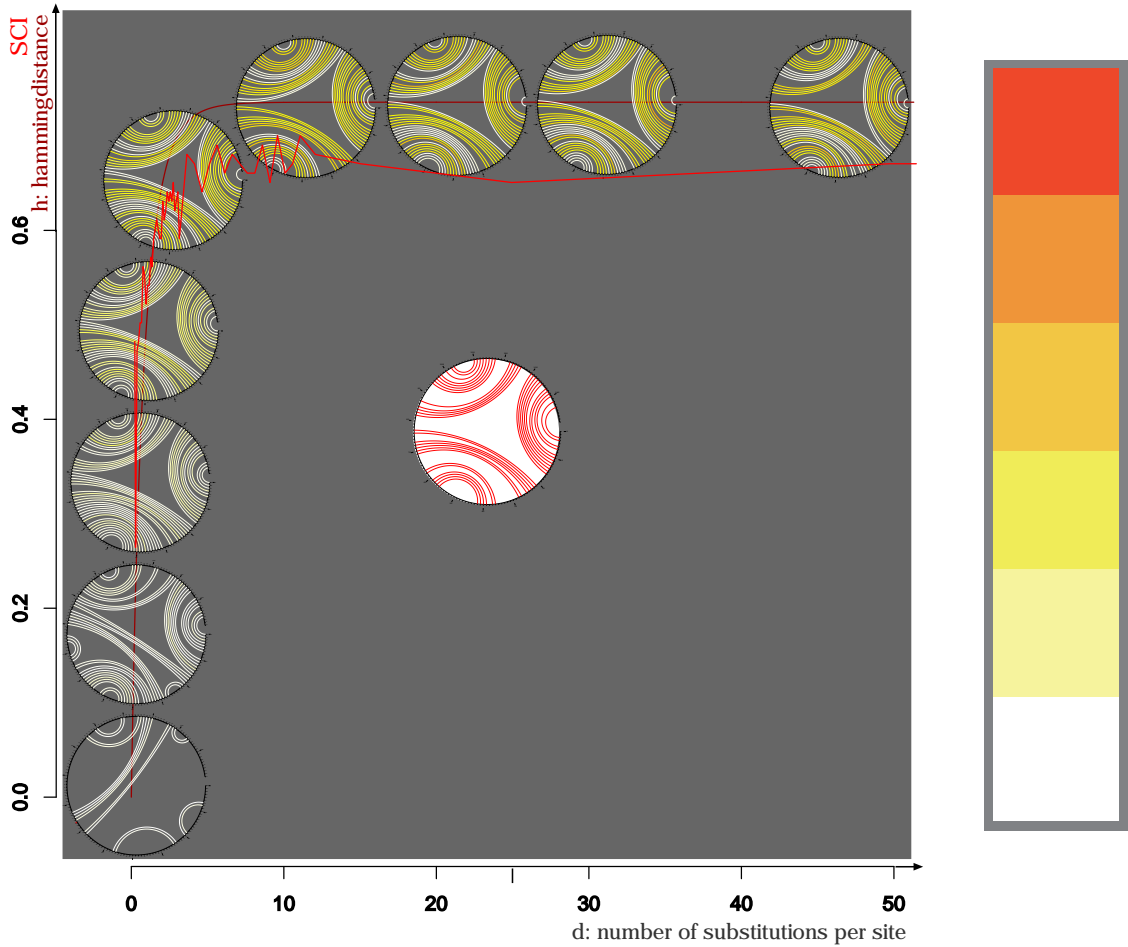


Figure 4.7: **A**: Structure prediction under MIC with simulations a star trees with 100 taxa, presented as circle plots. The middle shows the neighbourhood system of the simulations with the highest probability (red). On the x -axis there are the number of substitutions and the y -axis shows the Hamming distance. The red line describes the Hamming distance h as function of d . The second light red line shows the structure conservation index (SCI). Starting initially with very little information the realisations become more and more similar to the neighbourhood system over time. At the saturation of the Hamming distance the similarities are at largest and characterised by high sensitivity. The structure conservation index is correlated to the Hamming distance. **B** shows the colours used by ConStruct (Wilm *et al.*, 2008) for the base pair probability from white to yellow to red with the highest probability.

(Sec. 3.2). Here, it might be interesting to see implicitly how much thermodynamics is in covariance methods and furthermore to illustrate their differences. This may also help to address the differences of the predicted ncRNA candidates between the thermodynamic and probabilistic methods, which are still alternatives (Washietl *et al.*, 2007b). It illustrates that MIC and TH transformations show different structure information of a PS. However, these methods could complement one another dependent on the branch lengths.

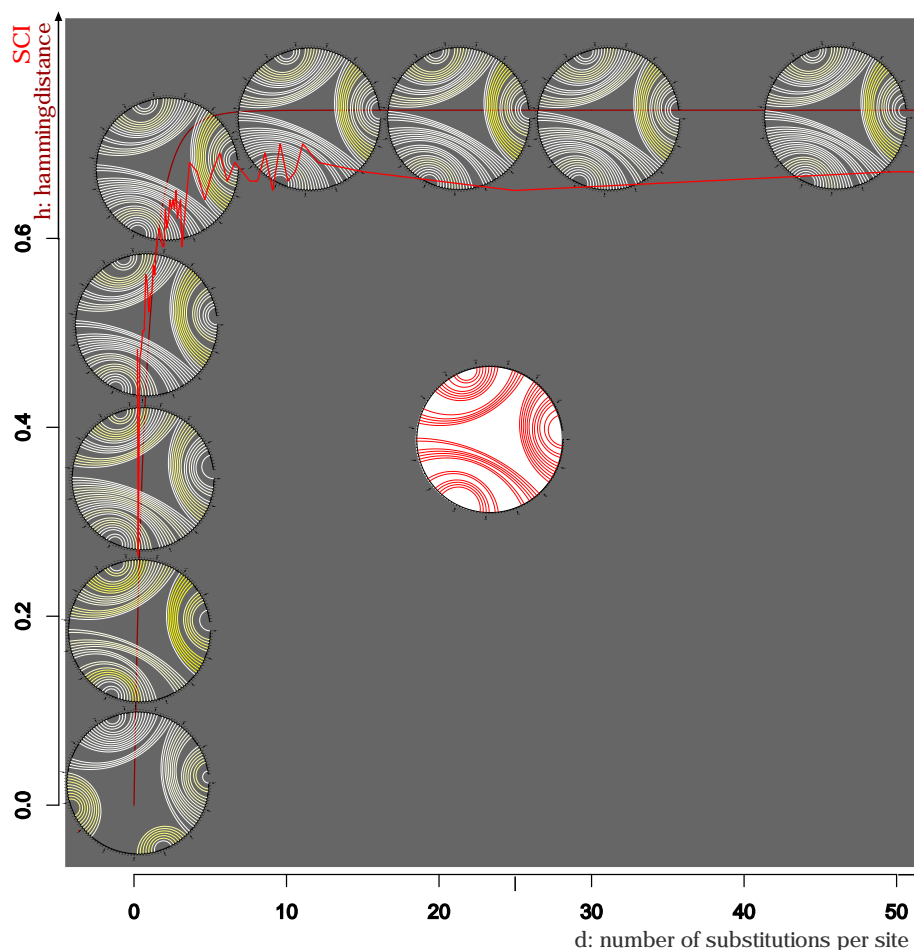


Figure 4.8: In contrast to the MIC realisations (Fig. 4.7), TH realisations are shown. At the saturation of the Hamming distance the similarities are at their greatest and have high sensitivity, but not so high as under the MIC methods.

Influence of the Three Aspects of a PS

So far, we have simulated along star-trees. Now, we simulate along three different tree topologies, but with the same mean branch length (Fig. 4.10). We have chosen the mean branch length of each topology according to Tab. 4.2. However, we divided the tree into four levels and changed the distributions of the branch lengths at each level. The first tree topology has very short external branch lengths, while the third tree topology has long external branch lengths. The second tree topology has a uniform distribution over all branch lengths. Fig. 4.10 shows the results for 40 runs with the same neighbourhood system and substitution model as the simulation study along star trees (Fig. 4.6). While

the accuracy of the thermodynamic consensus prediction doesn't show many differences between the simulation under different tree topologies (red lines in Fig. 4.10), the accuracy of the prediction under the mutual information content is different (black lines in Fig. 4.10). We see an increase in the sensitivity and selectivity from the first to the third topology. Fig. 4.9 highlighted the sensitivity of MIC predictions for the 8 taxa trees. In addition, the figure shows simulation runs with the 8 taxa trees with the neighbourhood system and the parameter of example IA (Tab. 4.1) and one with the same neighbourhood system like in Fig. 4.10, but the parameter for the substitution model of example IA . In comparison of the tree topologies the accuracy of the structure predictions with simulations along the first tree with short external branch lengths is always lesser than simulations along the other trees. Simulations along a tree with short external branch lengths, the so-called "bushes", generate a huge fraction of closely related sequences. With short branch length there are not enough mutations and the influence of the states at the internal nodes is higher. This can mislead the consensus structure predictions, if intermediate structures deviate by chance from the neighbourhood system, which we already called ancestral correlations. This might be the case for the tree topology one: However, with a higher number of taxa the tree topology has a lesser high influence (see Fig. 4.10). We observed no impact of the tree topologies on the accuracy of the thermodynamic consensus predictions.

If we change the neighbourhood system in the simulation runs with the PS of Fig. 4.1, the accuracy for TH is nearby one for sensitivity and selectivity in all simulations (data not shown). Also, the MIC predictions depends on the neighbourhood system (Fig. 4.9). The datasets under simulations with the artificial neighbourhood system is always the highest in selectivity and sensitivity than data sets under simulations with the neighbourhood system of example two (archaea 5S RNA). The parameters of the substitution model shows further influences on the structure accuracy.

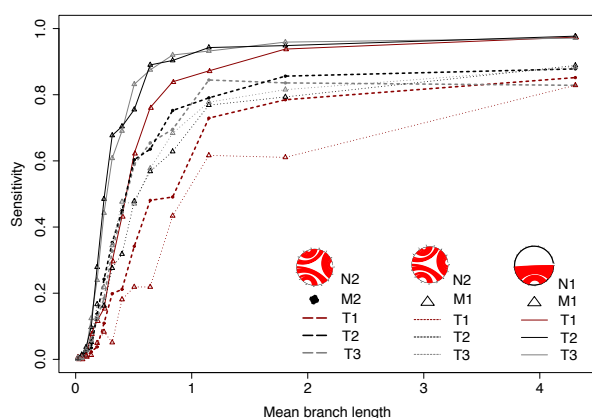
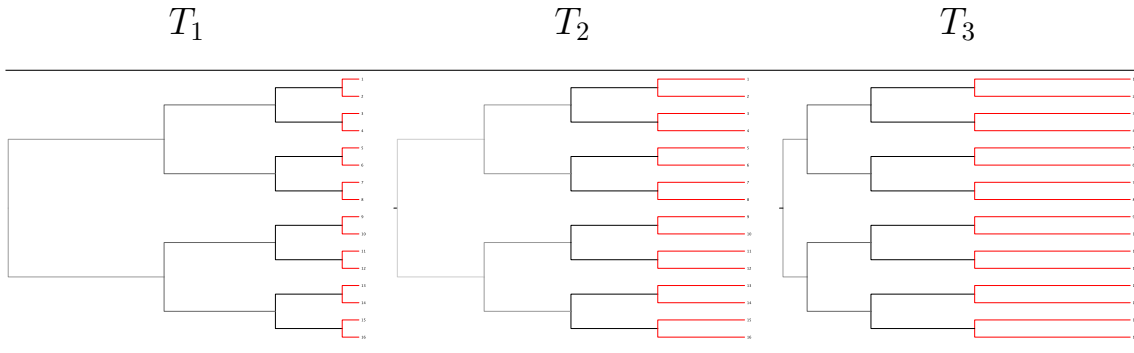


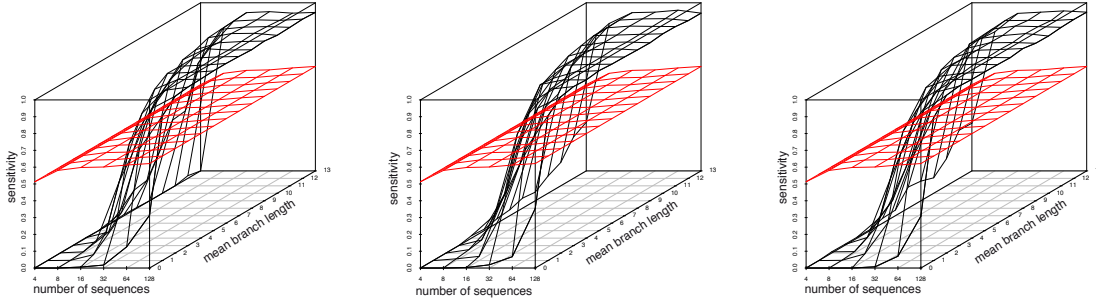
Figure 4.9: Details of Fig. 4.10 of the sensitivity of the MIC predictions for simulations along the 8 taxa trees. In addition, the MIC sensitivity for two different PSs is shown. Employing the neighbourhood system, we use the artificial neighbourhood system of the PS of Fig. 4.1 and as well as in Fig. 4.10 the archebacteria 5S RNA neighbourhood system (Fig. 4.6). The light point lines show the MIC sensitivity under simulations with the 5S RNA neighbourhood, while the substitution parameter (M1) correspond to example IA (Tab. 4.1).

Simulation Study along Different Tree Topologies



\mathcal{N}_{5S} + Substitution Model (Example II) + $T_{\{1,2,3\}}$

Sensitivity



Selectivity

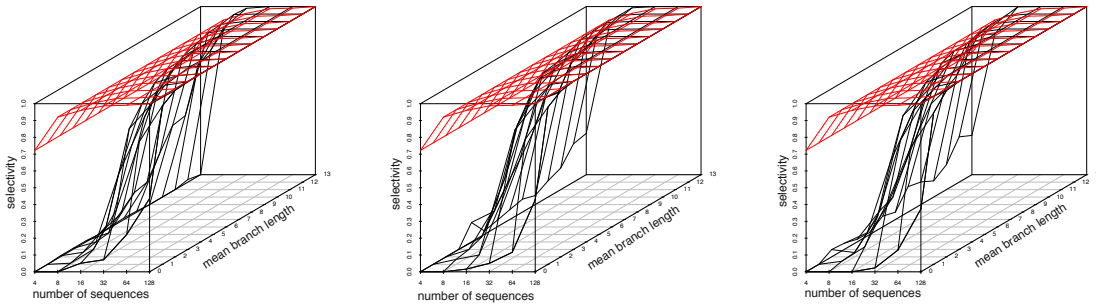


Figure 4.10: We simulate along three different tree topologies with the same over all mean branch length, but different distributions of the branch lengths along the tree. In order to do that, we have divided the tree into 4 levels, as indicated in the figures by different colours (red, black, dark and light grey). We chose the mean branch length according to Tab. 4.1, as well as the parameter for the substitution model and the neighbourhood system. Corresponding to the simulation study along star trees (Fig. 4.6), we did 40 runs and we considered the power of the methods using the measures of sensitivity and selectivity. Black lines: Mutual information context (MIC); Red lines: Thermodynamic consensus matrix.

4.4 (Un)structured RNAs

Here, we construct two PSs with the same neighbourhood system (the artificial one of Fig. 4.1), the second tree topology T_2 (Fig. 4.10), but different parameter for the substitution model using the permutation example (Tab. 4.1). We simulate alignments (Fig. 4.11A) under the substitution model parameter of example I A of Tab. 4.1A (observed frequencies of the ribosomal RNA) and the second alignments (Fig. 4.11B) under the permutation of the first two rows (Tab. 4.1B). The results of the consensus structure prediction under MIC and TH are shown in Fig. 4.11A and B. Although the matrices of the substitution model are different, the sensitivity and selectivity is high with high taxa and diverse alignments (only sensitivity is shown). In contrast, the accuracy of the TH under A shows a accuracy for sensitivity and selectivity of nearby one, while under B (permutation of the rows) the accuracy for TH is zero (Fig. 4.11B). Therewith we demonstrate our concept of (un)structured RNAs: **a PS can generate sequences, which although they show no structure conservation on one realisation level, they might show on another realisation level, very high structure conservation.**

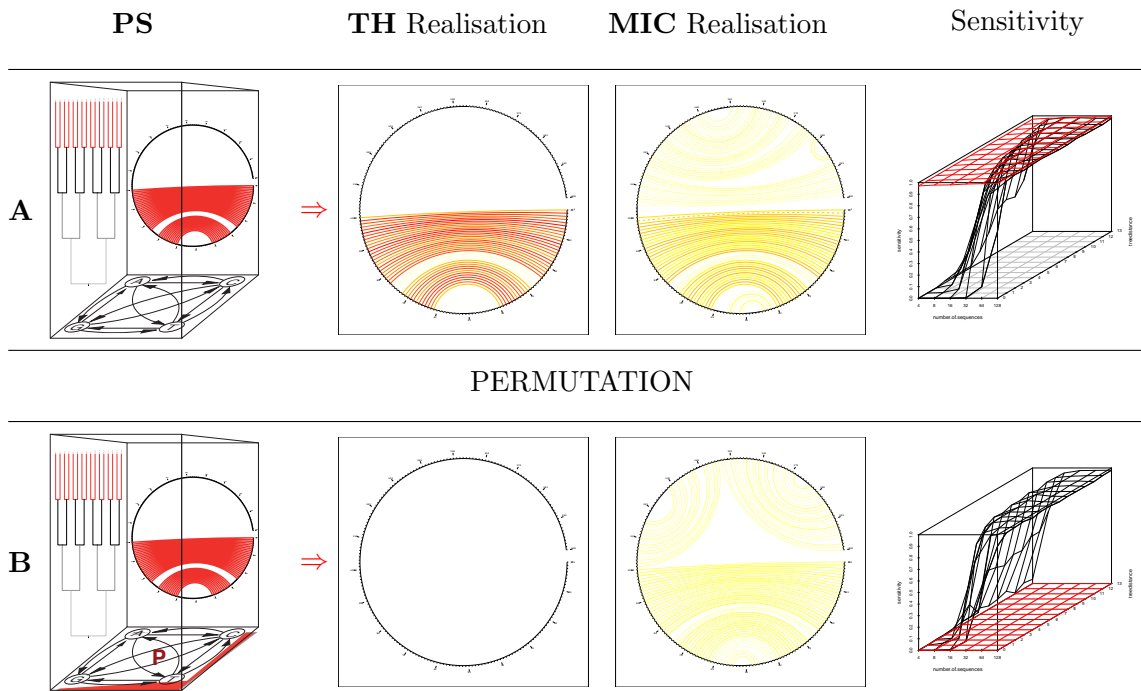


Figure 4.11: Permutation of the Substitution Model: Simulation Study, including MIC and TH consensus structure predictions. **A:** Simulation along a PS with the neighbourhood system and the parameter for the substitution model of example one A (Tab. 4.1) to simulate alignments along the second tree topology. We show one realisation for TH and MIC represented in circleplots. Both are examples for a 32 taxa tree with a mean branch length of 4 number of substitutions per site. In addition, we show the sensitivity for runs with different taxa and mean branch length along T_2 . **B:** Permutation Example: the first two columns of the substitution model of (A) are exchanged (Tab. 4.1B).

Lineage Specific Neighbourhood System (LSN)

We call a change in the neighbourhood system along the tree a *lineage specific neighbourhood system* (LSN). To illustrate what is meant by LSN, we made simulations with two different neighbourhood systems using the artificial example already discussed. We simply exchanged the paired sites with the unpaired sites for the neighbourhood system. The results can be seen in Fig. 4.12. If we compute the consensus structure from each simulation separately, we get a realisation corresponding to the appropriate neighbourhood system as explained previously. However, if we combine the simulation in one alignment and then compute the consensus structure, we can see, even though with less significant, pairs over the whole strands. Thus, *neighbourhood mutations*, respectively LSNs lead to misleading predictions by consensus structure predictions. Our illustration here is only a simplify example. More interwoven neighbourhoods with slower changes are likely, where the difference to the diversity of the realisations of a PS might be less and difficult to observe, or in contrast no base pairs can be predicted. To go in further detail here is out of the scope of this thesis. However, a phylogenetic definition of structure is a prerequisite for the definition of a LSN and consequently to structure evolution. Following our definition, *lineage specific evolution* might be also possible with the same neighbourhood system, for example with a different substitution model along the tree, but the same neighbourhood annotation. So far, the influence of the intertwined relationship of LSNs and lineage specific substitution models on structure evolution is not understood.

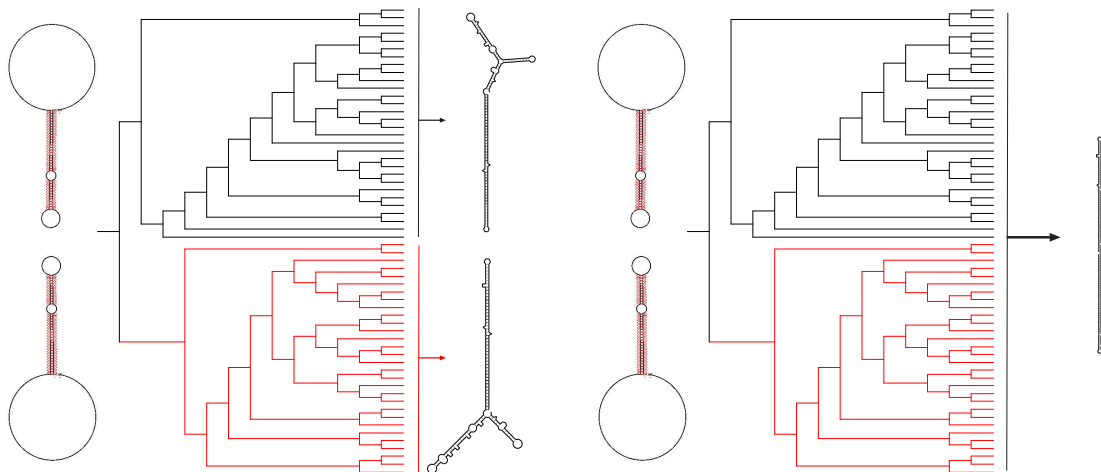


Figure 4.12: Lineage Specific Neighbourhood System (LSN): we made simulations with two different neighbourhood systems using the artificial example N_1 already discussed and another one N_2 , where we simply exchanged the paired sites with the unpaired sites. Left: consensus structures from each corresponding alignment under N_1 and N_2 predicted separately. Right: predicted consensus structure from the combined alignment.

4.5 The Definition Paradox

In this chapter we have introduced a *phylogenetic definition of structure*. Using SISSI we have shown that all three aspects of the definition: a neighbourhood system, a substitution model and a phylogenetic tree, build an intertwined relationship. All three aspects influence a realisation and the diversity of realisations of a PS.

Referring back to the Tab. 2.2 on page 8 of structure definitions, we are now able to differentiate between the definitions. To a large degree the original definitions in Tab. 2.2 are understood as structure definitions and can be related to a PS at different levels. Using the concept of a PS we shall distinguish between *descriptions, realisations, abstractions on descriptions or realisations*. Fig. 8.2 gives an overview of our classification of RNA definitions.

A PS provides different levels of abstraction, which allow the classification of existing structural definitions presented in the introduction.

Descriptions are *primary, secondary, tertiary and quaternary structure* as described in Chap. 2.

Realisations of a PS are related to structure predictions, which fold the whole sequence into a structure. From these realisations on each description level *coarse grained structures* and *consensus structure* can be abstracted, which can also be transformed into each other.

Coarse grained structures are basically an abstraction of the realisations of a PS.

One example are being abstract shapes: Giegerich *et al.* (2004) has developed a definition of the concept of *abstract shapes*, coarse-grained abstractions of full secondary structure (Voss *et al.*, 2006; Steffen *et al.*, 2006).

Consensus Structure is a common realisation of structure derived from two or more different RNA sequences and/or realisations. Please refer to Chap. 2, Sec. 2.3.2.

Here, we have focussed our attention on the consensus structure, using mutual information methods (MIC) compared to thermodynamic methods (TH) as one example. We have illustrated the interweavement of the three aspects of a PS using SISSI. Moreover, we have shown that a PS can generate sequences, which although they show no structure conservation with respect to thermodynamic structure, they show highly structural conservation with respect to the mutual information content. What does that mean if a sequence is “*unstructured*” on one realisation level, although it is highly “*structured*” on another realisation level?

Indeed, the potential function of ncRNAs appear to be extremely diverse and can regulate gene expressions at many levels by using a wide array of mechanism (Amaral *et al.*, 2008). In addition, recently, promising studies (Tafer *et al.*, 2008; Kertesz *et al.*, 2007; Hofacker, 2007) investigated the role of target-site accessibility, as determined by base-pairing interactions within the mRNA, in microRNA target recognition.

Consequently, it is important to take different realisations with different methods, into

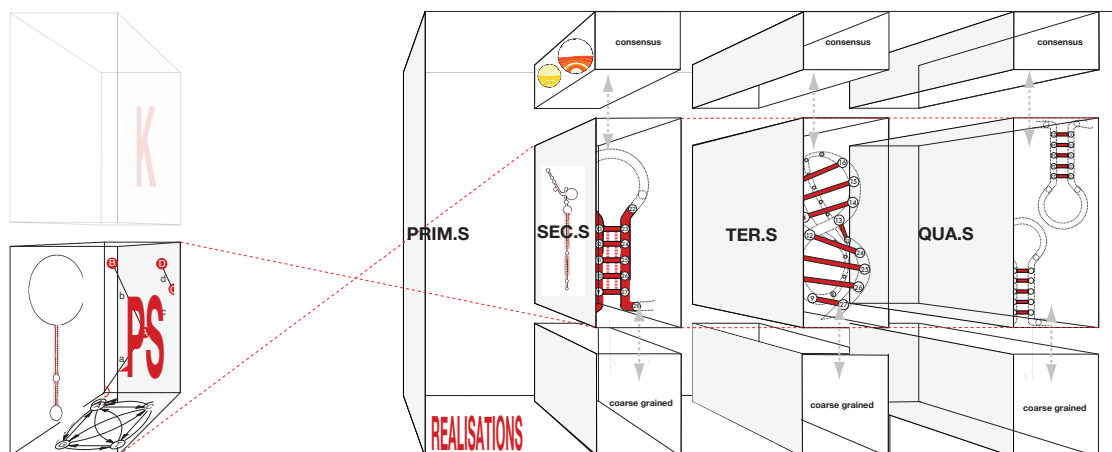


Figure 4.13: Classification of Structure Definition. **Left:** A PS (phylogenetic structure) as described in Fig. 4.1 transformed into different realisations. **Right:** The *realisations* can be described on the *description level* as primary, secondary, tertiary and quaternary structure, symbolised with graphics in the boxes. However, as described in Chap. 2, there are different types of transformations into realisations. As an example, we show the mfe-realisation of Fig. 4.2 on the “screen” (gray) of the secondary structure level. In addition, we can get different abstraction levels from each description or realisations, as *coarse grained structure* (at the bottom) and *consensus structure* (at the top).

We have focussed our attention on the secondary consensus structure with MIC (minimum free energy) and TH (thermodynamic) consensus structure predictions, illustrated with Fig. 4.11 on the secondary consensus structure “screen” at the top. There is limited experience in the realisation of tertiary structure and quaternary structure. At the moment, this is a largely unexplored area. Were these transformations to exist, we can go into the further abstraction levels, as *coarse grained structure* and *consensus structure* of tertiary and quaternary structure.

(left top: a kinetic processes (K), taking the kinetic folding time of RNA into account (e.g. Flamm and Hofacker, 2008), also have influence on realisations of a PS. However, this is not the focus of this thesis. The influence of the kinetic process on a PS and vice versa is not understood, but future research should combine both approaches in one framework.)

account to predict the variants of functions for ncRNAs.

Recently, several new tools and updates for computing consensus structures have been proposed. For instance, `RNAalifold` has substantially improved the accuracy of the consensus structure prediction. Bernhart *et al.* (2008) have included two parameters to fine-tune the impact of the covariance including the influence of the covariance score relative to the total folding energy. The new version of `ConStruct` (Wilm *et al.*, 2008) uses several mutual information scores and an improvement of the ability to predict tertiary interactions. However, open questions are optimal weighting between thermodynamics and covariance automatically and how to include the phylogeny without compromising the efficiency of the algorithm. Furthermore,

assuming a wide range of different realisations, we should combine thermodynamic models with other constraints, including the viewpoint of “(un)structure RNAs” and the dependencies of the three aspects of a PS.

This aim clearly also holds for single sequence secondary structure prediction and the other description levels. For several decades, free energy minimization methods have been the dominant strategy for single sequence RNA secondary structure prediction, where the most thermodynamical single secondary structure prediction programs make use of the nearest-neighbour models (Sec. 2.1.3) using the energy parameter of Mathews *et al.* (1999). Recently, probabilistic methodology has emerged as an alternative for modeling RNA structure (Do *et al.*, 2006; Hamada *et al.*, 2009). In addition recent approaches for optimising the energy parameters exist, e.g. constraint generation method based on a constraint optimisation problem (Andronescu *et al.*, 2007). Currently Boltzmann likelihood studies are ongoing, with prior knowledge of relationships between energy parameters and by integrating prediction results from a portfolio of algorithms, which use different energy models. Improvements in prediction accuracy are already obtained (Holger H. Hoos, personal communication). From a phylogenetic structure viewpoint the question arises as to the existence of one ‘perfect energy model’? In addition to structural data from a database of highly trusted secondary structures, the influence of different well-defined features during evolution should be considered. `SISSI` would be an ideal tool to define these features in a substitution model and an annotation of site-specific interactions and to generate the necessary test sets. Finally, the tree aspect should be included.

While bioinformatics research has made progress on the description level of RNA secondary structure in the past years, more recently the other description levels come more into the focus of attention. So far methods for predicting RNA interactions have focused on using single sequences, although existing RNA interactions might contain preserved or covarying patterns of interactions. In the case where energy is not the appropriate measure, we have to find another measure for structure conservation, as well as for different description levels, e.g. tertiary structure or quaternary structure like RNA-RNA, DNA-RNA, or Protein-RNA interactions.

Structural Conservation Measurement at Different Levels

So far, different measures have been developed for structural conservation on the description level of secondary structure, based on folding energies, on single structures or considering the entire folding space (cf. Gruber *et al.*, 2008). However, not all of these methods are applicable to the other description levels like tertiary and quarternary structure.

From the viewpoint of other description levels than secondary structure and to avoid alignment problems, the improvements of a comparative approach using graphic-theoretic measures is worthwhile. In Fig. 8.2, we have illustrated the tree representation of RNA secondary structure, which has the potential to be extended to the other description levels. For the tree representation of RNA secondary structure, tree editing methods, where the distance between two trees is defined as the minimal cost to transform any tree T_x into any other tree T_y , induces a metric in the space of RNA secondary structures. Tree editing is the only method that can act on structures generated by RNAs with length differences. A series of programs to compare RNA secondary structure have implemented different kind of tree edit algorithms (e.g Shapiro and Zhang, 1990b).

Consequently, the tree representation should be extended to the other RNA description levels. In Dehmer *et al.* (2008) we have developed a conceptual extension of the classical graph similarity problem in a way that the structural similarity of sets of graphs was determined. A style has been defined as a set of relational objects (graphs) where every graph processes certain characteristics imposed by a set of properties. The main assertion of Dehmer *et al.* (2008) was that we compared two styles by a comparative analysis of the underlying graphs. Suppose we have two styles representing a set of rooted trees, in our case RNA structures, our method might be an alternative strategy for measuring evolutionary conservation at different description levels in future research.

Phylogenetic Structural Conservation

Different opinions were published how to define RNA families. Homologs can be found on the basis of sequence homology, thus homologs may constitute a family. Furthermore, covariance models from a multiple alignment with structural annotation can be used. For instance, Rfam used stochastic context-free grammars (SCFGs) to assign new sequences to families. Compared to purely sequence based methods these tools detect also remote homologs. RNA can be grouped together, forming a ncRNA class whose members have no discernible homology at sequence level, but still share common structural and functional properties. Thus, structure based clustering might a promising way to define novel families of structured RNAs (e.g Will *et al.*, 2007).

In Fig. 4.14 the phylogenetic history suggests that D and E are a family. However, due the observed secondary structure realisation B and D build a class. To give a definition of RNA families would be out of the scope of this thesis, however, from our viewpoint for a definition of RNA families a *phylogenetic structure definition* should be taken into account. From this viewpoint, substitution models including site-specific interactions seem

necessary to define RNA families, as well as a phylogenetic structure conservation index. A PS improve our understanding of RNA structure evolution. For example, we have shown that we have to distinguish between different constraints, like ancestral and neighbourhood constraints, as well as different observable correlations such as ancestral correlations and functional (neighbourhood) correlations. In the viewpoint of an extension of the structure conservation measurement future research should aim to include all aspects of a PS into one measure, which one might call *phylogenetic structure conservation index* (PSci). However, it is not clear if such an index is definable. If a PSci is at all possible, then the values of the PSci will partition our PS into equivalence classes. Finally, we have defined lineage specific neighbourhood system (LSN) and a PSci should be able to detect LSNs. If this were possible, this would improve structure predictions and conservation in the future research. We will discuss that further in the outlook.

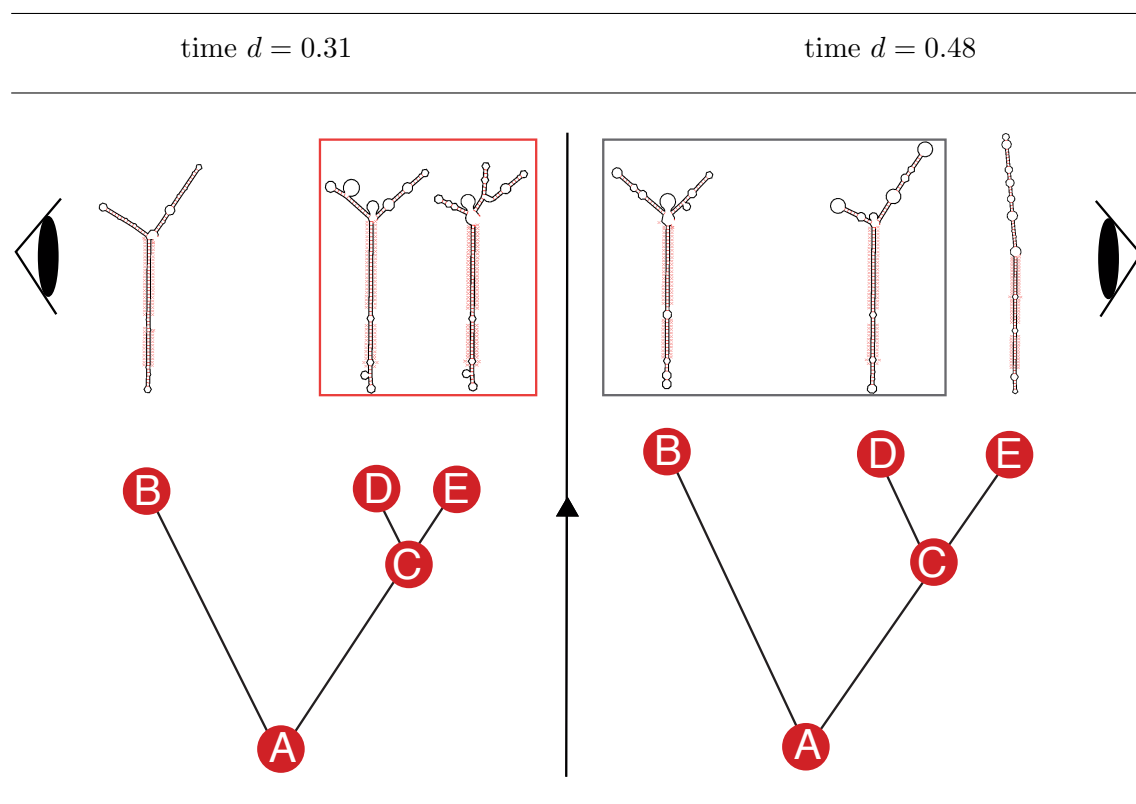


Figure 4.14: Example from our film, illustrated in Fig. 4.4. Left: short time after a speciation event, we have similarities based on the realisation at the speciation event C and not on the neighbourhood system in the upper part, see Fig. 4.4 for further details. Right: after more time has elapsed, B and D have more structural similarities than either has with E, although D and E are closer related in the phylogenetic tree. Thus, the phylogenetic history suggests that D and E are a family, while due the observed secondary structure realisation B and D (gray rectangle) can build a class as well as D and E (red rectangle).

So far, it is not clear, how the three aspects depends on each other and influence the accuracy of the structure prediction programs. The consensus structure prediction used here do not account for the phylogenetic relationship in a proper way. **SISSI** may help to address this particular issue in future research.

The next three chapters focus on particular examples, each chapter is devoted to one of the three aspects of a PS.....

Chapter 5



Mata Hari (1876-1917)

MATA's

Neighbourhood Aspect

... the Other is not simply the Other as coming from the outside so to speak. One is the one, I am the one, one is more or less the one and everyone is more or less the one and more or less one with him or herself. Which means that the Other is already inside, and has to be sheltered and welcomed in a certain way. We have to negotiate also, that's a complicated unconscious operation, to negotiate the hospitality within ourselves (Derrida in Bennington, 1997). TO TANJA WANDA

In Chap. 3 we introduced site-specific interactions using a neighbourhood system. It allows for a universal description of arbitrarily complex dependencies among sites. In Chap. 4 we discussed the neighbourhood system as one aspect of our definition of PS. Moreover, we illustrated that it is problematic to distinguish true positive correlations from false positive correlations (Fig. 4.5). Here, we consider the inference of a neighbourhood system with phylogenetic trees from a comparative viewpoint given a set of sequences, represented as a multiple sequence alignment.

We propose a method, MATA (Measurements of Accurate Thresholds of Alignments for Structure Prediction Programs) using the parametric bootstrap that enables the detection of functional correlations from a sequence alignment incorporating the phylogeny of the sequences. MATA identifies most positions in an alignment causing false positive associations. We show that false positive associations are due to positions for which the ancestral state of the root nucleotide cannot be assessed properly. Our method defines a null model that takes into account a phylogenetic tree on which sites evolved independently. Subsequently a χ^2 -statistics is applied to detect significant inter-site associations. Finally, we discuss MATA as a framework to measure accurate thresholds of alignments for other structure prediction programs. ¹

¹MATA, alias INF-DEP, is a cooperation with T. Schlegel and A. von Haeseler. It was partly presented in T. Schlegel (2007) thesis. We thank G. Steger for providing the alignment of the riboswitch.

5.1 Measurements of Accurate Thresholds of Alignments

MATA combines the advantages of the methods to predict structure from alignments with phylogenetic information and an automatic procedure to filter false positive correlations. While doing this, we assume that each sequence in the alignment has the same lineage specific neighbourhood system (LSN, Sec. 4, Chap. 4). Comparative methods investigate if the number of nucleotide pairs at two positions in a sequence alignment differs significantly from random expectation (see Chap. 2, Sec. 2.3.2). If this is the case, then both positions are called correlated. However, these approaches are only statistically valid if each sequence in the alignment represents an independent sample of the same evolutionary process (e.g. Chap. 2, Fig. 4.10). As sequences are related by a phylogeny, this assumption is obviously violated, unless the sequences are related by a “star” phylogeny. Therefore, such methods are too generous in suggesting associations (Lapedes *et al.*, 1999). Comparative methods detect not only base pairs in helical regions, like thermodynamic methods, but also so-called tertiary dependencies like pseudo-knots, or base triples (Gutell *et al.*, 1992a; Tabaska *et al.*, 1998; Ji *et al.*, 2004). They are able to suggest a *de novo* structure from an alignment. However it is very difficult to determine the appropriate significance level. Apart from the standard statistical problems that lead to false positive associations (Lapedes *et al.*, 1999; Pollock *et al.*, 1999), the influence of the topology of the tree and of its phylogenetic diversity (Faith, 1992) on the significance level is not understood.

Here we introduce MATA, a method that allows statistical inference of associated sites. The input for MATA is a multiple sequence alignment and a phylogenetic tree displaying the evolutionary relatedness of the data. MATA combines the advantages of the comparative method with an automatic procedure to filter false positive associations. In a first step, to demonstrate the workflow, we are concentrated on a comparative method, a χ^2 -statistic (Klingler and Brutlag, 1993).

With $\mathbf{D}_{data} = (D_1, \dots, D_l)$ we denote a sequence alignment of length l with n sequences. That is, D_i represents the nucleotides at the i th site of the alignment for each of the n sequences. \mathbf{D}_{data} constitutes the data we want to investigate.

We assume that the n sequences are related according to a rooted tree T where the leaves represent the sequences in the alignment and the branch lengths of T reflect the amount of evolution. The evolution of the nucleotides is then specified by a model of sequence evolution (Tavaré, 1986; Rodriguez *et al.*, 1990) consisting of a rate matrix and a stationary nucleotide distribution. The rate matrix typically belongs to the class of general time reversible models with stationary distribution $\pi = (\pi_x)_{x \in \mathcal{A}}$, where \mathcal{A} denotes a finite alphabete, i.e. nucleotides or amino acids. However, since the sequences are related by a tree, the base composition at any site in an alignment may deviate dramatically from the stationary distribution. Obviously, the degree of deviation depends on the branch lengths and the nucleotide at the root of the tree. Fig. 5.1 shows an overview of MATA’s framework.

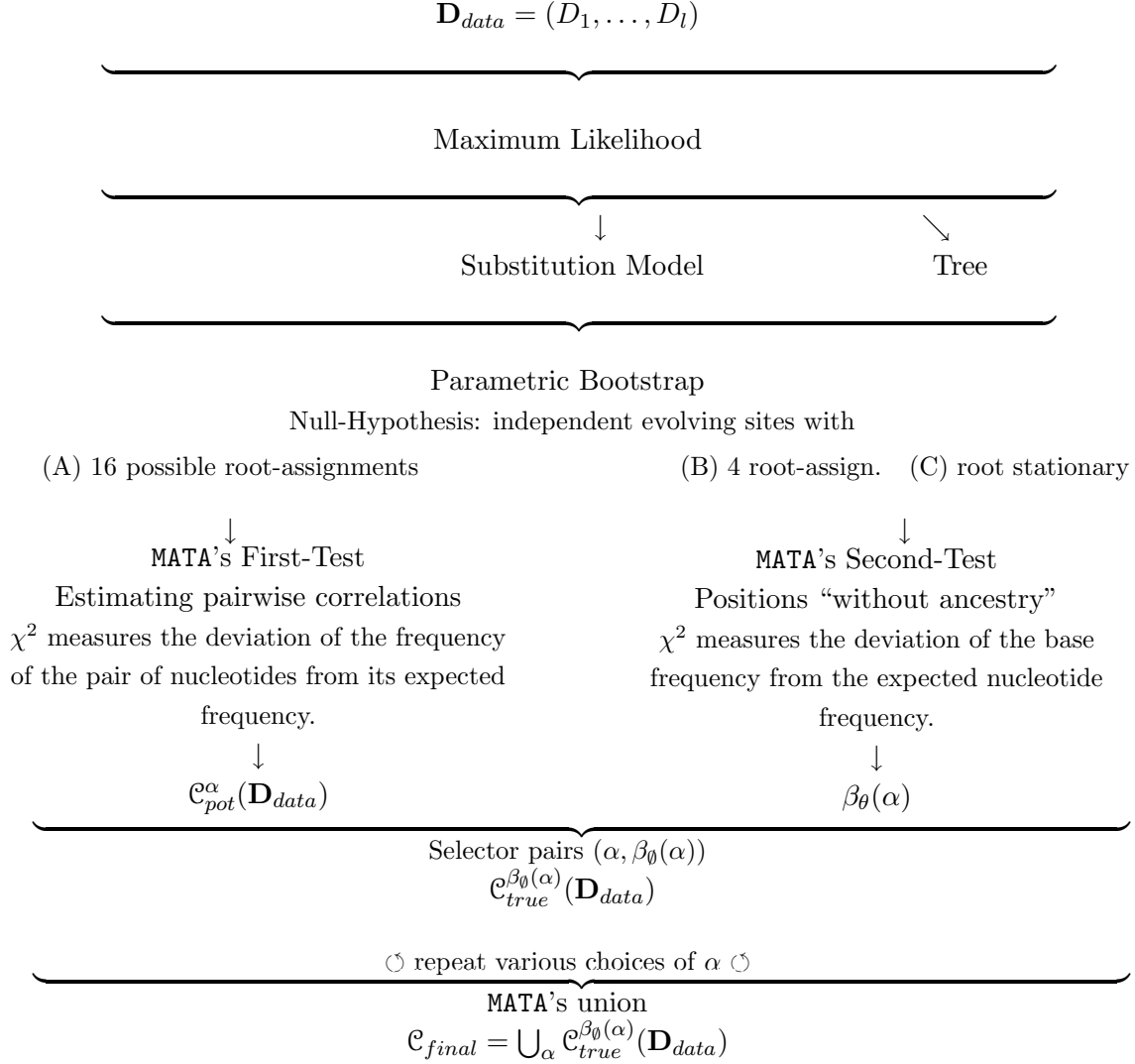


Figure 5.1: Overview of MATA's framework. Based on a given sequence alignment \mathbf{D}_{data} MATA computes two statistics. The first test suggests potential correlations, thus forming the set $\mathcal{C}_{pot}^\alpha(\mathbf{D}_{data})$ for a given significance level α . Then the second test filters false positive associations, where a parameter $\beta_\theta(\alpha)$ determines the amount of positions that are involved in false correlations. The remaining true correlations are collected in the set $\mathcal{C}_{true}^{\beta_\theta(\alpha)}(\mathbf{D}_{data})$. Because the value of $\beta_\theta(\alpha)$ depends on the choice of α in a complex way, we compute its value by generating a long alignment \mathbf{D}_{sim} using parametric bootstrap based on a maximum likelihood tree derived from \mathbf{D}_{data} assuming independence of positions. This procedure is repeated for various choices of α and the resulting collection of positions with significant correlation define the structure of the sequence. See text for details.

5.2 MATA's First-Test: Estimating Pairwise Correlations

To estimate the degree of association between two alignment positions we used a χ^2 -score. The χ^2 -score measures the deviation of the observed dinucleotide frequencies at alignment positions i and j from the expected dinucleotide frequencies assuming independent evolution of positions. The χ^2 -score is computed as:

$$\Delta_{uv}(D_i, D_j) = \sum_{x,y \in \mathcal{A}} \frac{\{O(x,y) - E_{uv}(x,y)\}^2}{E_{uv}(x,y)} \quad \text{for } u, v \in \mathcal{A}, \quad (5.1)$$

where $O(x,y)$ denotes the number of nucleotide pairs (x,y) that occur jointly in D_i, D_j and $E_{uv}(x,y)$ is the dinucleotide composition under independence conditional on the tree and the root, where (u,v) denotes the dinucleotide pair at the root. To analyse biological data we need to determine the null-distribution.

To this end, we determine the distributions of the $\Delta_{uv}(\cdot, \cdot)$ for each $u, v \in \mathcal{A}$. An analytical formula of the χ^2 -score distributions seems not feasible. Therefore the distributions are generated using the parametric bootstrap. We simulate the evolution of b independently evolving dinucleotide patterns $(X^k, Y^k), k = 1, 2, \dots, b$ along the phylogeny T with respect to the root nucleotides u and v . The expected nucleotide composition is then approximated by $E_{uv}(x,y) \approx \frac{1}{b} \sum_{k=1}^b O_{uv}(x^k, y^k)$ where $O_{uv}(\cdot, \cdot)$ is the same function as $O(\cdot, \cdot)$ in Equation (5.1) except that we keep track of the different dinucleotides at the root. The $\Delta_{uv}(X^k, Y^k)$ are computed according to Equ. 5.1. Thus, we get an approximation of the null-distribution of Δ_{uv} for each u, v , resulting in 16 distributions.

The p -value $p_{uv}(D_i, D_j)$ of the actually observed data $\Delta_{uv}(D_i, D_j)$ is then estimated by the proportion of simulated $\Delta_{uv}(X^k, Y^k)$ -values, $k = 1, 2, \dots, b$, equal to or larger than $\Delta_{uv}(D_i, D_j)$ for any fixed u, v . We obtain for the nucleotide patterns D_i and D_j at position i and j 16 p -values.

To classify alignment positions D_i and D_j as associated, we require that the null-hypotheses of independently evolving positions is rejected for the 16 possible root assignments on significance level α .

That is to say, if we assign at the root of D_i the nucleotide u and at the corresponding root of D_j the nucleotide v , then the p -values $p_{uv}(D_i, D_j)$ have to be smaller than α for all assignments of root nucleotides $u, v \in \mathcal{A}$, in other words:

$$\max_{(u,v) \in \mathcal{A}^2} \{p_{uv}(D_i, D_j)\} < \alpha. \quad (5.2)$$

We call positions i and j as associated if inequality 5.2 is true. Inequality 5.2 is based on the idea that only one $p_{uv}(D_i, D_j) \geq \alpha$ suffices to retain the null-hypothesis, i.e. explains co-occurrence of both patterns by means of independent evolution. The collection of potentially associated positions for alignment \mathbf{D}_{data} and a specified α is denoted by

$$\mathcal{C}_{pot}^\alpha(\mathbf{D}_{data}) = \{(i,j) | D_i, D_j \text{ fulfill inequality 5.2}\}. \quad (5.3)$$

We call this test **MATA's first-test** which estimate pairwise dependencies. Note that $\mathcal{C}_{pot}^\alpha(\mathbf{D}_{data})$ can be visualized in a circle plot graph, where alignment positions represent the nodes and $\mathcal{C}_{pot}^\alpha(\mathbf{D}_{data})$ defines the edges. In a nutshell: **MATA's first-test** describes a contingency test taking the tree T and the branch lengths into account. However, as we will discuss in the following, including the tree into the analysis does not suffice to reduce the number of false positive dependencies. Therefore, we need an additional step to further reduce the number of false positive inter-site associations. To this end, we introduce a second test.

5.3 MATA's Second-Test: Positions without Ancestry

Here we measure the deviation of the base composition from the expected nucleotide composition. Again, we use a χ^2 -score. The setting is similar to that of **MATA's first-test**.

$$\Delta_u(D_i) = \sum_{x \in \mathcal{A}} \frac{\{O(x) - E_u(x)\}^2}{E_u(x)} \quad \text{for } u \in \mathcal{A}, \quad (5.4)$$

where $O(\cdot)$ equals the observed nucleotide distribution at alignment site i and $E_u(\cdot)$ is the expected nucleotide distribution assuming nucleotide u at the root of the tree. We proceed as before and use the parametric bootstrap to generate the four distributions $\Delta_u(\cdot)$ ($u \in \mathcal{A}$). For a pattern D_i from alignment \mathbf{D}_{data} , we compute the p -values for each distribution. That is, we compute the proportion of $\Delta_u(X^k)$, $k = 1, \dots, b$ that is larger than $\Delta_u(D_i)$.

Given four possible root nucleotides u , one would intuitively expect to find one large p -value and three small p -values. The large p -value is the reverberation of the original ancestral root nucleotide, whereas the root nucleotides providing small p -values are probably not the true ancestral states. To capture this variation in p -values we compute the empirical standard deviation $\sigma(D_i)$ for the four p -values $\{p_u(D_i), u \in \mathcal{A}\}$. If $\sigma(D_i)$ is small, then we say the information about the ancestral nucleotide state is lost or not present.

To estimate the p -value for $\sigma(D_i)$ we determine the empirical distribution of $\sigma(\cdot)$. Again, we employ parametric bootstrap, that is, we randomly generate b nucleotide pattern X^k (typically $b = 1,000 - 10,000$) assuming stationarity at the root. For each pattern we compute $\sigma(X^k)$ as described. The fraction of $\sigma(X^k)$ smaller or equal than $\sigma(D_i)$ determines the p -value. If this value is smaller than a default value β then the site i is called false positive site.

Then, all pairs in $\mathcal{C}_{pot}^\alpha(\mathbf{D}_{data})$ comprising a false positive site are removed from the set. The choice of α and β influences the outcome of the tests and thus the set of potential associations. The **MATA** method outlined in the next section describes an algorithm to determine β as function of α .

5.4 Reconstructing Site-Specific Interactions with Phylogenetic Trees

Now we are ready to explain our strategy to determine the collection of associated sites. MATA needs as input the alignment, the phylogenetic tree T and the parameters for the nucleotide substitution model from the alignment \mathbf{D}_{data} . Based on the model and the tree, we generate an alignment \mathbf{D}_{sim} under independence using Seq-Gen (Rambaut and Grassly, 1997). From this alignment the distributions $\Delta_u(\cdot)$, $\Delta_{uv}(\cdot, \cdot)$ and $\sigma(\cdot)$ are estimated as described above.

In the following we describe the construction of a set of truly significant associated sites. Because \mathbf{D}_{sim} constitutes an alignment of independently evolving sites the set $\mathcal{C}_{pot}^\alpha(\mathbf{D}_{sim})$ should be empty. However, MATA’s **first-test** yields a set of false positive inter-site associations. Now, we apply MATA’s **second-test** to determine a value of β such that $\mathcal{C}_{pot}^\alpha(\mathbf{D}_{sim}) = \emptyset$. This value is denoted by $\beta_\emptyset(\alpha)$. This procedure is repeated for “every” α , ($0 < \alpha < 1$).

We end up with a collection of $(\alpha, \beta_\emptyset(\alpha))$ pairs, that serve as “selector pairs” to determine true associations in biological data \mathbf{D}_{data} . For each selector pair we define the set $\mathcal{C}_{true}^{\beta_\emptyset(\alpha)}(\mathbf{D}_{data})$ that comprises the collection of site-pairs (i, j) that could not be rejected for the given “selector pair”.

In a typical application, we start with small values of α , adjust $\beta_\emptyset(\alpha)$ accordingly and compute the number of associated sites for the data. Then we slightly increase α , adjust $\beta_\emptyset(\alpha)$, and again compute the set of associated sites. The union $\mathcal{C}_{final} = \bigcup_\alpha \mathcal{C}_{true}^{\beta_\emptyset(\alpha)}(\mathbf{D}_{data})$ of the resulting sets with true positive associates then constitutes our final collection of associated sites.

Performance of MATA

We investigated synthetic data and real data. The alignments of the **synthetic data** were generated using SISSI (Chap. 3) and one Phylogenetic Structure (PS) as an example, including a neighbourhood system, which is shown in Fig. 5.2 in red.

For **real data** we investigated a sequence alignment of 111 bacterial sequences (Gräf *et al.*, 2005) that included a purine riboswitch. The alignment length was 106 position, the riboswitch is located from position 19 to position 90. Riboswitches are genetic regulatory elements found in the 5’ untranslated region of messenger RNA (Batey *et al.*, 2004). The secondary structure of the *Bacillus subtilis* riboswitch (Batey *et al.*, 2004) consists of three helices that contain in total 20 base pairs. The circle plot of the secondary structure is displayed in Fig. 5.4. The phylogenetic tree and the parameter for the nucleotide substitution model were inferred using IQPNNI (Vinh and von Haeseler, 2004). We used the HKY nucleotide substitution model (Hasegawa *et al.*, 1985) for synthetic data and the riboswitch alignment. In addition, we used the general time-reversible substitution model with gamma distributed rates (Yang *et al.*, 1994).

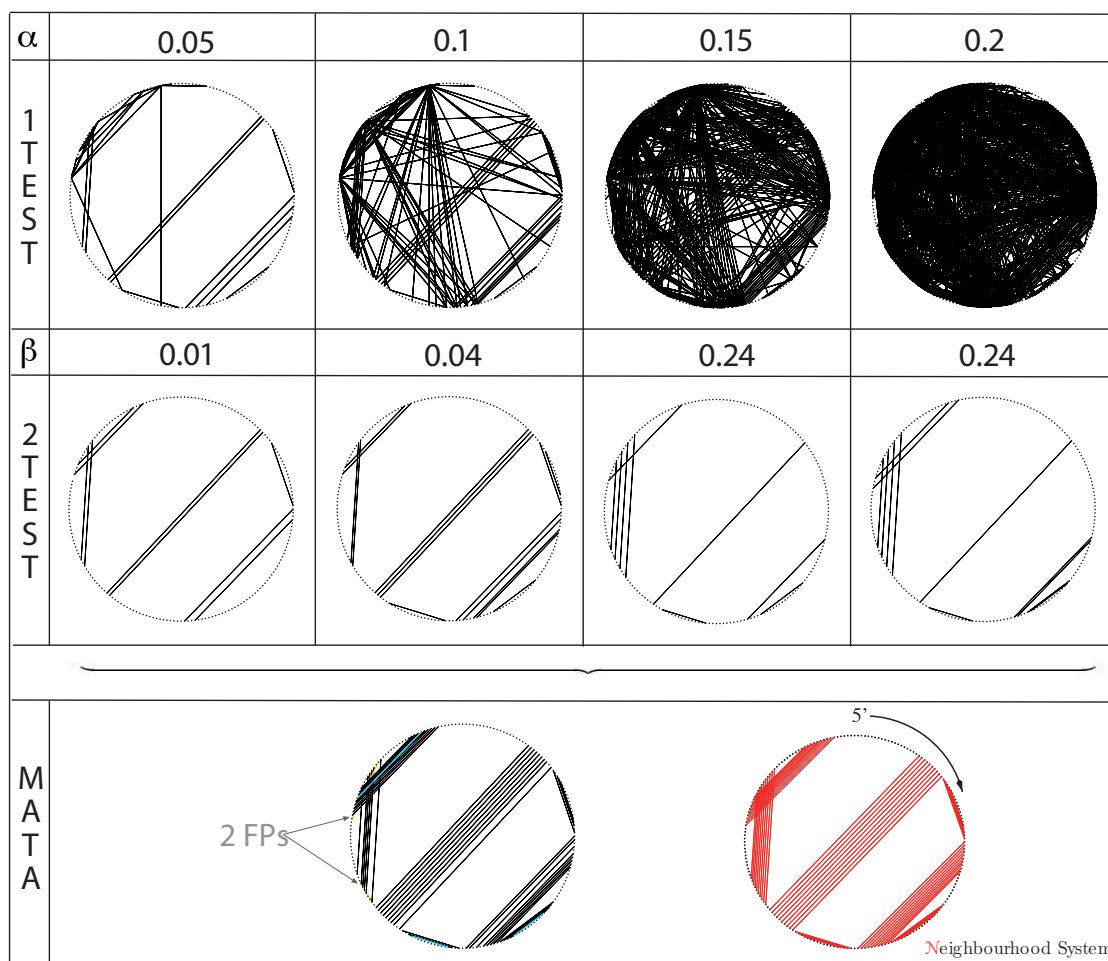


Figure 5.2: Dependency graphs of simulated data: The first row shows the potentially associated pairs \mathcal{C}_{pot}^α for different choices of α (MATA's first-test). The fourth row displays the remaining associated pairs $\mathcal{C}_{true}^{\alpha;\beta_0}$ after applying MATA's second-test, the corresponding β values are given in the third row. The set of final associations \mathcal{C}_{final} is obtained after superimposing all dependency graphs and displayed in the last row. The yellow lines represent the false positive predictions, the blue lines stand for predictions after a second round of MATA. Besides, the right dependency graph in red represents the neighbourhood system of the PS, which results in \mathbf{D}_{data} .

Performance of MATA using SISSI

We evaluated the ability of MATA to detect the inter-site associations of an RNA-molecule given a multiple sequence alignment \mathbf{D}_{data} . To this end we carried out a simulation using SISSI with one PS as an example: The neighbourhood system is 200 bases long and contains seven base paired regions, where one region represents a pseudo-knot (overlapping lines in the red dependency graph of Fig. 5.2. It contains 104 site with a cardinality $n_k = 1$ (including 54 base pairs) and 92 remaining sites with $n_k = 0$ (independent sites). The

α	$\beta_{\emptyset}(\alpha)$	$ \mathcal{C}_{pot}^{\alpha}(\mathbf{D}_{data}) $	$ \mathcal{C}_{true}^{\alpha, \beta_{\emptyset}}(\mathbf{D}_{data}) $	\mathcal{C}_{final}
0.01	0.0	0	0	0
0.05	0.01	26	9	9
0.1	0.04	106	15	19
0.15	0.24	354	9	34
0.2	0.24	724	12	40
0.25	0.46	1412	5	42
0.3	0.61	2137	4	43
0.35	0.66	3123	3	43
0.4	0.7	4025	3	43
0.45	0.8	5052	1	43

Table 5.1: MATA's results of the generated dataset: The first two columns display the selector pairs $\{\alpha, \beta_{\emptyset}(\alpha)\}$ used to determine inter-site associations. While α defines a significance level in the classical sense, the parameter β is estimated via simulation and determines the number of sites that are false positive predictions. For each selector pair we show the number of potential inter-site associations ($|\mathcal{C}_{pot}^{\alpha}(\mathbf{D}_{data})|$) and the number of true inter-site associations ($|\mathcal{C}_{true}^{\alpha, \beta_{\emptyset}}(\mathbf{D}_{data})|$) after applying the MATA's **second-test**. The last column (\mathcal{C}_{final}) displays the cumulative number of inter-site associations, that result from the union of the sets $\mathcal{C}_{true}^{\alpha, \beta_{\emptyset}}(\mathbf{D}_{data})$.

substitution model is the same as for the simulation in Chap. 4, an extended Fel model with the parameters in Tab. 4.1A. A phylogenetic tree with 100 leaves was generated under the Yule-Harding model (Harding, 1971). The branch lengths are randomly drawn from the interval $(0, 1)$. Based on this PS, \mathbf{D}_{data} constitutes a multiple sequence alignment, that is used for the subsequent analysis.

Using the inferred phylogenetic tree and the corresponding inferred substitution model of \mathbf{D}_{data} , an alignment \mathbf{D}_{sim} (length 1000) was generated, assuming independent sites. Then for $\alpha = 0.01, 0.05, 0.1, \dots, 0.45$ the corresponding $\beta_{\emptyset}(\alpha)$ values were determined. The results are summarized in Tab. 5.1. The second column specifies the value of parameter $\beta_{\emptyset}(\alpha)$ for varying α . The third column shows that the number of potential associations increases with growing α . This is obviously due to the increasing number of false positives. However, if we would fix a small α too few potentially associated sites will be detected. To remedy the deficiency the second test is invoked. MATA's **second-test** filters most of the false-positive prediction that result from a large α (column 4). With increasing α the number of associated sites corresponding to the neighbourhood system increases, reaching its maximum for $\alpha = 0.1$, then the number decrease. Finally, for $\alpha = 0.45$ no potential associations are detected. The last column shows the accumulation of true associations with increasing α . If α is larger than 0.3 no new associations are detected. Thus, simply testing for associations using one fixed α leads to an underestimation of true inter-site associations.

Fig. 4.5 illustrates the effect of first conducting MATA's **first-test** and subsequently applying MATA's **second-test**. The top row displays the circle plots resulting from the set of potential associations for $\alpha = 0.05, 0.1, 0.15$ and 0.2. The irregular structure of the graphs as compared to the circle plot showing the neighbourhood system (in the last row at the right of Fig. 5.2) clearly shows the large proportion of false positive predictions. The bottom row shows the cleaning effect of MATA's **second-test**. Different values of α and the corresponding β parameter lead to different inter-site associations. The final result of MATA prediction is displayed at the last row of Fig. 5.2. In summary, MATA suggested a

total of 41 associated pairs of sites compared to 54 base pairs in the sequence. Two of the 41 associated pairs are false positive predictions.

The accuracy of the prediction is enhanced when associated sites are excluded and MATA is applied to the reduced alignment. Here, we found three additional associations (blue in Fig. 5.2). Repeating this for the again reduced alignment did not yield additional predictions. Thus, we obtained in total 44 true positive base-pairs and two false positive predictions.

Influence of the Tree

We test the influence of the underlying tree on the capability to detect inter-site associations in an alignment. We generated one tree topology using the ape package (Paradis *et al.*, 2004; R Development Core Team, 2004). For this topology branch lengths were randomly drawn from a uniform distribution in the interval 0 to 1. The resulting tree with its fixed branch lengths was then rescaled to arrive at trees with mean branch length 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. The six trees are called $T_{0.05}, T_{0.1}, T_{0.2}, T_{0.3}, T_{0.4}, T_{0.5}$. For each tree 100 simulated data were generated and subsequently analysed with MATA.

More precisely, we counted the number of true positive associated sites and the number of false positive associated sites and display the corresponding box-plots (Fig. 5.3). First we observe, that the number of false positive predictions is virtually unaffected by the branch lengths of the trees. The median ranges between one and two. Moreover in 20%-41% of the simulated data no false positives were observed. Thus, false positive predictions are small with MATA and do not depend on the branch lengths.

The situation is different if we analysed the number of true positives. We note that for trees with short average branch lengths the rate of true positives is small. In $T_{0.05}$ the median is zero and the maximum number of associations equals three, compared to 54 base-pairs in the molecule. Only if the average branch lengths exceeds 0.3 we observed an

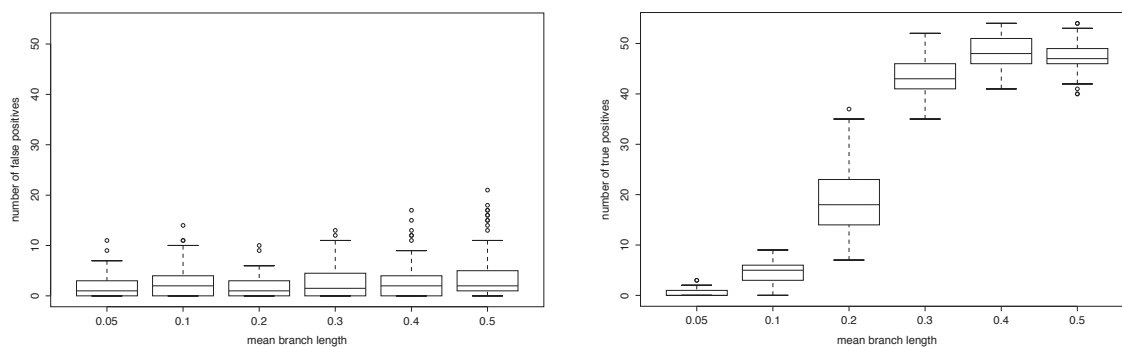


Figure 5.3: Number of Detected True Positive and False Positive Associations vs Mean Branch Length. Investigated are trees with mean branch length of 0.05, 0.1, 0.2, 0.3 and 0.5 substitutions per site. Lines in the box display the lower quartile, the median and the upper quartile. The whiskers are set to 1.5 times the interquartile range.

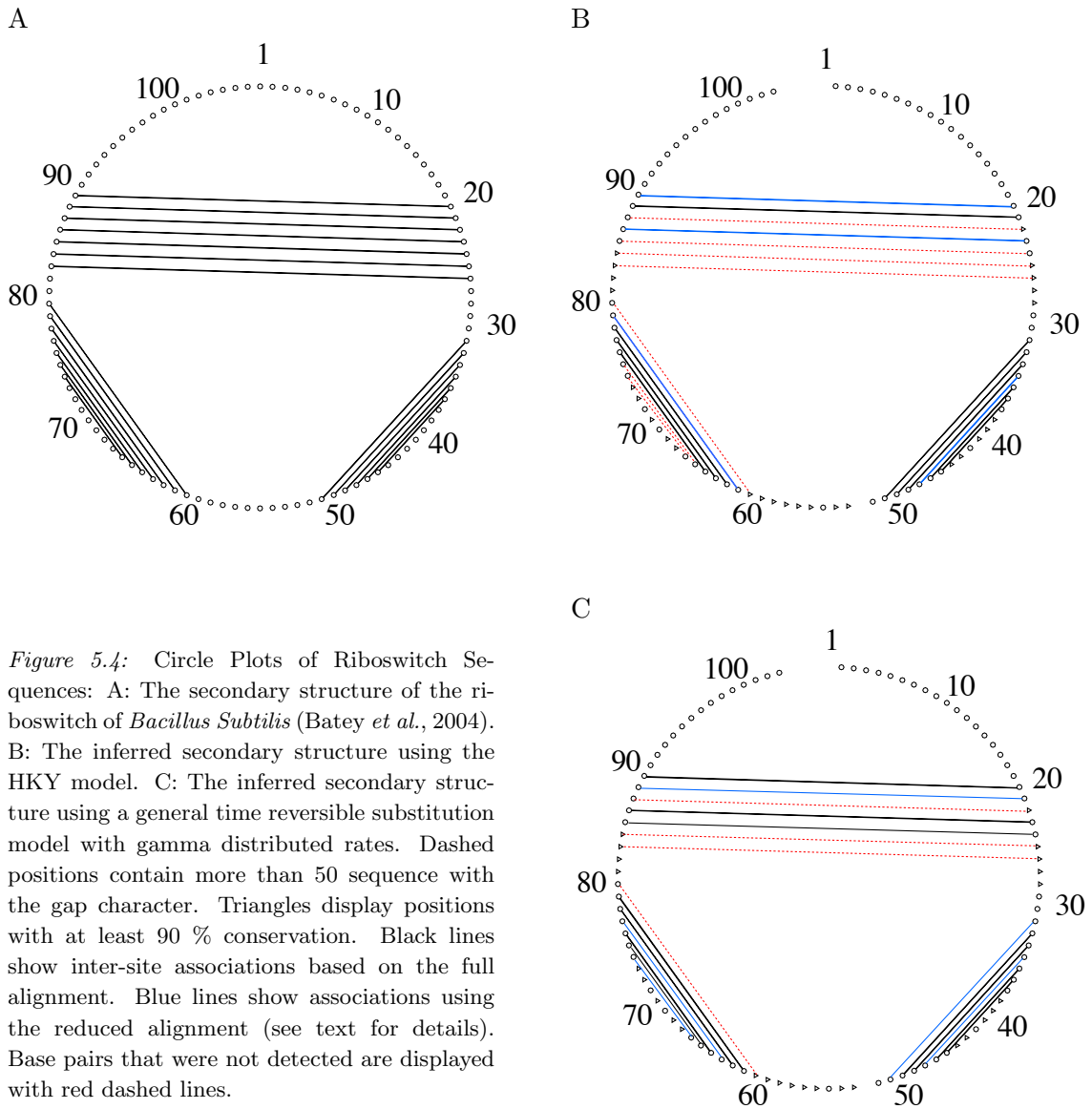


Figure 5.4: Circle Plots of Riboswitch Sequences: A: The secondary structure of the riboswitch of *Bacillus Subtilis* (Batey *et al.*, 2004). B: The inferred secondary structure using the HKY model. C: The inferred secondary structure using a general time reversible substitution model with gamma distributed rates. Dashed positions contain more than 50 sequence with the gap character. Triangles display positions with at least 90 % conservation. Black lines show inter-site associations based on the full alignment. Blue lines show associations using the reduced alignment (see text for details). Base pairs that were not detected are displayed with red dashed lines.

appreciable number of association, e.g. a median of 48 for $T_{0.4}$. The low detection rate of true positives is due to the lack of power. For a tree with zero branch lengths no statistical method has the ability to detect associated sites, since no substitution occurred. With increasing branch length the number of substitution accumulates and substitutions are reflected in the multiple alignment. Therefore the accumulation of different substitution patterns allows a better detection of associated sites.

Performance of MATA on a Purine Riboswitch

So far, we have only discussed the performance of MATA using artificially generated data. Here we show that MATA also works on biological data. As an example, we want to retrieve the inter-site association from a riboswitch molecule. Fig. 5.4A displays the secondary structure of the *Bacillus subtilis* riboswitch (Batey *et al.*, 2004).

Based on the multiple alignment, we applied MATA assuming in the first instance an HKY model (Hasegawa *et al.*, 1985). We detected nine associated pairs (black lines Fig. 5.2B). No false positive associations were suggested.

After having deleted the corresponding 18 sites from the alignment, we re-applied MATA to the shortened alignment. Recall that for the reduced alignment the tree is reconstructed again. The reduced alignment provided four additional associations (blue line Fig. 5.4B). Any further associations were not detected. In summary, we obtained 13 of 20 base pairs. The suggested structure reflects by and large the secondary structure of the *Bacillus subtilis* (Batey *et al.*, 2004) molecule. However, seven base pairs were not detected.

To investigate the influence of the substitution model on the ability to detect inter-site associations, we reapplied MATA to the riboswitch alignment using the general time reversible substitution model and gamma distributed rates (Tavaré, 1986; Rodriguez *et al.*, 1990). The results are displayed in Fig. 5.4C. Based on the more realistic substitution model we obtained a total of 16 associations and again no false positives were detected. Thus, only four inter-site associations were undetected. The four missed base pairs could not be detected, because for each base pair at least one site is constant. For example, the pair formed by alignment sites 25 and 84 consists entirely of U:A base pairs. Such inter-site associations between constant sites are impossible to detect by any statistical procedure, but easily by visual inspection.

More importantly the biological data shows that the model of sequence evolution also has an impact on the outcome of MATA. Thus, one should fit the best model of sequence evolution (Posada and Buckley, 2004) to the alignment before running MATA.

Conclusions

We introduced MATA as a method to detect inter-site associations from a sequence alignment using the parametric bootstrap. In contrast to other comparative methods with and without phylogeny (Chap. 2, Sec. 2.3.2), one main advantage of MATA is its self-consistency, i.e. no threshold needs to be set in advance to assess significance. It readily provides measurements of accurate thresholds corresponding to the given alignment MATA introduces two statistics: MATA's **first-test** and MATA's **second-test**. MATA's **first-test** suggests potential inter-site association. It measures the deviation of the frequency of the pair of nucleotides from its expected frequency. However, base pairing may not be the only cause for the deviation. Other causes e.g. unequal rates of evolution, inaccurate tree topology and branch length or other constraints at the individual sites may influence this

type of statistics.

To reduce the number of false positive associations **MATA's second-test** is applied. First the deviation from the expected frequency of nucleotides is calculated for each site and for each root nucleotide at the site. Then the standard deviation of the corresponding p -values is used as an heuristic to detect sites that cause false positive associations. If this deviation is small, then, the root nucleotide of this site cannot be accurately determined and these sites are excluded from the analysis. Besides, **MATA's second-test** may be too liberal in rejecting sites, especially when associated sites are constant as found in the riboswitch alignment. Only if some variability occurs in base paired sites comparative tests have a chance to suggest associations. To overcome this drawback of comparative methods we recommend a careful posterior analysis of sites that are close to associated sites. Then, additional arguments may be helpful to further exclude inter-site associations. In the case of RNA structure thermodynamics will also lead to an inclusion of undetected base pairs (e.g. helix Fig. 5.2). In any case, **MATA** can retrieve a rough picture of the neighbourhood system of a PS directly from an alignment and a tree. We have shown that the ability of **MATA** to detect associations depends on the branch lengths of the underlying tree. For trees with short branch lengths the detection is harder than for trees with long branch lengths. Also, the model of sequence evolution influences the predictive power of **MATA**. The riboswitch example shows this impressively.

Finally, repeating **MATA** in an iterative procedure also leads to an increase in the number of inter-site associations, without inflating the number of false positive predictions. Because **MATA** does not utilize thermodynamic arguments to infer structural constraints it is applicable to any kind of multiple sequence alignment. Here we have only demonstrated its application to RNA structure reconstruction. **MATA** could also work for amino acid sequence alignments. Then, however, many aligned sequences are necessary to infer the structure. Preliminary comparison to other programs shows that the performance is not as good as **ConStruct** (Lück *et al.*, 1996; Wilm *et al.*, 2008) and **RNAalifold** (Hofacker *et al.*, 2002a) (with and without thermodynamics, see Chap. 2). However, **MATA** offers a different viewpoint.

5.5 Double Life

Indeed, MATA in itself has to be improved and it gives rise to interesting biological and statistical questions. However, MATA's principle appears to be a useful complement to other existing tools. It is a heuristic mixture of model testing and simulations. Although, the statistic is not exactly, MATA balances between double associations in an alignment, called ancestral and functional correlations. False positive prediction are small with MATA, however, with low sensitivity. So far, MATA measures the deviation of the frequency of the pair of nucleotides from its expected frequency given an estimated pair at the root sequence and the tree. The generality of the methods and the simplicity of the model are an advantage in the sense of finding any correlations. Indeed, a rejection of the null hypothesis could have other causes than stem dependencies. A greater focus on the filtering procedure, which removes false positives, is necessary. Combining MATA's framework with advanced covariance measures has the power to improve other comparative methods in general (Chap. 2, Sec. 2.3.2), with a gain of selectivity in an automatic manner based on an iterative procedure and including phylogenetic information. Generally, MATA can be used for measurements of accurate thresholds of alignments for structure prediction programs. A combination with thermodynamic measures can be used to improve both current and future programs for secondary, as well as tertiary and quarternary structure prediction. In *RNAalifold* as well as *ConStruct* the values to quantify the contributions of compensatory and consistent mutations are chosen arbitrary. A combination of MATA's framework with these programs is a promising way to tackle this problem. However, parametric bootstrap with a null hypothesis taking overlapping dependencies into account would be necessary, see Fig. 5.5 . A program, which simulates overlapping dependencies along a phylogenetic tree directly combined with a consensus folding algorithm, is presented in the next chapter.

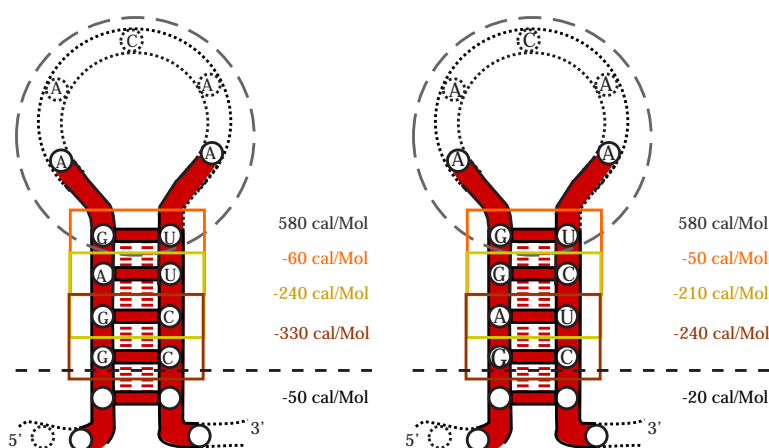


Figure 5.5: An example for the influence of the dinucleotide content on the folding stability. Although both helices have the same number and the same base-pairs the mfe is different, because the order of the two base pairs in the middle is changed. In the nearest neighbour model (Chap. 2, Sec. 2.1.3) energies are not assigned to single base-pairs but rather to neighbouring base-pairs that stack on each other.

Chapter 6



Kurt Gödel (1906-1978)

SISSIZ's Model Aspect

All models are wrong: icons of glamour and "perfect" beauty . . .

Brigitte Bardot is such a controversial figure - and she recalls Mother Courage. It's very interesting that she has this quality of being a model for everything. And yet she is always deconstructing her own role, . . . A model is not straightforward, not so clear: it is made out of circumstances, including your own perspective. (Rosemarie Trockel)

TO ANNE

Structure estimation as well as phylogeny inference is never free of assumptions. In phylogeny, for example, assumptions about the evolutionary process that produced the observed data form a *model* of character evolution which then yields estimates of evolutionary relationships. If a *conceptual model* is transformed into a *mathematically explicit formal model* of character change and applied to sequence data, then this is referred to as the model in phylogenetic literature. Practical introductions to formal models and model selection methods in molecular phylogenetics are given in several reviews (Whelan *et al.*, 2001; Sullivan and Swofford, 1997; Kelchner and Thomas, 2007; Posada and Buckley, 2004).

Regardless of this ongoing debate in phylogeny, we address the problem of finding an adequate null statistic under a background model for comparative noncoding RNA predictions using a simple and fast heuristics and a distance based approach. Doing this and including the assumptions about structure from the RNA community, the model question can be seen from another viewpoint, namely in the sense of *background or so-called null models*: Any experiment is only as good as its controls. What is true for experimental biology clearly also holds in the field of computational biology. The value of even the most sophisticated algorithm remains unclear if the significance of the results cannot be assessed properly. ¹

¹SISSIZ is a collaboration with S. Washietl.

6.1 Background Models for RNA Gene Prediction

Comparative genome analysis is currently a widely used strategy to detect and annotate noncoding RNAs (ncRNAs) (Griffiths-Jones, 2007; A. F. Bompfünowerer Consortium *et al.*, 2007). In the past few years a series of different algorithms have been developed that predict functional ncRNAs on the basis of conserved secondary structure (Rivas and Eddy, 2001; Coventry *et al.*, 2004; Washietl and Hofacker, 2004; Washietl *et al.*, 2005c; Pedersen *et al.*, 2006; Yao *et al.*, 2006; Uzilov *et al.*, 2006; Torarinsson *et al.*, 2006). Some of these methods have been used to discover novel ncRNAs on a genome wide scale (Washietl *et al.*, 2005b; Pedersen *et al.*, 2006; Missal *et al.*, 2005, 2006; Rose *et al.*, 2007). In combination with experimental verification (microarray, RT-PCR, Northern blot) these methods could successfully uncover many novel ncRNAs (Axmann *et al.*, 2005; Weile *et al.*, 2007; del Val *et al.*, 2007; Mourier *et al.*, 2007; Sandmann and Cohen, 2007; Washietl *et al.*, 2007b). However, in particular in large vertebrate genomes the signal-to-noise ratio of true predictions and false positives is thought to be relatively low (Washietl *et al.*, 2007b). In a recent paper, Babak and colleagues demonstrated that comparative ncRNA gene finders are strongly biased by the genomic dinucleotide content leading to an excess of false predictions (Babak *et al.*, 2007). Especially methods that are based on a thermodynamic folding model are sensitive to this effect: In the so-called nearest neighbour model (Chap. 2, Sec. 2.1.3) energies are not assigned to single base-pairs but rather to neighbouring base-pairs that stack on each other. As a consequence, the folding stability of genomic sequences does not only depend on the mononucleotide content but also the dinucleotide content (see Fig. 5.5).

To assess the significance of predicted structures, e.g. to estimate the false discovery rate in a genomic screen for ncRNAs, one should therefore compare the genomic predictions to the results obtained on randomised data with the same dinucleotide content. In the case of single sequences, there are well known and widely used algorithms to generate dinucleotide controlled random sequences either by shuffling or first order Markov chain simulation (Altschul and Erickson, 1985; Clote *et al.*, 2005). However, there is currently no algorithm to randomize multiple sequence alignments preserving the dinucleotide content. Babak and colleagues (Babak *et al.*, 2007) added the conservation of dinucleotides as an additional constraint to the commonly used (mononucleotide) shuffling algorithm `shuffle-aln.pl` (Washietl and Hofacker, 2004) and applied it to pairwise alignments. Their approach corresponds to a heuristic (Workman and Krogh, 1999), that is very inefficient as only a small subspace of the whole permutation space is covered. The heuristic exchanges only positions that have the same neighbours left and right. For the short sequence ACAGCCAA for example not a single permutation can be found that way. However, there are 11 permutations according to the Altschul & Erikson algorithm (Altschul and Erickson, 1985). But even a more efficient shuffling algorithm will soon run into difficulties on multiple alignments. Unless two neighbouring columns are 100% conserved, there are several different dinucleotide pairs in these columns. It is therefore impossible to exactly

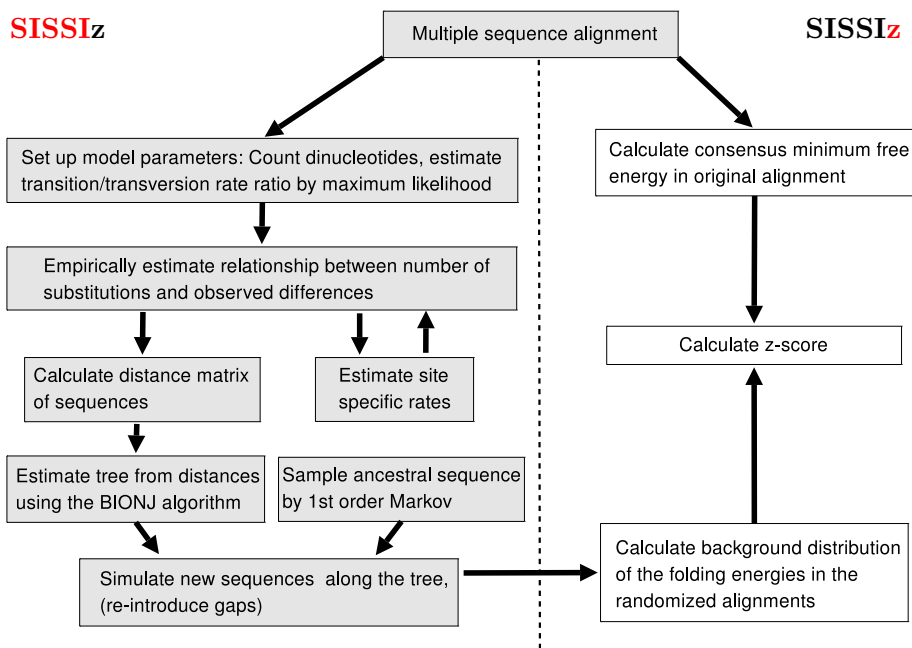


Figure 6.1: Overview of the algorithm SISSIz. Left: the steps of the randomisation procedure are shown. We extend the SISSI framework of Chap. 3. Right: In combination with RNAalifold consensus folding the randomisation procedure can be used to calculate z -scores and to predict significant RNA structures. See text for details.

preserve the dinucleotide content as in the single sequence case.

The present chapter addresses the problem of so-called randomisation problem with simulations using SISSI. In particular, we simulate alignments of a given dinucleotide content. We present a substitution model that captures the neighbour dependencies and other important alignment features except the signal in question. We describe a time efficient way to estimate a tree under this model that we use as a guide to simulate alignments of the desired properties. This new control strategy is tested on genomic alignments and the effect on thermodynamic RNA structure predictions is studied. In addition, we directly combined the new null model with the RNAalifold (Hofacker *et al.*, 2002a) consensus folding algorithm giving a new variant of a thermodynamic structure based RNA gene finding program that is not biased by the dinucleotide content.

A Short Overview: SISSIz & SISSIz

Fig. 6.1 gives a short outline of SISSIz. The whole randomisation procedure at the left side and its combination with RNAalifold consensus folding to calculate z -scores and to predict significant RNA structures at the right side.

We start by parametrisation of our model: we count the dinucleotides and calculate the corresponding stationary trinucleotide frequencies. A transition/transversion rate ratio for the alignment is estimated using maximum likelihood under a HKY+ Γ model. Having

set these parameters, we empirically estimate the relationship between substitutions and observed differences with equal rates for each site. This first estimate is used to calculate the site-specific rates, which are then used for the second estimation. In the next step, the pairwise distances between all sequences are calculated. For the calculation of the site-specific rates and the pairwise distances gap characters are treated in a special way as missing data. From the distance matrix a tree is built using the BIONJ algorithm (Gascuel, 1997). An ancestral sequence is generated from a first order Markov model parametrised according to the dinucleotide frequency in the original alignment. This is used as a starting sequence for the simulation that is guided by the tree. Finally, the gap pattern of the original alignment is introduced into the simulated one. Fig. 6.5 shows our rRNA example and two randomised versions obtained by this procedure. The simulated alignments can be used to calculate z -scores and predict significant RNA structures in combination with `RNAalifold`, outlined at the right side of Fig. 6.1 and explained on page 122 in detail. In the next section we present the randomisation procedure using `SISSI`.

6.2 Randomising Genomic Alignments

Requirements for an Adequate Null Statistic under a Background Model

An optimal null model preserves all the features of the original data with the exception of the signal under question that needs to be removed efficiently. In our case, the data are multiple alignments of homologous sequences and the signal of interest is an evolved RNA secondary structure. Correlations arising from base-pairing patterns need to be removed. Currently, alignments are usually randomised by shuffling the alignment columns (see ref. in Washietl and Hofacker (2004) for a discussion of this method). Although the shuffling approach has its limitations and considering dinucleotides seems difficult, it is an appealing approach because it is relatively simple, fast, and extremely conservative. Changing the order of the columns does not change the mutational patterns within the columns and thus the underlying phylogenetic tree is exactly preserved.

In this chapter we attempt to simulate new alignments from scratch. Even the most sophisticated model cannot capture all evolutionary processes and therefore a simulation approach will inevitably change the original data more than shuffling does. So much care has to be taken to preserve all the relevant characteristics of the data. To qualitatively assess the most important parameters that need to be considered in our model, we performed a series of simulation experiments. Using a simple tree with four taxa we simulated alignments under the HKY evolutionary model (Hasegawa *et al.*, 1985), described in Chap. 2, Sec. 2.2.1. We systematically varied model and tree parameters to study how they affect thermodynamic RNA consensus structure predictions in the alignment. We used `RNAalifold` (Hofacker *et al.*, 2002b) to predict consensus secondary structures which is the basis of the `AlifoldZ` (Washietl and Hofacker, 2004) and `RNAz` (Washietl *et al.*, 2005c) gene finders.

Not surprisingly, base composition is one of the parameters affecting the predicted folding energies strongest (Fig. 6.2A). High G+C content leads to more stable RNA predictions, while high A+T content gives less stable predictions. As mentioned before and in fact the main motivation of this chapter, also dinucleotide content affects folding energies. We used SISSI to simulate alignments of the same mononucleotide content but varying dinucleotide content. Fig. 6.2B shows for example that a three times enriched ApT content lead to more stable predictions. The excess of some other dinucleotides like for example GpT can cause the opposite effects leading to less stable predictions.

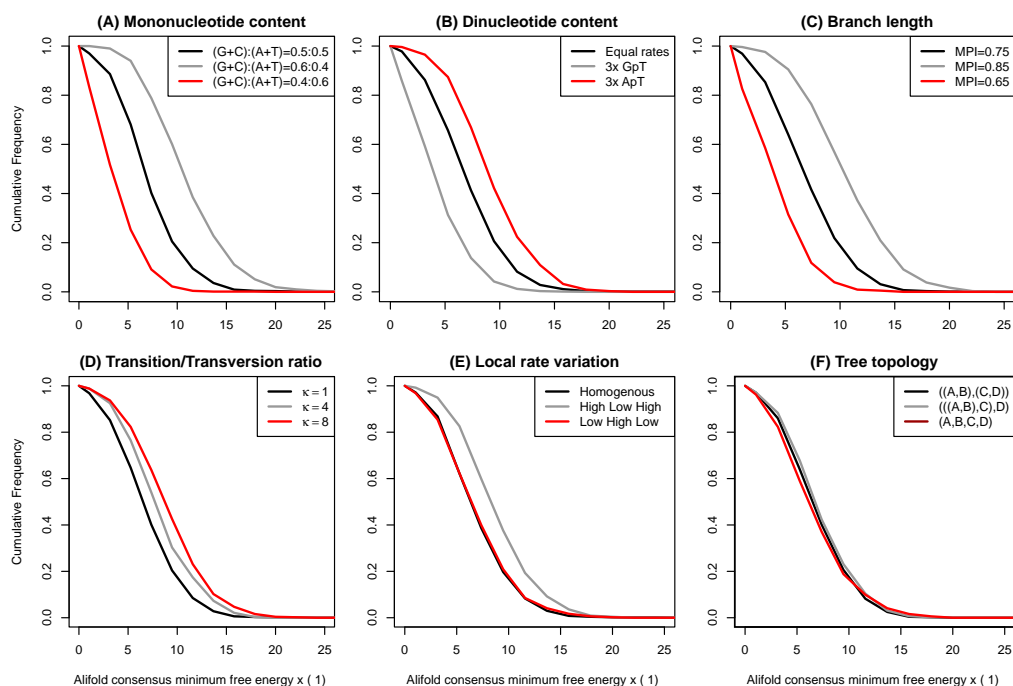


Figure 6.2: Parameters effecting thermodynamic consensus RNA structure predictions. As a basic parameter set we used equal base frequencies of 0.25, a transition/transversion rate ratio $\kappa = 1$, and the following tree $((A:0.09, B:0.09):0.09, (C:0.09, D:0.09):0.09)$. One parameter was varied at a time while others were kept constant. If necessary branch lengths were adjusted to keep a mean pairwise sequence identity (MPI) of 0.75 ± 0.01 . 1000 alignments of length 80 were simulated under each condition. Cumulative histograms for the `RNAalifold` consensus folding energies are shown. Please note that we plot negative minimum free energies, i.e. higher values correspond to more stable folds. **(A)** Base frequencies were varied to get high and low G+C content. **(B)** Two specific dinucleotide frequencies were elevated 3-fold while the mononucleotide content was kept constant. **(C)** Branch lengths were equally scaled to produce alignments with lower or higher MPI identity than for the basic tree. **(D)** The transition/transversion rate ratio was varied. $\kappa = 1$ means equal rates, while $\kappa > 1$ gives more transition than transversions. **(E)** The alignment of size 80 was divided into a central block of 40 and two flanking regions of 20. We set 100% conservation in the central block and low conservation in the flanks (rate “high-low-high”) and the other way round (“low-high-low”). The total average MPI was always 0.75. **(F)** We tested all possible topologies of this 4 taxa tree and adjusted the branch lengths to give a MPI of 0.75. For one given topology, all the branch lengths were of the same length.

Another major parameter that needs to be controlled is the diversity of the alignment. Variation of the branch lengths of the tree gives alignments with different sequence diversity which we usually measure as the mean pairwise sequence identity (MPI, also sometimes referred to as average pairwise sequence identity APSI). High diversity (i.e. low MPI) makes it difficult to predict a consensus structure. On the other hand, almost perfectly conserved sequences fold readily in some random structure even if there is no natural RNA structure present. Therefore we observe a strong dependency on the MPI (Fig. 6.2C).

One well known characteristic of natural mutation processes are the different rates for transitions and transversions (Felsenstein, 2004). Interestingly, this also affects the consensus structure predictions. A model with equal transition/transversion rates (parameter $\kappa = 1$ in the HKY model) gives less stable predictions than a model with more realistic rates (e.g $\kappa = 4$, Fig. 6.2D). This parameter affects the type of column patterns observed in the simulated alignments affects how well they can form consensus base pairs.

Natural mutation processes are not homogeneous across all sites, in particular in functional genomic regions. It was observed previously that mutation patterns within an alignment can affect structure predictions (Washietl and Hofacker, 2004). For example, an alignment containing a 100% conserved block with low mutation rate that is flanked by highly divergent regions of high mutation rate can have different folding energies compared to an alignment with homogeneous rates but the same overall MPI (Fig. 6.2E). The same is true for patterns of insertions and deletions which was also already discussed in reference (Washietl and Hofacker, 2004) and which we do not show here explicitly again.

We also tested the effect of different tree topologies, but did not find a significant influence of this parameter at least in our four taxa example.

Taken together, an accurate randomisation procedure needs to generate alignments that preserve (i) mono- and dinucleotide content, (ii) mean pairwise sequence identity, (iii) transition/transversion rate ratio (iv) site-specific mutation rates, and (v) gap patterns.

In the next section we extend the SISSI model of Chap. 3 to a model that is capable of simulating alignments under these constraints.

6.2.1 A Specific SISSI Model

Sequence evolution is usually described by a time-continuous Markov process (Tavaré, 1986; Felsenstein, 2004). The most commonly used models assume that all sites of a sequence evolve independently from each other rendering it impossible to model dinucleotide dependencies between neighbouring pairs. Various evolutionary models have been proposed in the past years to overcome this limitation (Jensen and Pedersen, 2000; Duret and Galtier, 2000; Pedersen and Jensen, 2001; Arndt *et al.*, 2003; Robinson *et al.*, 2003; Siepel and Haussler, 2004; Lunter and Hein, 2004; Christensen, 2006). We make use of the introduced framework SISSI (SImulating Site-Specific Interactions) of Chap. 3. SISSI allows to define site dependencies of arbitrary complexity in the form of a neighbourhood

our instantaneous rate matrix Q_k are defined by the mathematical requirement that the sum of each row is zero.

The entries of Q_k are thus given by

$$Q_k(\mathbf{s}_k, \mathbf{y}) = f_k \cdot \begin{cases} r_{(\mathbf{s}_k, \mathbf{y})} \cdot \pi_k(\mathbf{y}) & \text{if } H(\mathbf{s}_k, \mathbf{y}) = 1 \text{ and } x_k \neq y_2 \\ - \sum_{\substack{\mathbf{z} \in \mathcal{A}^3 \\ \mathbf{z} \neq \mathbf{s}_k}} Q_k(\mathbf{s}_k, \mathbf{z}) & \text{if } H(\mathbf{s}_k, \mathbf{y}) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (6.2)$$

where $\pi_k(\mathbf{y})$ is the stationary frequency of \mathbf{y} and the Hamming distance $H(\mathbf{s}_k, \mathbf{y})$ counts the number of differences between the sites of the triplets \mathbf{s}_k and \mathbf{y} .

In principle, we can choose any rate for the parameters $r_{(\mathbf{s}_k, \mathbf{y})}$. However, based on the requirement that we want to use the dinucleotide content as stationary distribution, we choose $r_{(\mathbf{s}_k, \mathbf{y})}$ so that the model becomes reversible. For our application, we use a transition/transversion rate ratio and set $r_{(\mathbf{s}_k, \mathbf{y})} = \kappa$ for transitions and $r_{(\mathbf{s}_k, \mathbf{y})} = 1$ for transversions. The restriction that a substitution is only possible at site k leads to sparse rate matrices. Q_k has only $|\mathcal{A}|^4$ non-zero entries. Hence, we can write Q_k in the form of 16 submatrices, which describe the admissible substitutions for site k depending on the left y_1 and right y_3 neighbours,

$$\begin{array}{c} y_1 A y_3 \quad y_1 C y_3 \quad y_1 G y_3 \quad y_1 U y_3 \\ \begin{array}{c} y_1 A y_3 \\ y_1 C y_3 \\ y_1 G y_3 \\ y_1 U y_3 \end{array} \begin{pmatrix} * & \pi_{y_1 C y_3} & \kappa \pi_{y_1 G y_3} & \pi_{y_1 U y_3} \\ \pi_{y_1 A y_3} & * & \pi_{y_1 G y_3} & \kappa \pi_{y_1 U y_3} \\ \kappa \pi_{y_1 A y_3} & \pi_{y_1 C y_3} & * & \pi_{y_1 U y_3} \\ \pi_{y_1 A y_3} & \kappa \pi_{y_1 C y_3} & \pi_{y_1 G y_3} & * \end{pmatrix} \end{array} \quad (6.3)$$

Finally, we scale Q_k such that the number of substitutions d_k equals 1:

$$d_k = - \sum_{\mathbf{z} \in \mathcal{A}^3} \pi_k(\mathbf{z}) \cdot Q_k(\mathbf{z}, \mathbf{z}) = 1. \quad (6.4)$$

and thus the total instantaneous substitution rate for a sequence \mathbf{x} can be written as the sum over each rate $Q_k(\mathbf{s}_k, \mathbf{s}_k)$ multiplied with the site-specific scaling factor f_k , with $k = 1, \dots, l$ (Fig. 6.3),

$$q(\mathbf{x}) = - \sum_{k=1}^l f_k \cdot Q_k(\mathbf{s}_k, \mathbf{s}_k). \quad (6.5)$$

Without dependencies on the neighbours, Q_k is of dimension 4×4 and the model reduces essentially to a HKY model with a specific rate for each site. We use this mononucleotide variant later in this chapter for testing and comparison to the dinucleotide model.

Simulation

For the simulation, we used the same algorithm as described in Chap. 3 with some modifications. During the simulation process, we pick a site k according to its relative mutability

$$P(k) = \frac{|f_k \cdot Q_k(\mathbf{s}_k, \mathbf{s}_k)|}{q(\mathbf{x})}, \quad (6.6)$$

and for the chosen site k , the nucleotide x_k will be replaced by a new nucleotide $y_2 \in \mathcal{A}$ from the corresponding triplet \mathbf{y} with probability:

$$P(x_k \rightarrow y_2) = \frac{f_k \cdot Q_k(\mathbf{s}_k, \mathbf{y})}{|f_k \cdot Q_k(\mathbf{s}_k, \mathbf{s}_k)|} = \frac{Q_k(\mathbf{s}_k, \mathbf{y})}{|Q_k(\mathbf{s}_k, \mathbf{s}_k)|} \quad (6.7)$$

In the general SISSI framework Q_k needs to be updated for all k sites every time one nucleotide in \mathbf{x} is substituted. However, in our case we can use the same instantaneous rate matrix Q_k for each site. As a consequence, we can fix $q(\mathbf{x})$ and do not need to sum over each rate of the site, which improves the running time of the algorithm.

6.2.2 Parameter Estimation

Ideally, all parameters are estimated simultaneously within a maximum likelihood framework. One problem is the high number of parameters since we want to estimate a specific rate for each site. A more fundamental issue is, however, that our model includes overlapping dependencies which breaks the independence assumption necessary for basic maximum likelihood estimation. Other possible techniques like Markov chain Monte Carlo in a Bayesian framework are not a viable alternative either. Speed is a critical issue as the algorithm is meant to be applied to data on a genome wide scale.

Facing these difficulties, we have developed heuristic approximations to estimate the parameters and use a distance based approach to estimate the tree. The method is fast and yet surprisingly accurate for our application.

Equilibrium Frequencies

The stationary frequencies of our model are set in a way that in equilibrium we obtain a dinucleotide frequency that is the same as the dinucleotide content of the alignment to be randomised. To this end, we first count the dinucleotide frequencies as an average of all

sequences in the original alignment. Then, we calculate the corresponding trinucleotide frequencies needed for Q_k as a function of the single and dinucleotide frequencies using an approximation based on simple conditional probabilities (Arndt *et al.*, 2003; Duret and Galtier, 2000):

$$\pi(\alpha\beta\gamma) = \frac{\pi(\alpha\beta)\pi(\beta\gamma)}{\pi(\beta)} \quad (6.8)$$

where $\pi(\alpha\beta\gamma)$ are the trinucleotide frequencies, $\pi(\alpha\beta)$ and $\pi(\beta\gamma)$ the counted dinucleotide frequencies and $\pi(\beta) = \sum_{\alpha} \pi(\alpha\beta) = \sum_{\alpha} \pi(\beta\alpha)$ with $\alpha, \beta, \gamma \in \{A, C, G, U\}$.

Fig. 6.4A shows an example of the dinucleotide frequency distribution of 1000 simulated alignments. We counted the dinucleotide frequencies of an alignment of 7 5.8 rRNA sequences and set the trinucleotide parameters of our model accordingly. On average, we get the same dinucleotide frequencies in the simulated alignments as in the original one.

Transition/transversion rate ratio

The transition/transversion rate ratio κ is a parameter in our model that cannot be simply counted as in the case of the dinucleotide frequencies, or determined like the branch lengths. Given that the influence of this parameter is not that critical as for example the branch length or base composition (see Fig. 6.2), one possibility might be to use a fixed transition/transversion ratio if a reasonable average value is known for the genome at hand. Alternatively, we found that a good estimate can be obtained by using maximum likelihood on an independent mononucleotide model. We used here the HKY model with Γ -distributed rates which is closest to our dinucleotide model.

Gaps

So far, gaps have been ignored completely. There are evolutionary models including deletion and insertions (Thorne *et al.*, 1991, 1992; Metzler, 2003; Miklós *et al.*, 2004; Fleißner *et al.*, 2005) and, we have suggested an algorithm in Chap. 3, Sec. 3.4, to combine the insertion-deletion dynamics with our model. However, this does not appear practical for randomising genomic alignments. Existing algorithms for joint estimation of phylogenies and alignments are not only very time-consuming (Fleißner *et al.*, 2005), it also seems difficult to estimate reasonable indel model parameters on relatively short alignment blocks which hold only little information. Moreover, alignment programs produce gap patterns that do not necessarily reflect phylogenetically reasonable insertion/deletion events and thus cannot always be captured by an idealized model that is motivated by evolutionary processes and ignores algorithmic idiosyncrasies of alignment programs.

So we follow here a very pragmatic strategy by Washietl and Hofacker (2004): We keep exactly the same gap pattern in our randomised alignments as in the original alignment. To this end, we simply treat gaps as missing data and simulate nucleotide characters for

the gapped positions. This is done in a way that the overall characteristics are not changed when they are replaced with gaps again at the end. ²

Distances and Tree Construction

To build a distance based tree, we first estimate the number of substitutions that have taken place between two sequences. In other words, we estimate the genetic or evolutionary distance d from the Hamming distances p under our model. To estimate the relationship between d and p , we simulate sequence pairs separated by different branch lengths d and calculate the corresponding Hamming distances p (Fig. 6.4B). We fit an exponential function to this curve:

$$p = \hat{a} \cdot (1 - e^{\hat{b} \cdot d}) \tag{6.9}$$

Using this function, all pairwise distances d are calculated for the sequences in the original alignment. ³

From this distance matrix a tree is constructed using the BIONJ algorithm (Gascuel, 1997). BIONJ is a variant of the well known neighbour joining algorithm and currently one of the most accurate algorithms for distance based tree building.

Given that the distances and the tree are accurately estimated, we observe on average the same mean pairwise identity in the simulated alignment as in the original one. Fig. 6.4C shows the distribution of MPIs of 1000 simulations of our example rRNA alignment. The average MPI of the simulations is exactly the same as the MPI 0.73 of the original alignment.

²When counting the dinucleotide content, dinucleotides including a gap (N-, -N, --) are ignored. During simulation, gap positions are filled with nucleotides and gaps are re-introduced afterwards. Note that this way, if two nucleotides N₁ and N₂ are separated by a gap (e.g. N₁---N₂) the dinucleotide N₁N₂ is not in equilibrium. Depending on how gaps are treated in the downstream analysis this might be or might not be of concern. In any case, since not every gap position but only every gap *opened* is affected, this (potential) error is generally very small for reasonable alignments. So we did not consider correcting for this effect which would require reconstructing the gap history and setting lineage specific neighbourhood systems.

³**Distances above the Level of Saturation:** When calculating genetic distances between two sequences the problem may occur that the observed number of differences is higher than the level of saturation. We found that this problem becomes severe when considering site-specific rates that generally lead to much lower levels of saturation (cf. Fig. 6.4B). We use the simple trick of Sec. 3.4 to overcome this limitation. We add additional sites during the simulation with site-specific rates that correspond to the average of the whole alignment (i.e. $\langle p_k \rangle$ is set to 1-MPI in Equ. 6.11 for all these additional sites). They act as “buffer sites” that reduces the number of mutation events that repeatedly hit the same sites of high rate leading to many double substitutions. As a consequence, the overall level of observed differences is higher and we do not run into problems building the distance matrix. In the end, the sites are removed again and since the relative rate ratios between the sites remained unchanged, we get the desired site-specific mutation patterns.

Site-Specific Rates

Setting different mutation rates at different sites gives us the possibility to preserve natural mutation patterns of the original alignment. The problem of finding accurate site-specific rates is illustrated in Fig. 6.4D. For each site in the alignment, the MPI of this site is plotted against the average MPI observed in the simulated alignments on the same site.

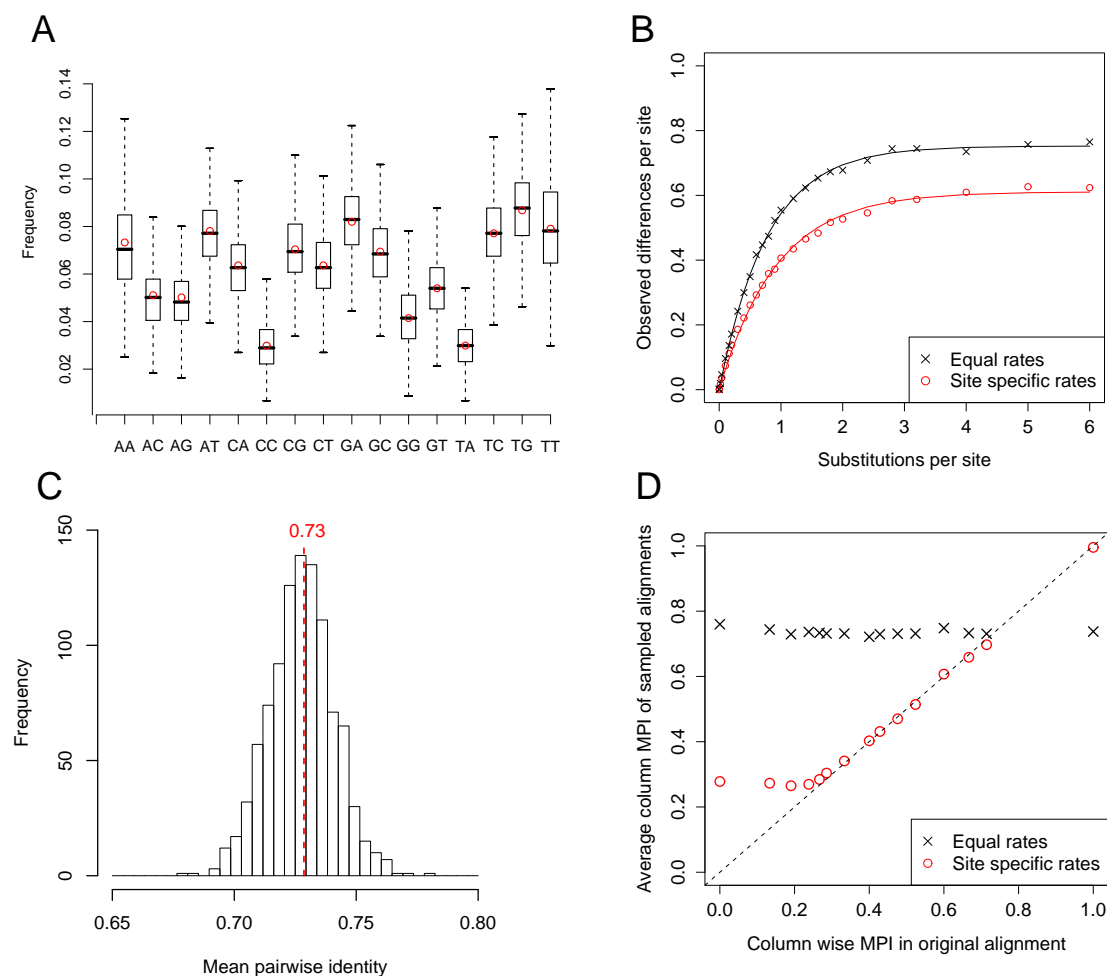


Figure 6.4: Key concepts of the algorithm shown on an example alignment of 5.8S rRNA. (A) Distribution of dinucleotide frequencies of 1000 simulated alignments are shown as box-plots (the line in the box indicates the median, the borders of the box the 25th and 75th quartile, and the dotted lines 1.5x the interquartile range). Red circles show the frequencies observed in the original alignment. (B) Relationship between the number of substitutions and observed differences empirically determined by sampling of 25 points. Each point shows the average of 10 simulations. Note that the short distances are sampled more densely. These settings are the default values and used throughout the thesis. (C) Distribution of mean pairwise identities for 1000 random samples. The MPI of the original alignment is shown in red. (D) Comparison of site-wise MPIs in the original alignment and the average of the corresponding sites of 1000 random alignments.

If we consider equal rates for all sites, each site will have the same average MPI which is of course equal to the overall MPI of 0.73 of the whole alignment. Ideally, the average MPI for each simulated site is the same as the original MPI at this site. In this case, the points in the plot are on a diagonal indicating that we have found accurate estimates for the rates.

The substitution rate at a site is related to the observed sequence diversity at this site. If a site is highly conserved the rate is low, whereas high sequence diversity indicates a high mutation rate. So in a first step, we calculate the average number of pairwise differences $\langle p_k \rangle$ for each site k in the alignment with n sequences:

$$\langle p_k \rangle = \frac{2}{n(n-1)} \sum_i^n \sum_{j>i}^n \delta_{ij}^k; \quad \text{with } \delta_{ij}^k = \begin{cases} 1 & \text{if nucleotides in sequences } i, j \\ & \text{differ at site } k \\ 0 & \text{otherwise} \end{cases} \quad (6.10)$$

If we naively choose our rates proportional to $\langle p_k \rangle$ we would underestimate high rates while overestimating low rates. We therefore use the relationship in equation 6.9 to correct for this effect and calculate estimates \hat{f}_k for the rates at site k as follows:

$$\hat{f}_k = \frac{1}{\hat{b}} \cdot \ln \left(1 - \frac{\langle p_k \rangle}{\hat{a}} \right) \quad (6.11)$$

with $f_k = 1$ and $\langle p_k \rangle < \hat{a}$. It must be pointed out that the site-specific rates change the relationship between genetic distance and observed differences (Fig. 6.4B). For correcting the site-specific rates we use the estimates for \hat{a} and \hat{b} from our model *without* site-specific rates. So this is only an approximation and one could think about iteratively refining the estimates. However, we found that this approach already yields rates within one step as can be seen in Fig. 6.4D. Using the model with site-specific rates, the simulated alignments have on average almost exactly the same site-wise MPI as the original one.

The reader will notice that the first three points deviate from the diagonal. This illustrates a limitation of our method. We can on average only reach the level of saturation even if we use very high rates. It is possible, however, that the original data contains sites below the level of saturation. For example in a four way alignment a column can be ACGT, i.e. MPI=0. However, we cannot simulate on average columns with MPI=0, since the MPI is bounded below by zero and our simulations will always contain columns with MPI > 0. In practice this does not seem to cause any obvious problems in particular when we have many sequences where it is unlikely to see columns below saturation. ⁴

⁴ For calculating the site-specific rates, we also treat gaps as missing data and calculate $\langle p_k \rangle$ in Equ. 6.10 only over non-gap positions. After the simulation, the whole column has on average $\langle p_k \rangle$ estimated from the non-gap positions that does not change when originally gapped positions are masked again. For calculating the observed differences p between two sequences we set positions that includes gaps to the average $\langle p_k \rangle$ at this site.

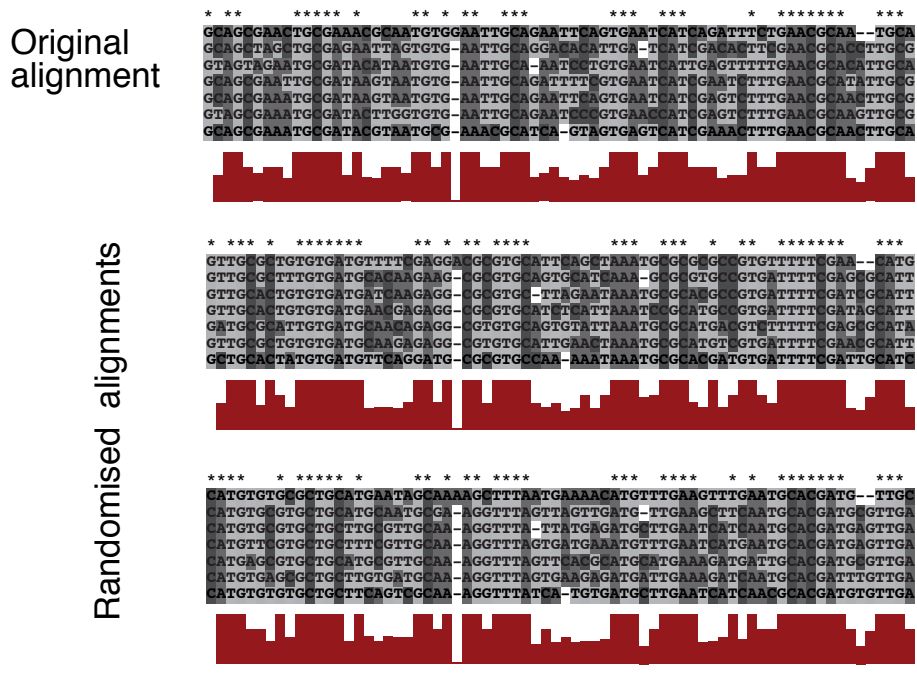


Figure 6.5: Example of randomised alignments. Part of the example alignment used in Fig. 6.4 are shown. The red bars indicate the level of local conservation. Exactly conserved sites are marked by asterisks.

6.2.3 Randomising Vertebrate Genomic Alignments

We tested our randomisation method on vertebrate genomic alignments in terms of how well the simulations reflect the properties of the original data. In a setting similar to recent genomic screens in vertebrates (Washietl *et al.*, 2005b, 2007b), we extracted Multiz (Blanchette *et al.*, 2004) alignment blocks from human chromosome 1. We randomly selected 1000 alignment blocks between 70 and 120 nt in length and between 4 and 10 sequences without considering annotation information. These alignments are meant to represent an unbiased “genomic background” that may also contain functional elements like coding exons or structured RNAs depending on their frequency in the genome.

The alignments were randomised using our new simulation procedure with both the dinucleotide and the mononucleotide model. In addition, we shuffled the alignments using `shuffle-aln.pl`. The global distribution of dinucleotides for the original and randomised data is shown in Fig. 6.6. As expected, the shuffling approach and the mononucleotide simulation give the same results. The dinucleotide distribution obtained by these methods, however, differs from the distribution in the native alignments. One can see for example the well known under-representation of CpGs in the native genomic data. Using our dinucleotide based model, we obtain simulated alignments which are statistically indistinguishable from the native data in terms of their average dinucleotide content.

Also the observed sequence diversity of the simulated alignments closely follows the original

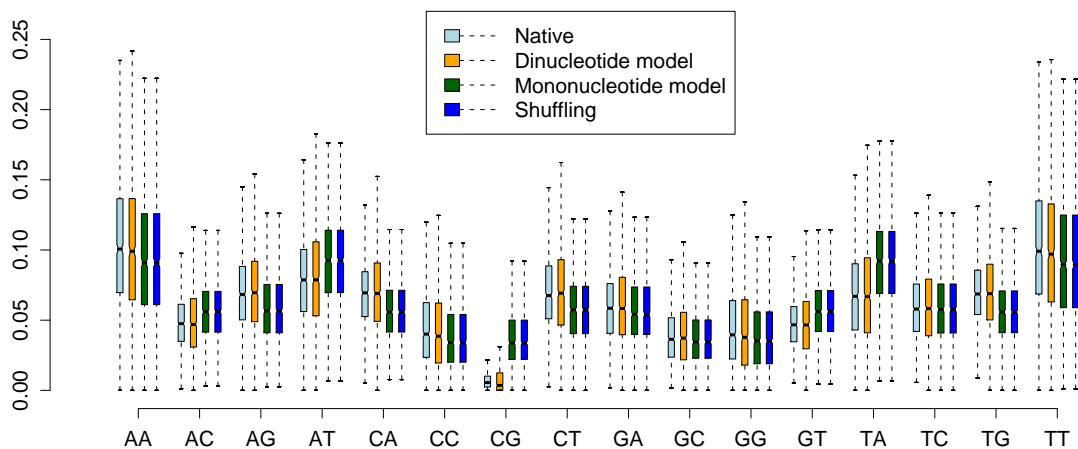


Figure 6.6: Dinucleotide frequencies of genomic alignments. 1000 vertebrate genome alignments were randomised using three different methods. The dinucleotide frequency of the native and randomised data is shown as box-plots.

data as shown in Fig. 6.7. 98.7% of the simulated alignments are within a range of ± 0.05 mean pairwise identity compared to the original alignments. It must be noted, that the distribution in Fig. 6.7 has a mean of $+0.007$ which indicates a subtle bias of the simulations towards higher MPIs. We suspect that this is a result of the way we estimate site-specific rates and related to the issue of sites below saturation discussed before. However, this deviation does not have any practical consequences since it represents a conservative bias in the context of RNA folding controls and, more importantly, seems to be too small to have any noticeable effect at all.

6.2.4 Influence of Randomisation Procedure on RNA Predictions

The main motivation of this chapter is to provide dinucleotide based controls for comparative RNA gene predictions. Therefore, we ran `RNAalifold` and `RNAz` on the alignments to demonstrate how different randomisation procedures affect the results. Fig. 6.8A shows the distribution of `RNAalifold` consensus MFEs on the genomic alignments and their different randomisations. One can see that the genomic alignments show the most stable structures. There is a clear difference between the native genomic alignments and the shuffled and mononucleotide simulated ones. However, the folding energies of the dinucleotide simulated alignments are much closer to the native data. This difference between the di- and mononucleotide simulations reflects the bias caused by the genomic dinucleotide content. The difference between the native and the dinucleotide controls indicates the existence of RNA signals in the genome or, alternatively, another as yet unidentified bias.

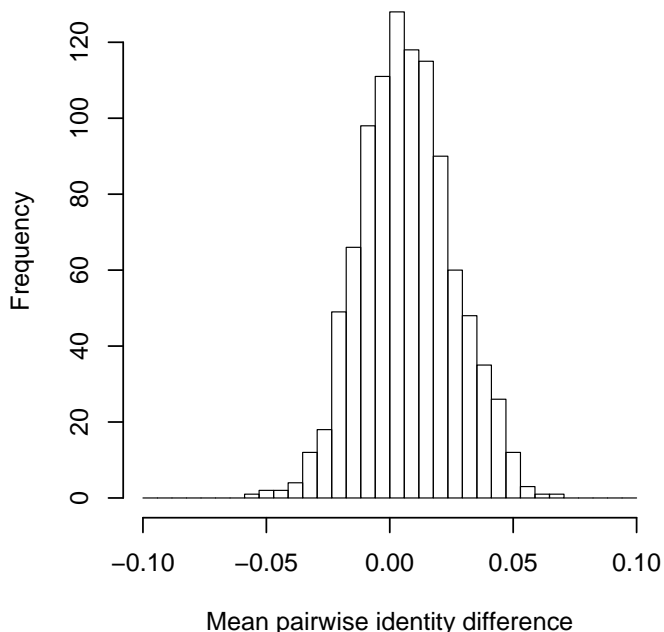


Figure 6.7: Mean pairwise identity in randomised genomic alignments. The distribution of the difference of the mean pairwise identity between the original genomic alignments and the simulated ones (dinucleotide model) is shown.

Clearly, the differences shown here in these cumulative histograms might appear very subtle. The results for the RNAz predictions, however, show that such differences can strongly affect the statistics of RNA gene predictions (Fig. 6.8B). On this particular test set, RNAz predicts RNA signals in 4.3% of the native alignments. Using the conventional shuffling strategy or mononucleotide based simulation one would estimate a false positive rate of 0.8% or 0.7%, respectively. Using the more conservative dinucleotide based model the estimate would be 2.1%, i.e. three times higher. This is consistent with the results obtained by Babak *et al.* using their dinucleotide shuffling approach on pairwise alignments.⁵

⁵ **Limiting base composition variation:** During the testing of the influence of the randomisation procedure on RNA folding, we made an interesting observation. As expected, the variance of the folding energies of randomised data is higher with simulation than with shuffling. However, we also observed that there is difference in the mean. Simulation leads to slightly higher (i.e. less stable) folding energies than shuffling. We observed this behaviour not only on multiple alignments but also on single sequences using shuffling vs. first order Markov simulation. We suspect that extreme deviations in the base composition that can occur in simulated data do not symmetrically lead to the same deviations of the folding energies but preferentially impair the formation of RNA structures. To compensate for this effect, we have introduced an option in our software that only outputs simulated alignments, that are within a specific range of mononucleotide frequencies. We can thus limit our output to mononucleotide frequencies that are almost exactly as in the original data. As a distance measure we use the Euclidean distance $\sum_{\alpha \in A, G, C, T} \sqrt{(\pi_{\alpha}^s - \pi_{\alpha})^2}$ with π_{α} the desired frequency of nucleotide α in the original alignment, and π_{α}^s the observed frequency in the simulation. For all the data shown in Figs. 6.8, 6.9 and Tab. 6.1 we used simulations with this cutoff set to 0.05.

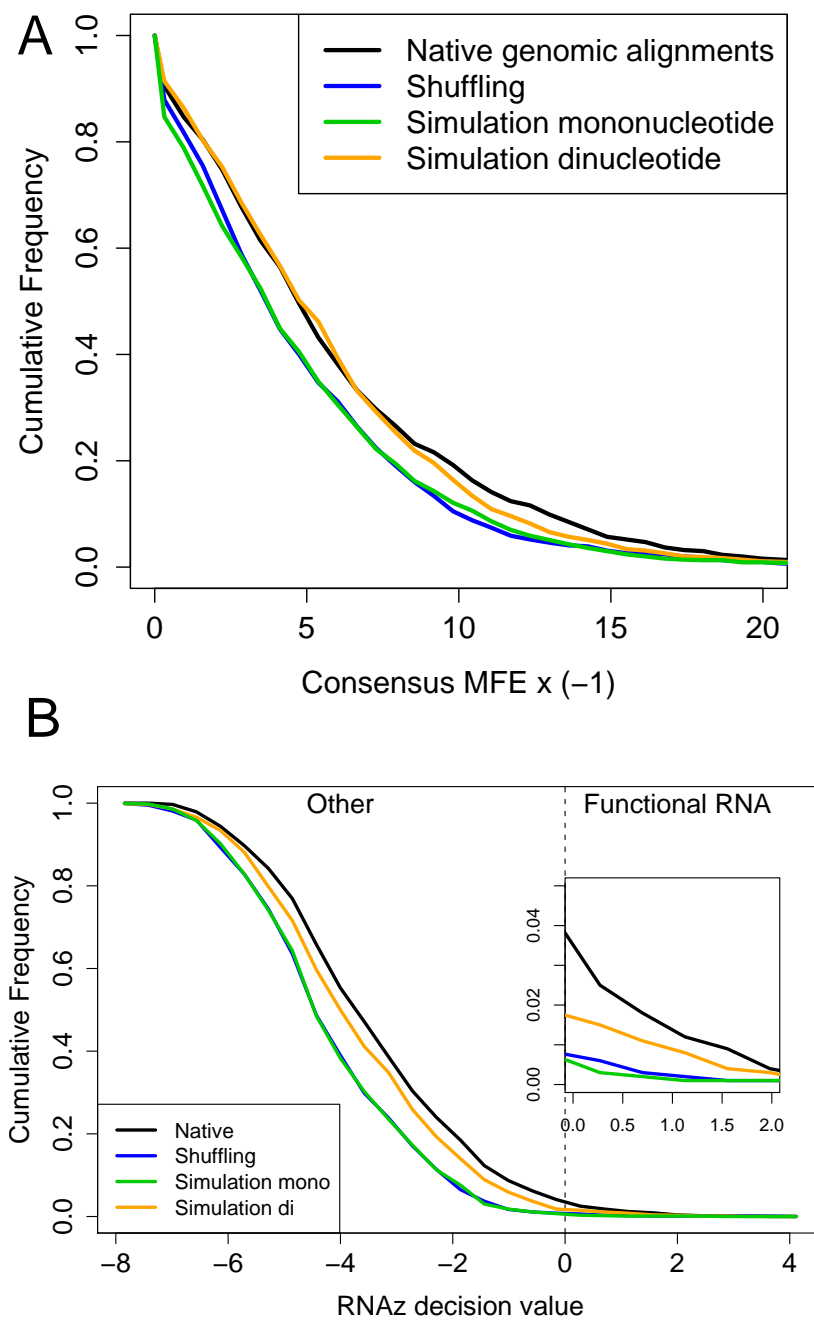


Figure 6.8: Influence of the randomisation procedure on RNA predictions. (A) Cumulative frequency distribution of `RNAalifold` consensus folding energies for the native and randomised alignments. (B) Cumulative frequency distribution of `RNAz` scores. The “decision-value” is the result of the support vector machine classification. Positive values indicate a potential functional RNA while negative values indicate no significant fold. The positive tail is magnified.

6.3 SSSIz: First Dinucleotide Based RNA Gene Finder

Calculating z-Scores to Predict Structural RNAs

We can directly assess the significance of a predicted RNA by calculating a z -score. The folding energy of the native data m and the mean μ and standard deviation σ from randomised data is calculated. The stability of the native fold can then be expressed as $z = (m - \mu)/\sigma$, i.e. the number of standard deviations from the mean (see Chap. 2, Sec. 2.3). This score has been repeatedly used on single sequences applying mono- or dinucleotide shuffling or simulation using a zero or first order Markov model (Clote *et al.*, 2005; Workman and Krogh, 1999). Using shuffled alignments as a null model, this approach is implemented in the RNA gene finding program `AlifoldZ` (Washietl and Hofacker, 2004). The same strategy can be used in combination with our dinucleotide base randomisation strategy without any further modifications (Fig. 6.1).

To test the effectiveness of this approach, we conducted a benchmark similar to those used previously (Washietl and Hofacker, 2004; Washietl *et al.*, 2005c) for testing `AlifoldZ` and `RNAz`. We used multiple sequence alignments of eight different structural RNA families taken from the Rfam database (Griffiths-Jones *et al.*, 2005). The alignments contained three to six sequences and had a mean pairwise identity between 50% and 100%. For the tests of `AlifoldZ` and `RNAz`, shuffled alignments were used as negative controls. For obvious reasons, this is not possible here. So we used genomic alignments from random locations of the human genome (see Methods). Using the “genomic background” as negative controls in this test implies the assumption that the genome does not contain any structural RNAs at all, which is clearly not valid. However, if we assume true structural RNAs to be sparse in the genome this assumption seems to be a sensible choice.

We calculated z -scores with a sample size of 1000 randomisations for both sets of true structured RNAs and the genomic background using three different randomisation meth-

Table 6.1: z -scores and classification performance

Data type	N	RNAz			AlifoldZ			SSSIz (mono)			SSSIz (di)		
		z	S _{0.01}	S _{0.05}	z	S _{0.01}	S _{0.05}	z	S _{0.01}	S _{0.05}	z	S _{0.01}	S _{0.05}
5S rRNA	368	n/a	0.77	0.98	-6.72	0.84	0.98	-6.35	0.86	0.98	-6.35	0.93	1.00
tRNA	382	n/a	0.74	0.98	-6.29	0.75	0.98	-6.24	0.74	0.98	-5.86	0.88	0.99
U2 snRNA	458	n/a	0.76	1.00	-7.17	0.89	0.99	-5.92	0.84	0.97	-5.22	0.93	0.99
U3 snRNA	377	n/a	0.52	0.92	-5.11	0.74	0.86	-4.47	0.69	0.83	-4.23	0.76	0.86
U5 snRNA	424	n/a	0.90	0.96	-5.61	0.77	0.96	-5.10	0.69	0.89	-4.43	0.76	0.91
Hammerhead	499	n/a	0.78	1.00	-6.68	0.85	1.00	-6.67	0.90	1.00	-6.66	0.99	1.00
Group II intron	480	n/a	0.68	0.82	-6.58	0.74	0.81	-6.77	0.72	0.81	-6.29	0.77	0.82
micro RNA precursor	571	n/a	0.75	1.00	-8.89	1.00	1.00	-8.84	1.00	1.00	-7.58	1.00	1.00
Total of all classes	3559	n/a	0.80	0.96	-6.75	0.87	0.95	-6.43	0.85	0.94	-5.93	0.90	0.95
Genomic background	3559	n/a	n/a	n/a	-0.44	n/a	n/a	-0.58	n/a	n/a	-0.15	n/a	n/a

S_{0.01}, S_{0.05}... Sensitivity at a false positive rate of 0.01 and 0.05, respectively.

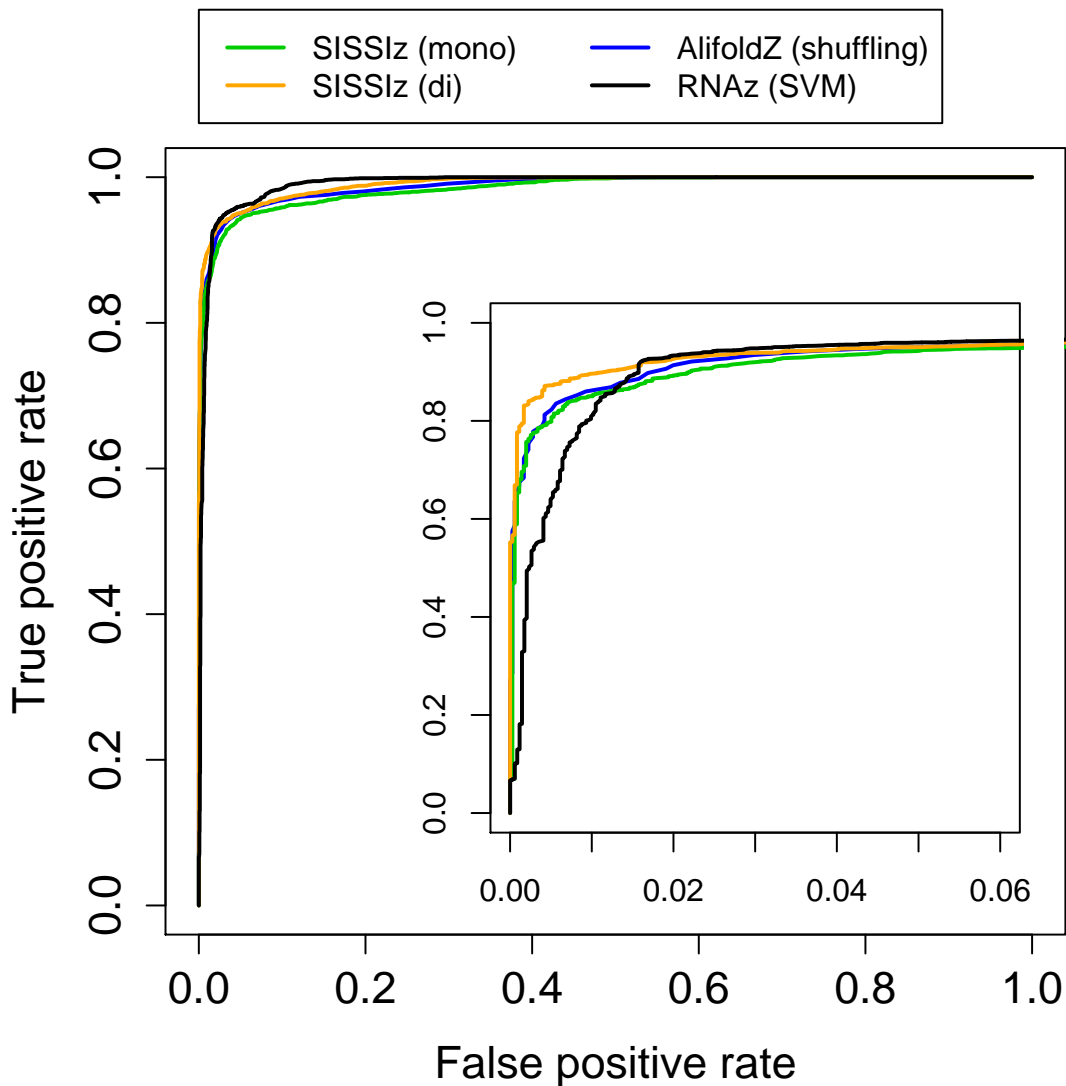


Figure 6.9: Accuracy of z -score based classification of structured RNAs. As positive examples, alignments from eight different classes of structural RNAs were used. As negative examples, random locations from genome wide vertebrate alignments were chosen. ROC curves are shown in dependence on the null model used. In addition, the results of the RNAz support vector machine are shown. The high specificity which is of special interest is magnified.

ods: Shuffling (AlifoldZ), simulation using a mononucleotide model (SISSIz mono) and simulation using the dinucleotide model (SISSIz di). The results are summarized in Tab. 6.1.

Using mononucleotide based randomisation the z -scores of the genomic background are approximately half a standard deviation from zero (-0.44 and -0.58 , for shuffling and

mononucleotide simulation respectively). This shows the relatively strong influence of the genomic background that causes false positive predictions as shown in the previous section and in reference (Babak *et al.*, 2007). Albeit the signal does not vanish completely, the dinucleotide based z -scores are much closer to zero (-0.15).

The z -scores of the structural RNAs in this test set are on average well below -4 indicating a clear structural signal. Also here, we observe that mononucleotide simulated z -score distributions are lower than the dinucleotide simulated ones. In this case, a dinucleotide content that favors stable RNA structures is clearly not only a general background effect of the genomic base composition but a feature of structural RNAs. However, this signal is lost if the dinucleotide based null model is used.

There is also a clear difference between the two mononucleotide randomisation procedures: Shuffling leads to more significant z -scores than simulation. The main reason is the fact that simulation results in higher standard deviations than shuffling which in turn can lead to different z -scores.

This shows that there are many effects that have to be taken into account. To assess the overall classification performance we generated receiver operating characteristic curves based on the three different methods for the z -statistics, as well as the support vector machine score from **RNAz** (Fig. 6.9). In addition, we calculated the sensitivity at two different levels of specificity (0.01 and 0.05) for all four approaches (Tab. 6.1).

The ROC curve shows that all the methods perform very well on this test set. The curve further suggests that there is not as much difference between them. However, differences become evident when looking at the region of high specificity, the only relevant region for practical applications (see inset Fig. 6.9). Here, the dinucleotide based approach generally outperforms the mononucleotide based methods. The improvement is small but clearly noticeable: At a false positive rate of 0.01%, dinucleotide based simulation shows the highest sensitivity for 7 of the 8 RNA classes. For example, in the tRNA group the sensitivity is 13% higher than **AlifoldZ** and **RNAz**. The latter performs significantly worse than all other methods at this level. At a false positive rate of 0.05%, dinucleotide simulation still performs slightly better than mononucleotide shuffling/simulation but is on the same level as **RNAz** that performs significantly better here.

6.3.1 Discussion

In this chapter we addressed the problem of finding an adequate control strategy for comparative noncoding RNA predictions, which are started to get widely used for genome annotation.

Babak *et al.* (2007) demonstrated that currently used null models based on mononucleotide shuffling lead to an underestimation of the false positive rate in such screens. Although individual opinions may be different (Forsdyke, 2007), it is generally accepted that in the context of RNA gene prediction one should consider dinucleotide content as “background” rather than “signal”. However, while there have been dinucleotide controlled randomisa-

tion algorithms for single sequences for more than 20 years, it is a non-trivial problem in the case of multiple sequence alignments.

Here we devised a simulation procedure that produces alignments that have on average a given dinucleotide frequency and sequence diversity (globally and locally). The corresponding model needs to be relatively complex including overlapping dependencies and site-specific rates. Clearly, this model with a high number of parameters would not be a reasonable choice for use in phylogenetic analysis, but it turned out to be a good choice for this specific application.

We have to use heuristics and simplifications to estimate the tree and parameters for this model in reasonable time. The accuracy of our approach is measured in terms of how well the simulations reflect the properties of the original data. In this respect, we found that our strategy performs very well. Again, phylogenetic analysis was not the goal here, but some of the techniques introduced here might be of interest in this context. For example, we found that in the mononucleotide case our estimations for site-specific rates are surprisingly competitive when compared to the currently best maximum likelihood methods (data not shown).

The influence of the null model for genomic RNA predictions was found to be remarkable. Consistent with Babak and colleagues' findings on pairwise alignments, we observed three times more false positives using dinucleotide controls than using mononucleotide controls. This clearly shows that the new approach should be the method of choice to get more sensible estimates of the significance of comparative RNA predictions.

The next obvious step, is to use the new null model to improve current RNA gene prediction algorithms. In analogy to `AlifoldZ`, we combined our new simulation procedure with the `RNAalifold` consensus structure prediction algorithm. `SISSIZ` calculates z -scores that are not biased by the genomic dinucleotide content and it is thus the first comparative gene finding program, that explicitly corrects for this effect. However, by using this conservative null model we also lose part of the signal in true structured RNAs. This might be the main reason, why the observed improvements in the overall classification performance were only relatively small.

In general, the support vector machine approach used by `RNAz` is preferable over the `AlifoldZ` approach, since it is orders of magnitude faster. However, it turned out to be difficult to create a dinucleotide based version of `RNAz` mainly for two reasons. Until now, there was no way to produce a dinucleotide controlled negative test set that is necessary for training the two class support vector machine (Washietl *et al.*, 2005c). With the method presented here, we have solved this problem and it is now possible to create test sets with specific dinucleotide properties. However, it remains an unsolved question how to compute dinucleotide based z -scores efficiently without shuffling. `RNAz` uses a regression approach to solve this problem for mononucleotides, which, unfortunately, does not scale well to the high dimensional dinucleotide case.

A promising alternative to the thermodynamic RNA prediction methods used in this chapter, are probabilistic methods. The `EvoFold` algorithm (Pedersen *et al.*, 2006) uses

phylogenetic stochastic context-free grammars and, in its core, depends on a null model which is essentially an independent mononucleotide model. Since the folding grammar of `EvoFold` does not explicitly model stacking interactions there is no need for using a null model with overlapping dinucleotides as we have described here. However, also `EvoFold` was found to be affected to some degree by the dinucleotide content for reasons that are not immediately obvious (Babak *et al.*, 2007). A dinucleotide background model together with an advanced folding grammar that considers stacks can thus be expected to improve performance. However, it would require considerable effort to include such a null model into the sophisticated probabilistic framework of `EvoFold`.

Finally, we want to add that our randomisation algorithm is not only of interest in the context of RNA gene prediction. It can be used for other comparative genomics applications whenever random alignments are needed as control. One could consider other applications in the context of RNA structures (e.g. prediction of conserved miRNA target sites) but also in different context (e.g. conserved sequence motifs). Currently `SISSIz` implements a mono- and dinucleotide model which should be sufficient for many applications. In principle, however, it is also possible to consider higher order correlations within the `SISSI` framework.

6.4 The Beauty of Elephants

“With enough parameters you can fit an elephant”.⁶

Simplicity versus complexity is a general discussion point. Thus, also in the phylogenetic community a level of concern has been noticed with the growth of formal models use in phylogenetics, also in terms of modelling structure and over-fitting the data. However, one should note that even those methods that do not formalise a model, and thus claim to be model-free, e.g. parsimony, make significant, and sometimes, incorrect assumptions about evolution. A recent review by Kelchner and Thomas (2007) focuses on nine key questions in phylogenetics. Beside these interesting phylogenetic questions, this review address only independent models and does not consider biochemical perspectives. Generally, `SISSI` can address these questions of performance of tree building methods under dependencies. In addition, in Chap. 4 we have used `SISSI` for testing the performance of structure prediction methods and the understanding of the intertwined relationship between structure and substitution process. Here, with `SISSIz` we focus on a special substitution model that captures the neighbour dependencies and other important alignment features except the signal in question in relation to structure prediction of ncRNAs of genomic alignments. Thus, it based on a biochemical assumptions, which we test in terms of how well the simulations reflect the properties of the original data. In phylogenetics, the usage of Markov models is aimed at the accurate reconstruction but also to model the process of sequence evolution itself. Thus, we were following a quote of Simon Tavaré (Steel, 2005):

⁶folklore quote from physics

“Talk to the biochemists !”

Similar to `SISSIZ` for `RNAz`, in general `SISSI` has the potential to create test sets under many different constraints for other applications. For example, recently, a computational approach to RNA free energy parameter estimation was developed that can be efficiently trained on large sets of structural as well as thermodynamic data. On biologically sound data, Andronescu *et al.* (2007) have obtained revised parameters for the Turner99 energy model (Mathews *et al.*, 1999). From a viewpoint of a phylogenetic definition of structure of Chap. 4 the question arises; exist “one perfect energy model” or are there more than one? `SISSI` could simulate test datasets to optimise the energy parameters under different constraints.

Chapter 7



R. Franklin (1920-58)

OSM's Tree Aspects

*The possibility of Franklin having played a major role was not revealed until Watson (1968) wrote his personal account, *The Double Helix*, which subsequently inspired several people to investigate the history of the discovery of the structure of DNA and Franklin's contribution (Franklin and Gosling, 1953).*

TO ANDREA & SEBASTIAN JACOBI

Phylogenetic research questions about RNA mostly direct the focus on the performance of the tree reconstruction methods, like Maximum Likelihood (ML), Neighbour Joining (NJ) or Maximum Parsimony (MP), when sequence sites are not independently evolving. In the ML framework several mixture models or models for overlapping dependencies were developed. While mixture models have progressed overlapping dependencies are still a serious problem. Most approaches are based on (Bayesian) Markov Chain Monte Carlo methods (MCMC) (Huelsenbeck and Bollback, 2001), which are still not directly applicable to RNA research, especially on a genome-wide scale, since the running time is too long.

In contrast, we introduce another view on sequence evolution. So far, our phylogenetic definition of structure of Chap. 4 has specified the evolutionary process of nucleotide evolution with site-specific interactions. However, is there a complementary framework to SISSI, which merges the substitution matrix directly with a phylogenetic tree ?

This is discussed as a description in pattern space. Doing this, the substitution matrix encodes the phylogenetic tree directly. From the viewpoint of the tree aspect we consider available tree reconstruction principles in relation to our *One Step Mutation* (OSM) description. Although, so far we cannot include the neighbourhood system directly into the OSM matrix, OSM has the potential to reverse back to previous chapters: for example to Chap. 5, where an analytical formula of the pattern distributions did not seem feasible and to Chap. 6, where we have discussed how far a model explains the sequences. ¹

¹OSM is a collaboration with S. Klaere and A. von Haeseler.

7.1 OSM: One Step Mutation Matrices

Here, we will introduce another description of the evolutionary process on trees. Contrary to other approaches we model the substitution process in two steps. First we assume (arbitrary) scaled branch lengths on a given phylogenetic tree. Second we allocate a Poisson distributed number of substitutions on the branches. The probability to place a mutation on a branch is proportional to its relative branch length. More importantly the action of a single mutation on an alignment column is described as a double stochastic matrix, the so-called one-step mutation matrix (OSM). This matrix leads to analytical formulas for the posterior probability distribution of the number of substitutions for an alignment column. More precisely, given a phylogenetic tree and an alignment that evolved along the tree, we now ask the following question: How does the alignment change, if an additional substitution on an arbitrary branch of the tree took place? In other words, consider a collection of morphological traits that are either in an ancestral (0) or derived (1) state. Each derived character state characterizes a monophyletic group and represents a cluster in the tree. For such a data matrix (or alignment) the tree reconstruction problem is easy. However, stochastic effects that act somewhere on the branches of the tree may disturb this signal. This noise is modeled by the assumption of throwing an arbitrary number of changes on the tree and measuring their impact on the otherwise perfect data matrix. To this end, we construct a OSM matrix.

7.1.1 The Binary model on an n -taxon tree

We consider a set of n taxa $S = \{1, \dots, n\}$. With S comes along some information about the common properties and differences of the taxa, typically displayed in an alignment. In the following a (sequence) *alignment* \mathbb{A} is an $n \times \ell$ -array with entries either 0 or 1, where ℓ is the length of the alignment. Each of the ℓ columns (*sites*) \mathbf{a}_j of the alignment represents a *pattern* of n homologous characters, where $a_{ij} \in \{0, 1\}$ is the state of character j in taxon i . For binary character states 2^n patterns are possible.

We are interested in the evolution of such patterns along a (rooted) *tree* $T = (V, E)$ with node set V and branch set $E \subset V \times V$ (Semple and Steel, 2003). The node set V contains the taxon set S that forms the leaf set. Avoiding the technical details, each branch is uniquely encoded by the subsets X of S that originates from the branch. Such a set X specified by a branch will be called *cluster*. A leaf is a trivial cluster.

Finally, we introduce a function $\lambda: E \rightarrow \mathbb{R}_+$, such that $\lambda(e) > 0$ represents the *length* of a branch $e \in E$. The *tree-length* Λ_T is the sum of the branch lengths. The *relative branch length*

$$p_e = \frac{\lambda(e)}{\Lambda_T} \tag{7.1}$$

denotes the probability that a substitution hits branch e of the rooted tree T .

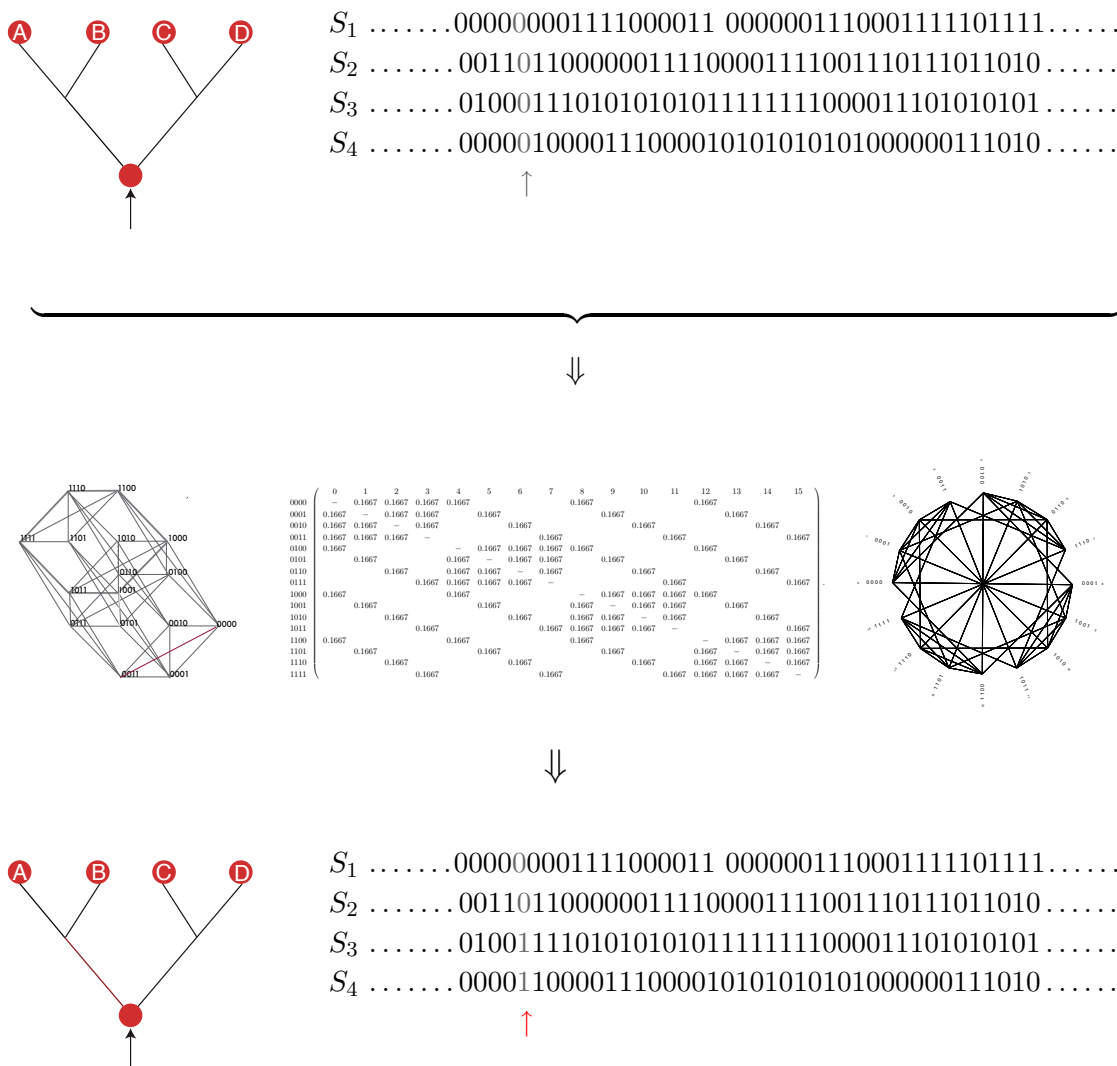


Figure 7.1: We are interested in the evolution of patterns in an alignment along a tree. Example of a four taxa tree with the corresponding OSM-cube, the OSM-corresponding adjacency graph and the OSM matrix: The node set V contains the taxon set S that forms the leaf set. The branches are encoded by the subsets of X that originate from the branch. This can be easily done by coding the leaf labels with powers of 2 and recursively labeling the inner nodes by the sum of the labels of its immediate descendant (see further examples and more details in Fig. 7.4). However, the random walk is very different from the standard random walk on the hypercube (Eigen *et al.*, 1988). Instead of looking at the process through time along the branches, we describe how a single mutation that occurs anywhere on tree T changes the character states on the leaves. F : Thus, if a mutation hits an alignment site (grey arrow), then the corresponding pattern (light gray arrow) will change to a new pattern (red arrow), where the new pattern is determined according to the branch, where the mutation occurs. The mutation path is actually a path through a cube, where elements are connected by an edge, if the OSM matrix is greater zero.

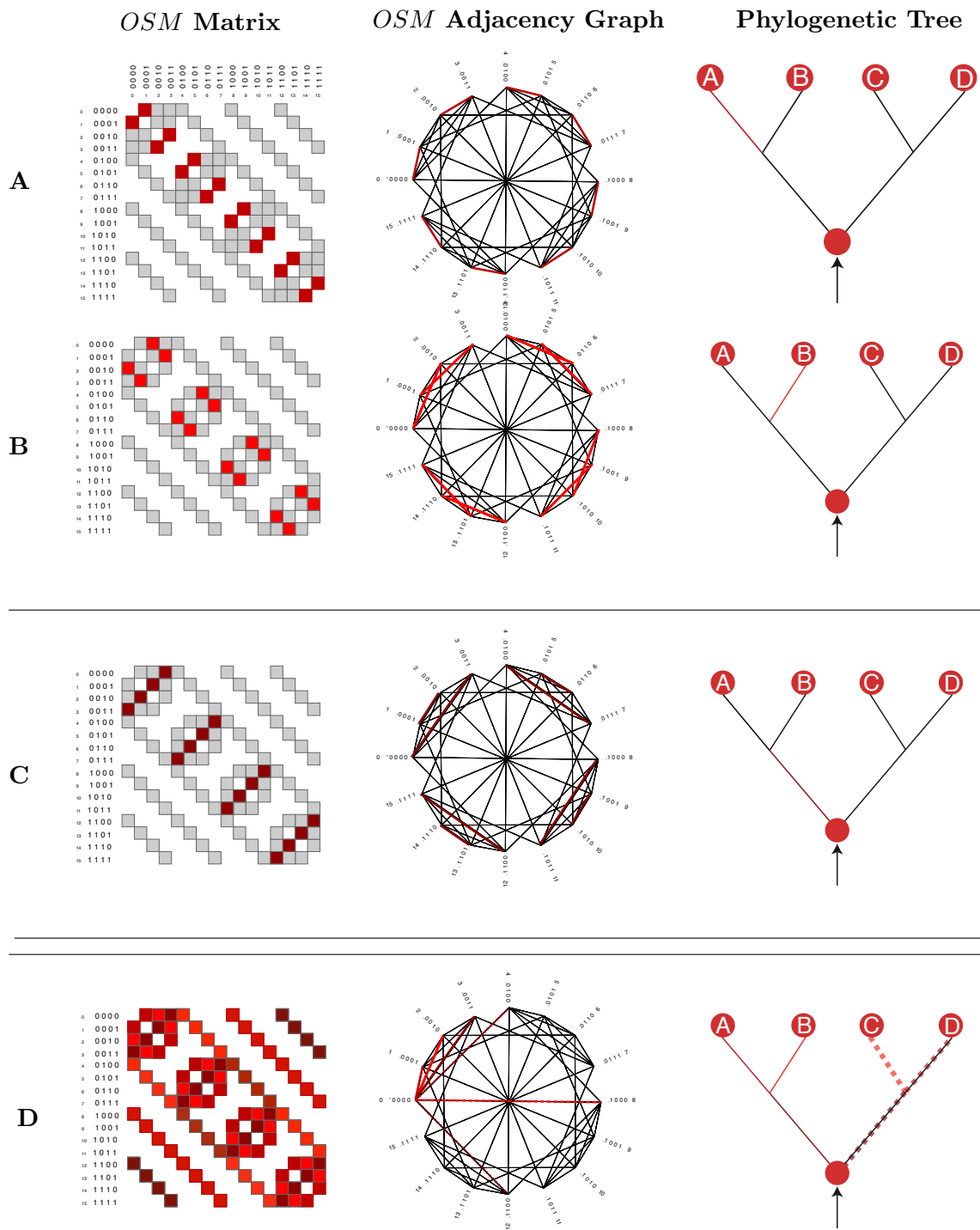


Figure 7.2: OSMmatrix with substitutions on different branches of the phylogenetic tree T_4 at the right. The corresponding branches defined by the corresponding clusters are highlighted. A: The branch defined by cluster $\{A\}$ is highlighted. A substitution on this branch gives rise to the unique change in patterns depicted in the graph or its corresponding adjacency matrix. B: Likewise cluster $\{B\}$. C: Here, the cluster $\{AB\}$ is pointed out. D: The whole OSM matrix of T_4 : Corresponding components are identified by common colours.

7.1.2 The effect of substitutions on an alignment

We now describe how a single mutation on the tree changes the current character states at the leaves. Obviously the outcome will depend on the branch where the substitution occurred. Moreover, each of the 2^n possible patterns will be affected differently by such a substitution. Therefore we introduce a $2^n \times 2^n$ matrix that describes the action of a substitution on the patterns for a specific branch. Fig. 7.1 illustrates the simple four taxa case.

Fig. 7.2 describes the model on an example tree T_4 with four taxa. For instance, a substitution at the branch defined by cluster $\{A\}$ changes the pattern 1011 to the pattern 1010 because only the character of taxon A is affected. Please note that the order of taxa is (D, C, B, A) for each pattern. All possible changes between the patterns identified by a substitution on branch e_A are depicted in the substitution graph (Fig. 7.2A). The corresponding adjacency matrix σ_A is displayed, where a red square stands for one (the patterns are connected by an edge in the substitution graph) and a white square represents zero. The structure of matrix σ_A constitutes an example of the so-called *permutation matrices* with entries equal to one if the substitution converts one pattern into another (Bona, 2004, pp.75). Respectively, Fig. 7.2B shows the *permutation matrices* σ_B defined by cluster $\{B\}$, and the *permutation matrix* σ_{AB} defined by cluster $\{AB\}$ is shown in Fig. 7.2C.

For each branch we easily construct the corresponding permutation matrix. We point out that the permutation matrix for a non-trivial cluster is the product of the permutation matrices of its elements. In other words the action of one mutation on a branch e can be replaced by any partition of the cluster associated with e , such that each set of the partition is represented by a branch in the tree. For tree T_4 we obtain six permutation matrices

$$\sigma_A, \sigma_B, \sigma_C, \sigma_D \text{ and } \sigma_{AB} = \sigma_A \cdot \sigma_B, \sigma_{CD} = \sigma_C \cdot \sigma_D. \quad (7.2)$$

To take the relative contribution of the branch lengths into account, we weight each permutation matrix with p_e as described in (7.1). Such matrices are a special case of the *generalized* permutation matrices. Then the so-called *one step mutation* (OSM) matrix of the tree T_4 is simply the convex sum:

$$M_4 = p_A \cdot \sigma_A + p_B \cdot \sigma_B + p_C \cdot \sigma_C + p_D \cdot \sigma_D + p_{AB} \cdot \sigma_{AB} + p_{CD} \cdot \sigma_{CD}. \quad (7.3)$$

Fig. 7.2D shows the result of this computation. The substitution graph in Fig. 7.2D displays the effect of a substitution on all the branches of the tree on the patterns. Two patterns are connected by an edge if a substitution switches between the two patterns.

For an arbitrary phylogenetic tree T on n taxa the OSM matrix is obtained by:

$$M_T = \sum_{e \in E} p_e \sigma_{\mathcal{C}(e)}, \quad (7.4)$$

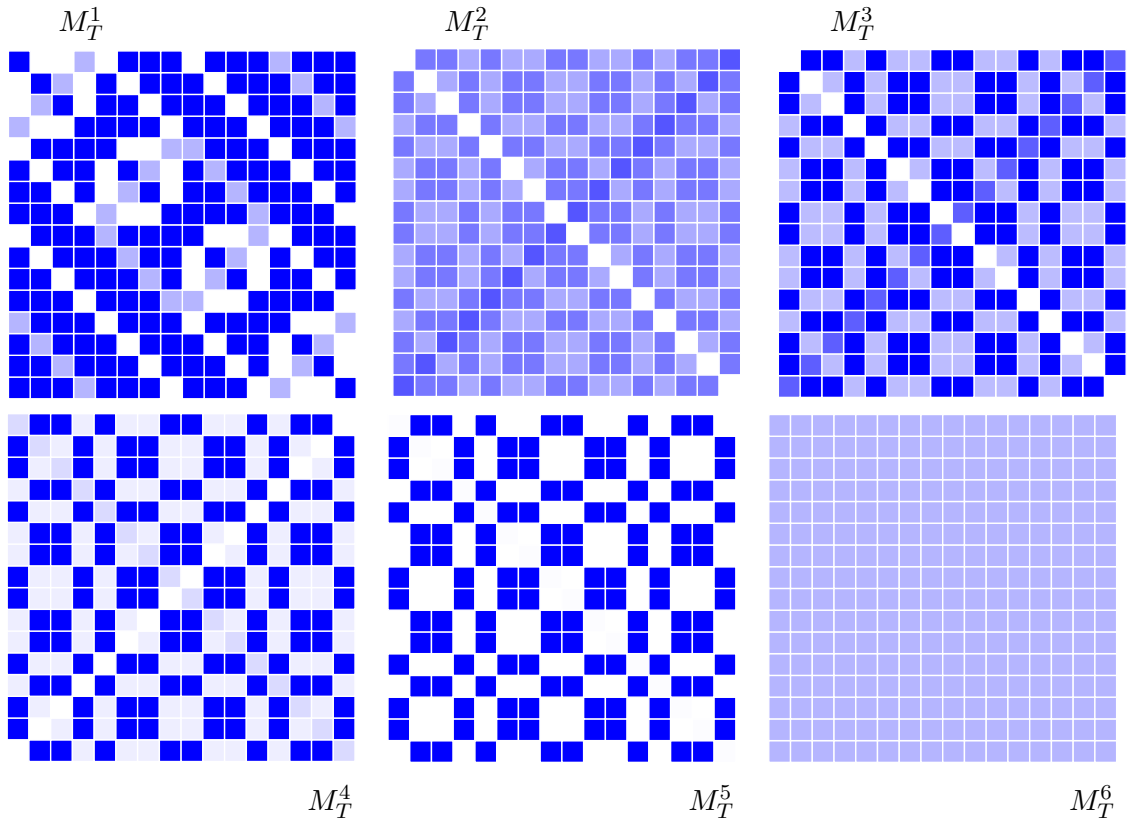


Figure 7.3: A OSM matrix with substitutions on different branches of the four rooted tree T_4 as illustrated in Fig. 7.2. Here, the OSM matrices are shown as density plots with the colour code from white (highest probability) to dark blue (zero). All six branch lengths are assumed to be equal. The K -th power M_T^K provides the probabilities to move from one pattern to another in K substitutions. If K is large, then M_T^K will lose the phylogenetic information of the original alignment and will approximate the uniform distribution, that is each pattern occurs with the same frequencies (all entries are light blue at the right bottom).

where $\mathcal{C}(e)$ is the cluster identified by branch $e \in E$.

The entry $M_T(\mathbf{i}, \mathbf{j})$ is positive if the tree T contains a cluster where a substitution on the corresponding branch implies that pattern \mathbf{i} is changed to \mathbf{j} . Hence, each row and each column has $2n - 2$ non-zero entries, one entry for each branch in the tree. Thus, the OSM matrix belongs to the class of doubly stochastic Markov transition matrices, where the relative branch lengths are represented exactly once in each row and each column. Consequently, the K -th power M_T^K provides the probabilities to move from one pattern to another in K substitutions. Thus, the repeated application of M_T describes a *random walk* on the state space of the 2^n patterns (Fig. 7.4). This random walk is very different from the standard random walk on the hypercube (Eigen *et al.*, 1988). If K is large, then M_T^K will lose the phylogenetic information of the original alignment and will approximate the uniform distribution, that is each pattern occurs at the same frequency like illustrated in Fig. 7.3. However, our setting does not yet assume a probability distribution for the

number of substitutions on the tree. Similar to our random walk in sequence space we assume that the number of substitutions is Poisson and we compute the average OSM matrix by

$$\overline{M}_T = \sum_{K=0}^{\infty} \frac{\exp(-\Lambda_T)\Lambda_T^K}{K!} M_T^K, \quad (7.5)$$

which is equivalent to

$$\overline{M}_T = \exp(-\Lambda_T) \cdot \exp[\Lambda_T M_T].$$

The exponential of the matrix $\Lambda_T M_T$ is easy to compute, because M_T is a sum of generalized permutation matrices (Equ. 7.4), which commute with respect to matrix multiplication.

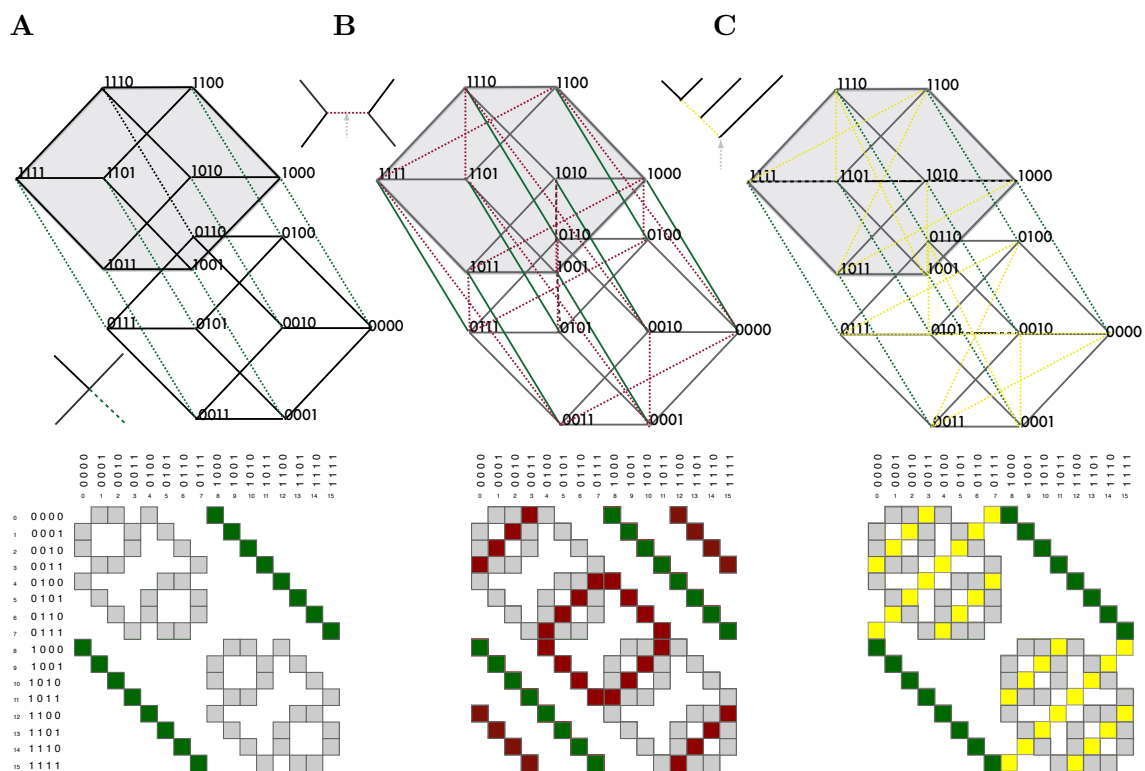


Figure 7.4: Random walk through pattern space & OSM matrices on different four-taxon trees: The nodes of the cube represent the 2^n (16) patterns. The edges represent nearest neighbours. **A** shows the iterative buildup of the normal sequence space. Each additional taxa requires a doubling of the former diagram (gray) and to connect points in both diagrams (dash line). **B+C**: The different tree topologies induce different graphs on the hypercube in Fig. A. Thus, the random walk is very different from the standard random walk. Here, the mutation path is actually a path through a cube, where elements are connected by an edge, if the OSM is greater than zero. The figure shows the tree topology of a balanced four taxa tree (B) and caterpillar four taxa tree (C) introduce different OSM adjacency matrices.

7.1.3 Relation to tree reconstruction

The OSM matrix leads to a very general description of character based phylogenetic inference techniques. Moreover, the explicit model assumptions in maximum likelihood and the implicit assumptions in maximum parsimony are directly comparable.

The OSM matrix and its powers describe the substitution process between arbitrary patterns. However, in classical phylogeny the starting point of a substitution process are ancestral states on trees. In particular, one assumes a stationary distribution $\pi = (\pi_0, \pi_1)$ of character states at the root, and the characters evolve along the tree according to a Markov transition matrix (Tavaré, 1986). In our framework this is equivalent to starting in the *constant* patterns $\mathbf{0} = (0, \dots, 0)$ or $\mathbf{1} = (1, \dots, 1)$ and letting it evolve according to the OSM matrix. This process has a non-stationary pattern distribution π_T^K which starts at $\pi_T^0 = (\pi_0, 0, \dots, 0, \pi_1)$, i.e. with zero substitutions only constant patterns exist, and in each step the pattern distribution is given by $\pi_T^K = M_T^K \pi_T^0$. If the number of substitutions is not weighted as in Equ. 7.5, then π_T^K will approach the uniform distribution as K goes to infinity. To overcome the loss of phylogenetic signal, we assume in the following that the number of substitutions on a tree is Poisson distributed with parameter Λ_T . Moreover, we assume that the substitution process is described by the symmetric Cavender-Farris-Neyman mutation model (CFN, Cavender, 1978; Farris, 1973; Neyman, 1971). Under these assumptions the probability of observing pattern \mathbf{a} when starting in a constant pattern is then calculated employing Equ. 7.5

$$\mathbb{P}[\mathbf{a}|\{\mathbf{0}, \mathbf{1}\}] = \pi_0 \overline{M}_T(\mathbf{0}, \mathbf{a}) + \pi_1 \overline{M}_T(\mathbf{1}, \mathbf{a}), \quad (7.6)$$

where π_0 and π_1 are taken from the stationary distribution of character states. The resulting probability distribution for all possible patterns is then identical to the standard way of computing the probabilities of pattern (Felsenstein, 2004).

Distance Approaches

Now, we briefly illustrate how to derive distance corrections from the OSM matrix. To this end, we consider the rooted tree with two leaves A, B and branch lengths λ_A and λ_B . Then the corresponding OSM matrix \mathbf{M}_2 has the following structure

$$\mathbf{M}_2 = \begin{pmatrix} 0 & p_A & p_B & 0 \\ p_A & 0 & 0 & p_B \\ p_B & 0 & 0 & p_A \\ 0 & p_B & p_A & 0 \end{pmatrix},$$

where p_A and p_B are computed according to (7.1). If evolution started with character state 0 or 1 at the root of the tree and the character states are in equilibrium ($\pi_0 = \pi_1$), then we quickly compute

$$\mathbb{P}(\mathbf{0}, \mathbf{1}) = \frac{1}{2} (1 + \exp(-2\Lambda))$$

as the probability to observe a constant pattern in an alignment. Similarly we compute the probability to observe different character states between taxa A and B . From this it is straightforward to get the distance correction of the CFN model.

Maximum Likelihood

The maximum likelihood principle for an alignment \mathbb{A} and a tree T is easily formulated in terms of the OSM matrix. We introduce as parameter vector θ the branch lengths of T . Then the probability of \mathbb{A} is given by

$$L(\mathbb{A}|T) = \prod_{i=1}^{\ell} \mathbb{P}[\mathbf{a}_i | \{\mathbf{0}, \mathbf{1}\}] \quad (7.7)$$

where the factors on the right-hand side are defined by equation (7.6). The parameter vector θ enters the equation via the OSM and $\Lambda = \sum \theta_i$ in the obvious way. As usual, we want to find parameter assignments such that Equ.7.7 is maximized.

Maximum Parsimony

We associate the adjacency matrix \mathbf{A}_{OSM} (e.g., Cormen *et al.*, 2001, sect. 22.1), or simply \mathbf{A} , with the OSM matrix. \mathbf{A} is obtained as the unweighted sum of the permutation matrices $\sigma_{\mathcal{C}(e)}$. Hence, an entry \mathbf{A}_{ij} is equal to one, when there is a branch in the tree which changes pattern \mathbf{i} into pattern \mathbf{j} , and is zero otherwise. Finally, we note that $\mathbf{A}^K(\mathbf{i}, \mathbf{j})$ describes the number of paths of length K between pattern \mathbf{i} and \mathbf{j} . Each path specifies a series of branches in the tree where a substitution occurred.

Now, fix a column \mathbf{a}_i in alignment \mathbb{A} , and a tree T . We ask for the minimal number K_{\min} such that $\mathbf{A}^{K_{\min}}(\mathbf{a}_i, \mathbf{0})$ or $\mathbf{A}^{K_{\min}}(\mathbf{a}_i, \mathbf{1})$ is greater than zero. In other words, for an alignment column \mathbf{a}_i the minimal number of mutations on T equals

$$MP(\mathbf{a}_i) = \min\{K \in \mathbb{N} | \mathbf{A}^K(\mathbf{a}_i, \mathbf{0}) > 0 \text{ or } \mathbf{A}^K(\mathbf{a}_i, \mathbf{1}) > 0\}.$$

Thus the minimal number of mutations for an alignment $\mathbb{A} = (\mathbf{a}_1, \dots, \mathbf{a}_\ell)$ equals

$$MP(T) = \sum_{i=1}^{\ell} MP(\mathbf{a}_i). \quad (7.8)$$

This is another description of the maximum parsimony principle.

7.1.4 Mapping Substitutions

From the computation of the powers of the OSM matrix it is possible to derive the (posterior) probability distribution, $\text{ppdf}(K|\mathbf{x})$, of the number of mutations that generated an observed pattern \mathbf{x} , when the process started in patterns $\mathbf{0}$ or $\mathbf{1}$. The posterior probabilities have been estimated before employing Bayesian simulation methods (Nielsen, 2002; Huelsenbeck *et al.*, 2003; Minin and Suchard, 2008), but an analytic approach has not been attempted before.

In general, the posterior probabilities $\text{ppdf}(K|\mathbf{a})$ for a pattern \mathbf{a} are calculated in the following way using Equ. 7.6:

$$\text{ppdf}(K|\mathbf{a}) = \frac{e^{-\Lambda T} \Lambda_T^K (\pi_0 M_T^K(\mathbf{0}, \mathbf{a}) + \pi_1 M_T^K(\mathbf{1}, \mathbf{a}))}{K! \mathbb{P}[\mathbf{a}|\{\mathbf{0}, \mathbf{1}\}]},$$

i.e. we compute for pattern \mathbf{a} the proportion of its occurrence after K substitutions.

Only if Λ is large, then the posterior mean number of substitutions will approach the expected number of substitutions per site Λ . For a constant pattern the posterior mean is always smaller than Λ and for non-constant patterns the posterior mean is larger than Λ . Similarly we extend the calculations to a four-taxon tree. For instance, consider the four-taxon tree T_4 (Fig. 7.2). This tree has two non-trivial clusters $\{A, B\}$ and $\{C, D\}$. We want to compute the posterior probability of the number of substitutions if the constant pattern 0000 is observed. Let us assume that the two character states occur with uniform probability, then we can compute:

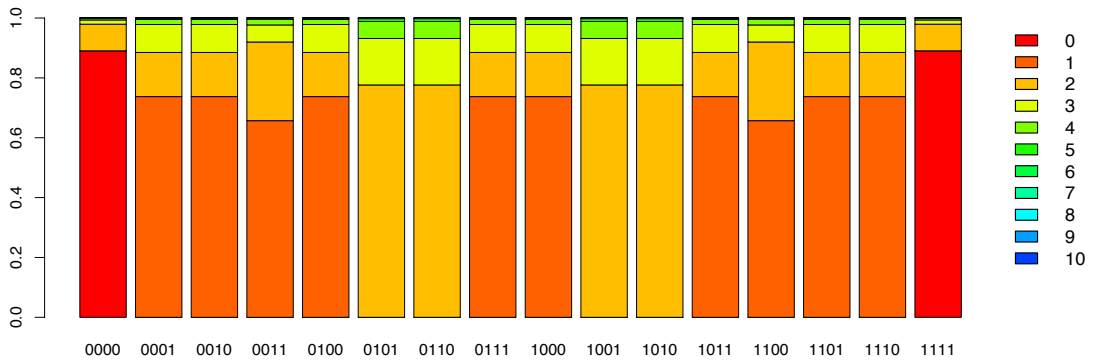


Figure 7.5: Posterior probabilities for the eight symmetric patterns of the four-taxon tree T_4 with branch lengths $\lambda_A = \lambda_B = \lambda_C = \lambda_D = 0.1$ and $\lambda_{AB} = \lambda_{CD} = 0.05$, and character distribution $\pi_0 = \pi_1 = 1/2$. Note that the symmetry in the posterior probabilities is due the uniform stationary character distribution. The 8 patterns can be classified in constant, parsimonious uninformative, compatible and incompatible patterns shown in Fig. 7.6.

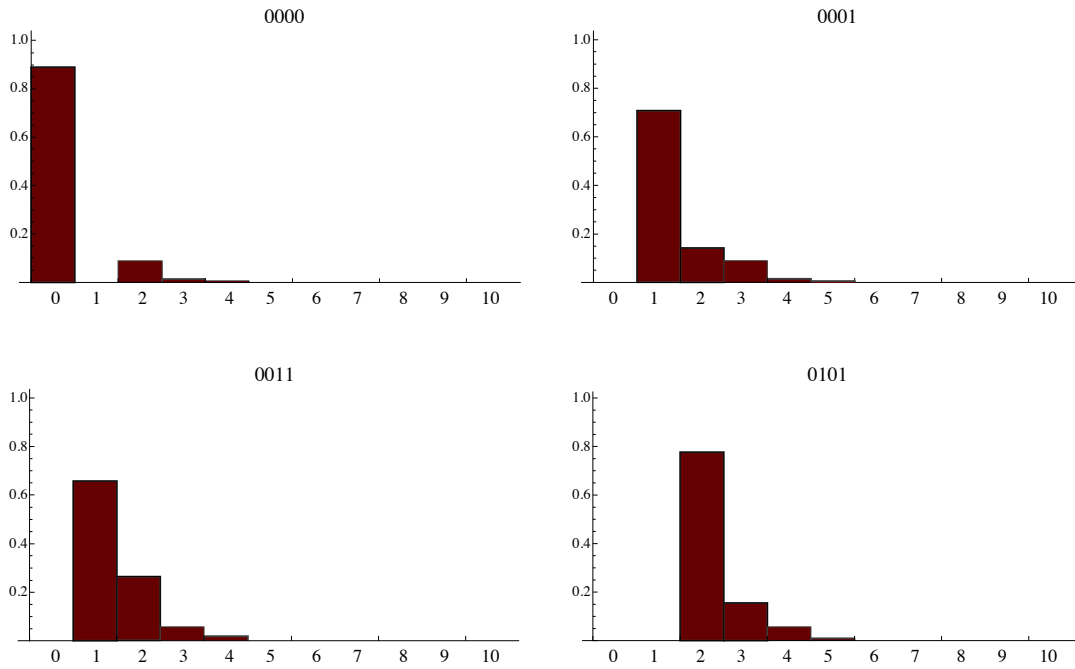


Figure 7.6: Posterior probabilities for representative patterns of the four-taxon tree T_4 with branch lengths $\lambda_A = \lambda_B = \lambda_C = \lambda_D = 0.2$ and $\lambda_{AB} = \lambda_{CD} = 0.1$, and character distribution $\pi_0 = \pi_1 = 1/2$. The selected patterns 0000, 0001, 0011, and 0101 represent constant, parsimonious uninformative, compatible and incompatible patterns, respectively.

$$\mathbb{P}[0000|\{\mathbf{0}, \mathbf{1}\}] = \frac{1}{16 e^\Lambda} (e^{\lambda_A - \lambda_B + \lambda_C - \lambda_D - \lambda_X} + e^{\lambda_A - \lambda_B - \lambda_C + \lambda_D - \lambda_X} + e^{-\lambda_A + \lambda_B - \lambda_C + \lambda_D - \lambda_X} + e^{\lambda_A + \lambda_B - \lambda_C - \lambda_D + \lambda_X} + e^{-\lambda_A - \lambda_B + \lambda_C + \lambda_D + \lambda_X} + e^{-\lambda_A - \lambda_B - \lambda_C - \lambda_D + \lambda_X} + e^{-\lambda_A + \lambda_B + \lambda_C - \lambda_D - \lambda_X} + e^{\lambda_A + \lambda_B + \lambda_C + \lambda_D + \lambda_X}),$$

where $\lambda_X = \lambda_{AB} + \lambda_{CD}$ is the sum of the lengths of the interior branches. Now Taylor expansion leads to the desired posterior probability distribution. Fig. 7.6 shows the resulting posterior probability distributions for the 16 possible patterns, assuming branch lengths $\lambda_A = \lambda_B = \lambda_C = \lambda_D = 0.2$ and $\lambda_{AB} = \lambda_{CD} = 0.1$. The symmetries in the CFN model are reflected in the symmetries of the posterior distributions. Complementary patterns (i.e. 0000 and 1111) show the same distribution. Because the tree is clock-like the parsimonious uninformative patterns (0001, 0010, 0100, 1000) and their complements show identical distributions, as do the patterns that need at least two substitutions (0101, 0110, 1010, 1001) on T_4 . Posterior probabilities may be used to compute for instance the number of unvaried sites (Fitch and Ayala, 1994), that is exactly the proportion of the constant patterns with zero substitutions. In our example we expect about 42% constant patterns of which approximately 90% are unvaried. This is only one application for posterior probabilities of the number of substitutions. As in the two taxon case we compute the posterior mean

of substitutions for pattern \mathbf{a} as

$$\mu(\mathbf{a}) = \sum_{K=0}^{\infty} K \cdot \text{ppdf}(K|\mathbf{a}).$$

Fig. 7.7(a) shows the posterior mean number of substitutions for the topology of T_4 with branch probabilities $p_A = p_B = p_C = p_D = 0.2$ and $p_{AB} = p_{CD} = 0.1$ for a constant pattern (0000), a pattern compatible with an interior branch (0011) and a pattern incompatible with the tree (0110). The difference between posterior mean and tree lengths is smaller than 0.01 if the tree lengths exceeds 10 substitutions per site. Fig. 7.7(b) displays the posterior means for a tree with branch probabilities $p_A = p_D = 0.47$, $p_B = p_C = 0.02$ and $p_{AB} = p_{CD} = 0.01$. The proportion of p_A and p_D is so large that the incompatible pattern 0110 will be observed more often than the pattern 0011, that is compatible with a branch of the tree. Thus, this tree is an instance, where maximum parsimony will reconstruct the wrong tree (Felsenstein, 1978). The figure also shows that the compatible pattern 0011 has a lower posterior mean number of substitutions than 0110 for short tree lengths. However, if the tree lengths exceeded 1.64 substitutions per site, then the situation is reversed. The posterior mean of the incompatible pattern quickly approaches the tree length, whereas the mean posterior substitutions of the compatible pattern is only close to the tree length if $\Lambda \geq 54$ substitutions per site. In other words if we observe a compatible pattern, than this pattern has typically experienced more substitutions than the incompatible pattern.

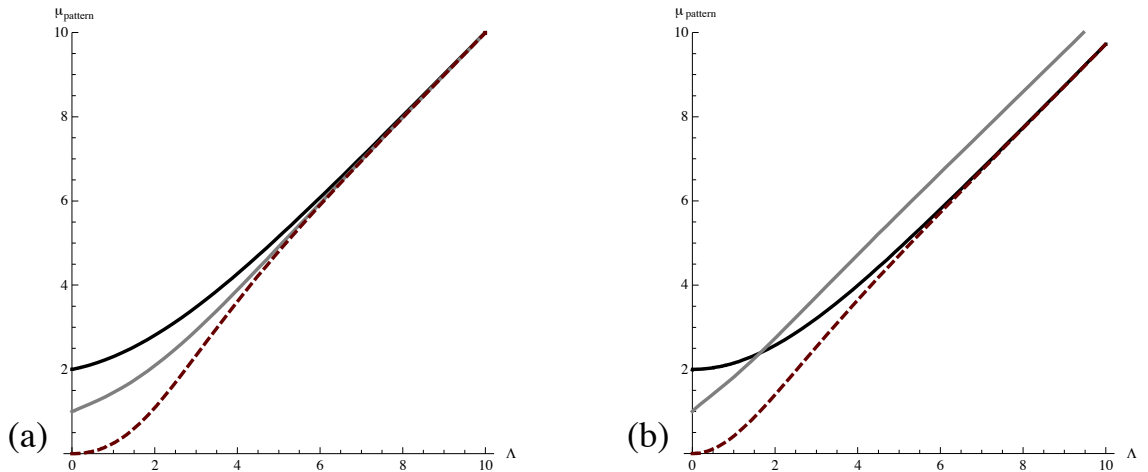


Figure 7.7: Posterior mean number of substitutions as function of the tree length Λ for the tree topology T_4 . The posterior means for patterns 0000 (red dashed line), 0011 (grey solid line) and 0110 (black solid line) are shown. Fig. (a) shows the posterior means on the tree with relative branch lengths $p_A = p_B = p_C = p_D = 0.2$ and $p_{AB} = p_{CD} = 0.1$. Fig. (b) shows the result for relative branch lengths $p_A = p_D = 0.47$, $p_B = p_C = 0.02$ and $p_{AB} = p_{CD} = 0.01$.

OSM Matrices for a Nucleotide Alphabet

While we have outlined only the most simple model of sequence evolution, several extensions are easily possible. The OSM approach can be augmented to the Kimura 3st model (Kimura, 1981); see Fig. 7.8 for an illustration. In this framework every substitution class (transition (a), transversion 1 (b) and transversion 2 (c)) uniquely generates a fix-point free $4^n \times 4^n$ -dimensional permutation matrix for each branch in a tree (a-c). Let $\alpha_1 + \alpha_2 + \alpha_3 = 1$ denote the probabilities for the three substitution classes (d), then the OSM matrix for the Kimura 3st model is defined as:

$$M_T = \sum_{K=1}^3 \alpha_K \sum_{e \in E} p_e \cdot \sigma_{\mathcal{C}(e)}^K, \quad (7.9)$$

i.e. we look at the sum of generalized permutation matrices. Fig. 7.8(e) shows an OSM Kimura matrix on a rooted triplet tree (Fig. 7.8(f)). Each row and each column contains $12 = (\text{number of branches}) \times (\text{number of substitution classes})$ non-zero entries, where each entry is the product of a mutation class parameter α_i and a branch probability p_e , Equ. 7.9. All results for binary character state models can be expanded to the Kimura 3st model.

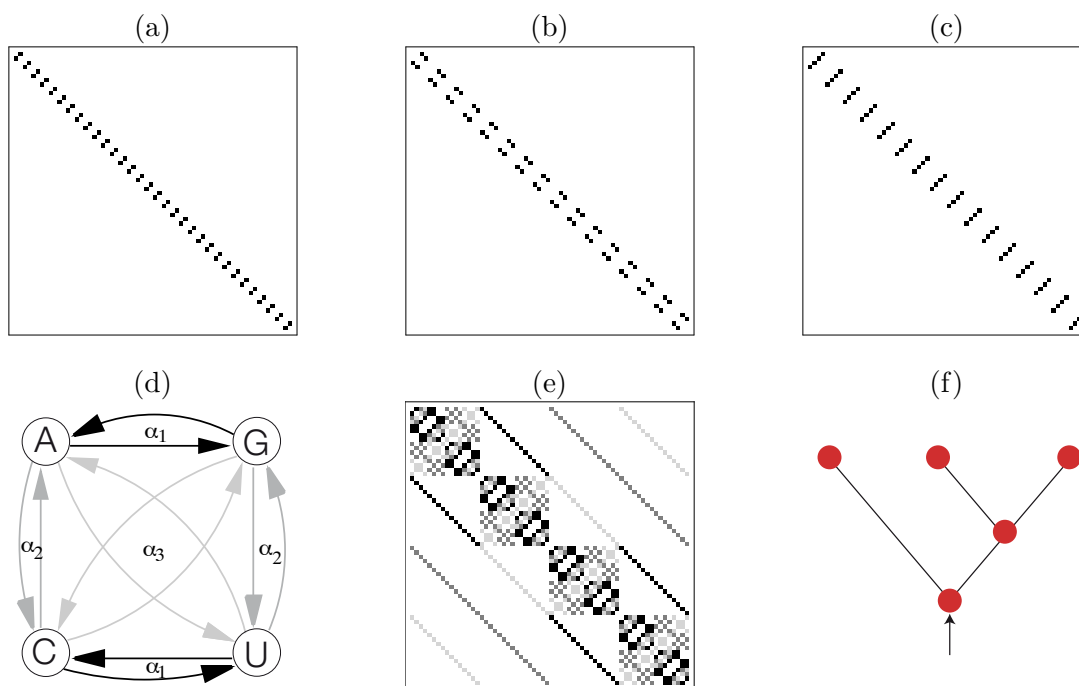


Figure 7.8: Kimura 3st model: (a-c): The three permutation matrices defined by a cluster of a rooted triplet tree T_3 (f). The white squares are zero. The transition scheme (d) and an example of an OSM matrix (e) under the Kimura 3st model of a rooted triplet tree (f). All four branch lengths are taken to be equal. Hence, black squares indicate a transition, dark gray squares an α_2 transversion and light gray squares an α_3 transversion.

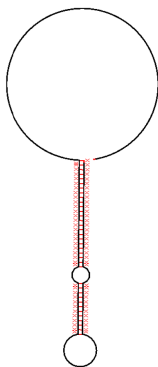
7.2 A Phylogenetic Definition in Pattern Space

Unfortunately, it seems to be difficult to include the neighbourhood system directly into the OSM matrix because this could destroy the permutation characteristic of the OSM matrices. This area requires further research. At this stage, this means we have to include the neighbourhood system implicitly into our phylogenetic definition of structure, as we did before with the tree given in the definition of Chap. 4. We recap again our basic definition of a phylogenetic structure (PS) from Chap. 4 as consisting of three aspects: a substitution matrix, a neighbourhood system and a phylogenetic tree. Here, the model specifies the evolutionary process of patterns with a focus on the third aspect of a phylogenetic definition of structure, the phylogenetic tree T .

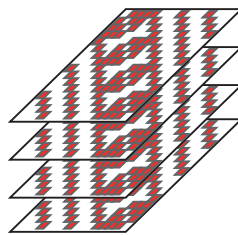
Def.: A phylogenetic structure (PS) is an abstract object which is defined by a neighbourhood system, a substitution model and a phylogenetic tree.

Secondly, in comparison to Chap. 4 the three aspects are defined as:

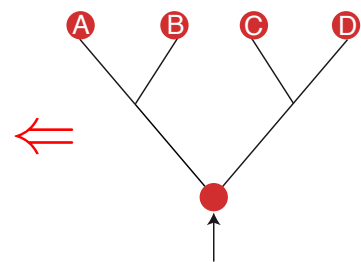
- I The neighbourhood system \mathcal{N} as in 3.1.1 with $\mathcal{N} = (N_k)_{k=1,2,\dots,l}$ for each site $k = 1, \dots, l$ in a sequence, respectively each alignment site (column) in an alignment \mathbb{A} .
- II A substitution model constitutes a collection of possibly different substitution processes acting on phylogenetic trees and an annotation of (arbitrary) scaled branch lengths as described in 7.1.1.
- III A rooted phylogenetic tree T as defined in 2.2.3.



Neighbourhood System



Substitution Model



Phylogenetic Tree

Figure 7.9: An example of a phylogenetic structure. At the left a neighbourhood system and the right a phylogenetic tree. In comparison to Fig. 4.1 of Chap. 4, here the model constitutes a collection of possible different substitution models action on trees, illustrated by the red arrow.

While we have used a sequence evolution model acting on the nucleotides in Chap. 4, this second step presents an alternative description of a phylogenetic definition of structure and merges a substitution model directly with a tree. The matrix is influenced by the tree that defines the pattern of evolution at alignment site k . The neighbourhood system assigns a phylogenetic tree T_k to a alignment site k .

That is, speaking from a simulation viewpoint: our SISSI framework of Chap. 4 includes a model that represents a universal description of arbitrary complex dependencies among sites which is finally applied recursively through a given rooted or unrooted tree topology. In contrast, we present a framework including evolutionary models on trees through the annotation of a neighbourhood system $\mathcal{N} = (N_k)_{k=1,2,\dots,l}$ for each site $k = 1, \dots, l$. This process acts on an alignment of length l . Thus, here the substitution model constitutes a collection of possibly different evolutionary models acting on patterns with an annotation of the neighbourhood system \mathcal{N} to the assignment for a phylogenetic tree T_k for each site k .

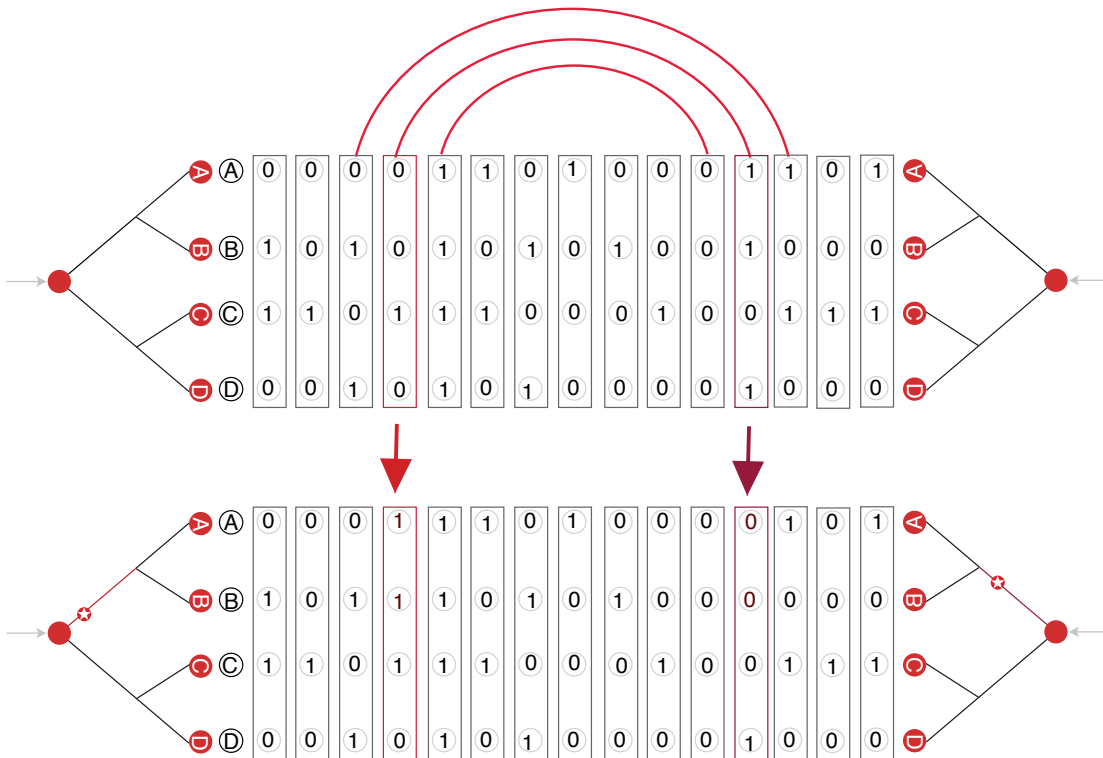


Figure 7.10: **Compensatory patterns** using base-pairing rules of binary sequences, e.g. with $\mathcal{A} = \{A, U\}$ or $\mathcal{A} = \{G, C\}$. Then, $0 - 1$ or $1 - 0$ can build a base pair. A single mutation on the tree changes the current pattern in the alignment and the outcome will depend on the branch where the substitution occurred. If the sites have site-specific interactions, *compensatory patterns* should be observed, that is we assume mutations on the same branch for both sites at the current state. As an example one compensatory pattern is highlighted with the corresponding branches on the trees in light and dark red.

So far, the OSM framework does not include neighbourhood systems directly into the OSM matrix. However, this task is more or less analogous to including rate-heterogeneity for partitions in modelling sequence evolution, e.g. for dependent and independent site or different codon positions. However, instead of looking at the process through time along the branches, we describe how a single mutation that occurs anywhere on tree T changes the character states at the leaves. Thus, if a mutation hits alignment site k , then the corresponding pattern will change to a new pattern, where the new pattern is determined according to the branch where the mutation occurs. Thus, the OSM framework has the potential to model **compensatory patterns**, illustrated in Fig. 7.10: as an example we can use simple base-pairing rules for binary sequences, e.g. with $\mathcal{A} = \{A, U\}$ or $\mathcal{A} = \{G, C\}$. Thus, $0 - 1$, respectively $0 - 1$ build a base pair, while $0 - 0$ as well as $1 - 1$ cannot build one. Then, a single mutation on the tree changes the current pattern in the alignment and the outcome will depend on the branch where the substitution occurred. If the sites have site-specific interactions, *compensatory patterns* should be observed, e.g. 1110 is a compensatory pattern to 0001. Thus, we assume mutations on the same branch for both sites at the current state of each site, illustrated in Fig. 7.10. However, that is one approach, which include only the half of an RNA helix. In future work, we want to develop a compensatory pattern framework including a four letters alphabet $\mathcal{A} = \{A, U, G, C\}$. Thus, in this chapter we have completed our definition in a second cross-step introducing another view of sequence evolution and combining that afterwards through the assignment of different trees T_k and the assignment of patterns for each site k through the annotation of the neighbourhood system \mathcal{N} . In the outlook of this thesis the potential of the completeness of a phylogenetic definition of structure is discussed.

7.3 Awesome Times

With the walk in pattern space, we have presented an alternative description how to model sequence evolution on a tree. Our approach lifts the commonly used stochastic models of sequence evolution that act on nucleotides to the set of all possible patterns for n taxa. We have shown that available tree reconstruction principles are included in our description of the process. Moreover, the definition of the OSM matrix leads to analytical formulas to compute the posterior probability distribution of the number of substitutions for each pattern. From this distribution it is then straightforward to compute the posterior mean of the substitutions.

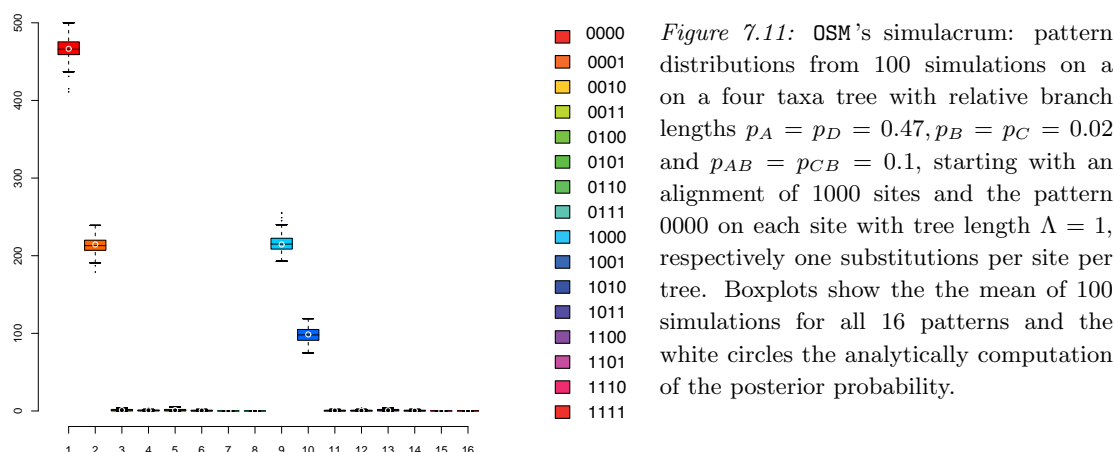
An immediate application of OSM matrices is the analytical computation of posterior probabilities that count the number of evolutionary changes on a tree. So far, these posterior probabilities have been estimated using Bayesian simulation (Nielsen, 2002; Huelsenbeck *et al.*, 2003) or by applying the theory of counting processes (Minin and Suchard, 2008). In the outlook of this thesis we discuss this application towards improvements of structure prediction programs with phylogenetic trees, as well as including structure evolution.

After the possibility of analytically computation, the question arises what kind of simulations under the OSM framework are necessary, respectively what we are doing if we put mutations on an alignment, instead of looking at the process through time along the branches. Beside the analytical computation, we have implemented a simulation program, OSM's simulacrum, see Fig. 7.11. It will be extended for nucleotides and for a process including compensatory patterns for further research on structure evolution: for example to study the influence of trees to different realisations of a PS. Furthermore, OSM's and SISSI's simulacrum together can contribute to the understanding about what can and what cannot be inferred from a sequence alignments including questions to site-specific processes and to the risk for stochastic effects, for example that act somewhere on the branches of the tree.

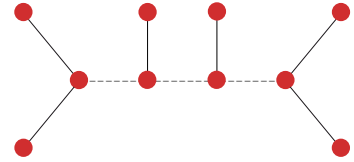
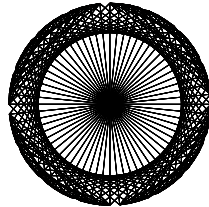
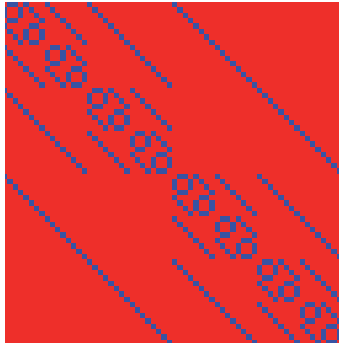
If one wants to abandon the assumption that evolution proceeds along a tree, then this is also possible within the OSM framework. Consider a set of rooted trees which give rise to a collection \mathfrak{C} of possibly conflicting clusters. The associated OSM matrix is then given by:

$$M_{\mathfrak{C}} = \sum_{\mathfrak{C} \in \mathfrak{C}} p_{\mathfrak{C}} \sigma_{\mathfrak{C}},$$

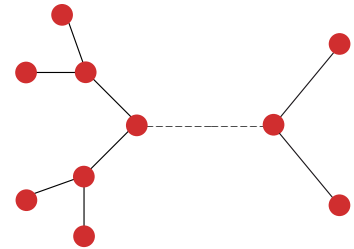
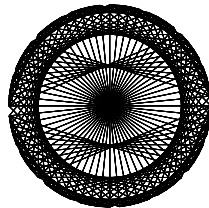
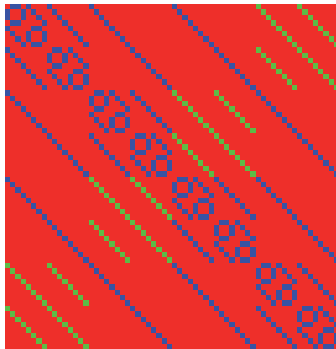
where $p_{\mathfrak{C}}$ is the normalized sum of branch lengths from those trees in which the branch depicting \mathfrak{C} is existent. Fig. 7.12 illustrates an example. This extension bears some similarity to a maximum likelihood reconstruction of networks (von Haeseler and Churchill, 1993). Furthermore, there is to the best of our knowledge so far no program for simulations on networks. It would be possible with an extension of OSM's simulacrum. Here issues like the meaning of the overall length of the cluster set, or the meaning of a root in such sets



A



B



C

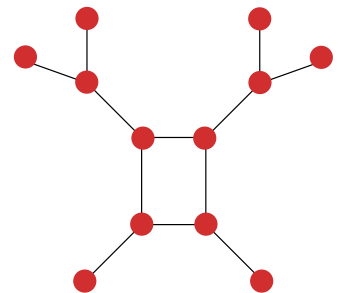
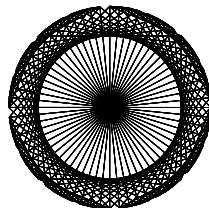
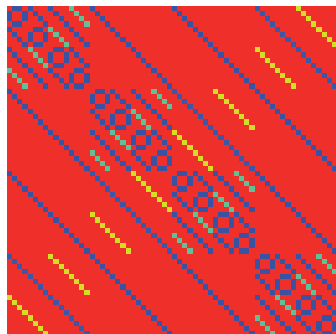


Figure 7.12: OSM's Simulacrum: If one wants to abandon the assumption that evolution proceeds along a tree, then this is also possible within the OSM framework. Here, we are considering two six taxa trees as an example. The associated OSM matrix is illustrated in C, which is the sum of $A + B$.

need to be discussed. Another question regards the Poisson weights for the number substitutions. Generally, the argument is that the process of distributing substitutions along a tree is memoryless and therefore the number of substitutions is Poisson distributed. Our framework permits to assign a different probability distribution to the number of substitution. From the viewpoint including site-specific interactions this is an important issue, where the dependencies should be taken into the weighting scheme. One possible weighting scheme could be a contagious distribution, which had previously been used to evaluate accident data (Kemp, 1967). This approach might provide alternative description of the evolutionary history of an alignment.

Finally, these frameworks lead to a cross-step description of our phylogenetic definition of structure. In the outlook of this thesis we discuss the potential of forming a complete phylogenetic description of structure.

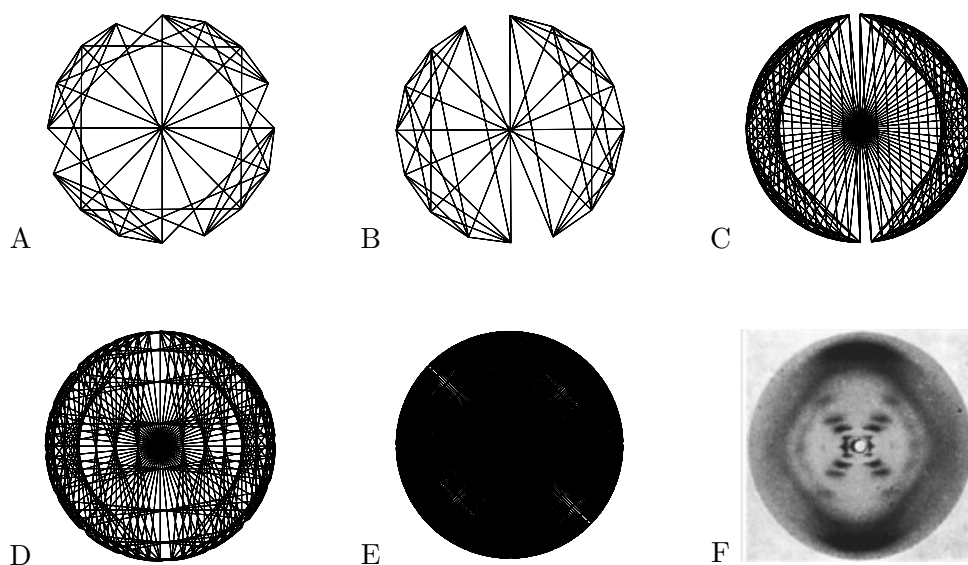


Figure 7.13: Examples of OSM adjacency graphs: This figure shows how the inclusion of trees introduces different dependencies inside the OSM matrices. A: Balanced four-taxon tree, B: Four-taxon caterpillar tree, C: Six-taxon caterpillar tree, D: Six-taxon tree under the Yule-Harding model (Harding, 1971) and E: Eight-taxon tree under the Yule-Harding model (shift 45°). F: Side note, there is a surprising similarity to the Franklin's X-ray of DNA (Franklin and Gosling, 1953).

Chapter 8



Solaris, (Tarkovsky, 1972)

Outlook

I have a little time and must tell you something and warn you. By now you know about me. If not, S. will tell you. What's happened to me is not important. Or rather, it's indescribable. I fear that this is just the beginning. I hate the idea but here it can probably happen to anyone. Only, don't think I've lost my mind. You know me well. If I have time, I'll tell you everything. If it happens to you, just know that it's not madness ... That's the main thing. As for further research, I lean towards S.' suggestion subjecting the ocean to radiation. That has been forbidden. But there's no other way. We ... you ... will only get bogged down. Radiation may get us out of deadlock. It is the only way to deal with this monster. No other way. (Tarkovsky, 1972)

TO THE OCEAN¹

¹el oceano de felicidad & his sisters

8.1 (In)complete Phylogenetic Definition of Structure

“How can we define structure ?” In this doctoral thesis, I have been focusing on this question in the context of molecular biology, more precisely RNA structure. The clear definition of discrete states for nucleotides allows a mechanistic definition of structure. I shall recap again our basic definition of Chap. 4 and Chap. 7 as consisting of three aspects.

The three aspects are defined as:

I The neighbourhood system \mathcal{N} as in 3.1.1 with $\mathcal{N} = (N_k)_{k=1,2,\dots,l}$ for each site $k = 1, \dots, l$ in a sequence, respectively each alignment site (column) in an alignment \mathbb{A} .

II A substitution model constitutes a collection of possibly different substitution processes.

4.1 acting on the sequence and an annotation of site-specific interactions among sites as described in 3.1.2.

7.2 acting on phylogenetic trees and an annotation of (arbitrary) scaled branch lengths as described in 7.1.1.

III The phylogenetic tree T as defined in 2.2.3. While in 4.1 the phylogenetic tree is a rooted tree, in 4.1 rooted as well as unrooted trees are allowed.

⇒ Def.: A phylogenetic structure (PS) is an abstract object which is defined by a neighbourhood system, a substitution model and a phylogenetic tree.

We have presented two examples how a PS may be defined. While the first and third aspect are defined identically in Chap. 4 and in Chap. 7, the second aspect, the substitution matrix, is specified as process of nucleotide evolution in Chap. 4.1 but as process of pattern evolution in 7.2.

So far I focused on the three aspects of the phylogenetic definition of structure in pairs: the substitution matrix is influenced by the neighbourhood system that defines the interactions among the sites in 4.1, while 7.2 merges the substitution matrix directly with a phylogenetic tree. The next stage is to bring all three aspects together to see if they can fit into one substitution model (Fig. 8.1). This will require further research especially in relation to the question: *how far does the model explain a structure ?*

The main goal for future research is to include all aspects directly into one substitution matrix, though it is neither clear if that is possible nor if it would explain PS adequately.

Neighbourhood System

Substitution Model

Phylogenetic Tree

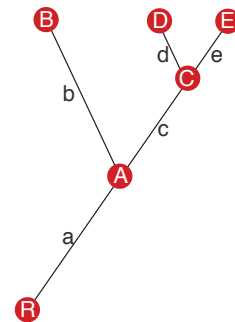
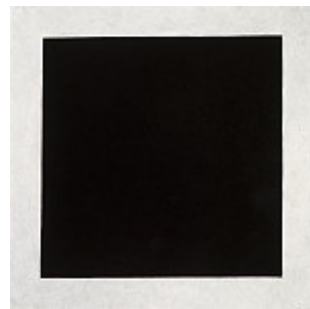
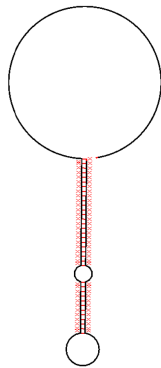
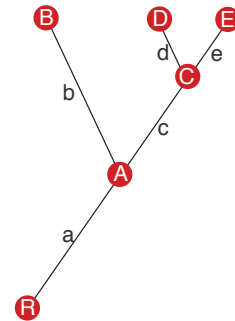
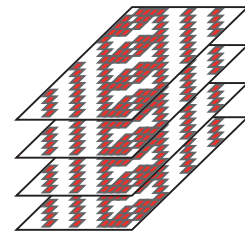
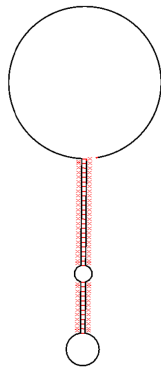
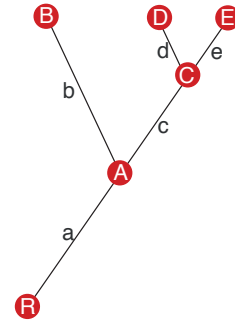
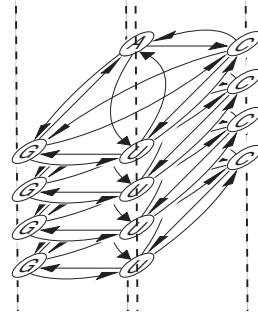
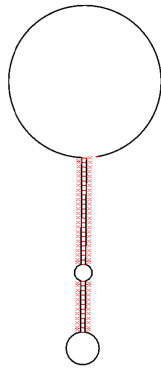


Figure 8.1: Definition: A phylogenetic structure (PS) is an abstract object which is defined by a neighbourhood system \mathcal{N} , a substitution model and a phylogenetic tree T . **Left:** Neighbourhood System; **Right:** Phylogenetic Tree; **Middle:** The substitution model constitutes a collection of possibly different substitution models acting on correlations among sites illustrated in the first row and acting on phylogenetic trees illustrated in the second row. Details are described in 8.1. However, the main goal for future research is to include all aspects into one substitution model directly. This is illustrated in the middle of the last row with the Black Square (Kazimir Malevich, 1915, Oil on Canvas, State Russian Museum, St.Petersburg).

8.2 Towards Structure Evolution (‘‘Happy Together’’)

In Chap. 4, we focused the definition 3.1.2 of a PS to existing definitions of RNA structure (see Tab. 2.2), illustrated in Fig. 4.2. Here, I extended this figure to Fig. 8.3: It is known that sequence evolution is subject to noise, e.g. based on short sequences or different tree topologies. Deterministic approaches cannot capture the potentially significant effect of factors that cause stochasticity. Indeed, it seems clear that even with a phylogenetic definition of structure the description will be incomplete. However, there is a toolbox with which I can approach my point of interest (PI) (Fig 8.2 and Fig. 8.3).

Incomplete happy, I cannot find the PI due to the noise between the realisations of a PS, but with *guidable* tools I can converge around this point closer. In future research, we need theoretical tools to understand the mechanism of the intertwined relationship between the three aspects of a PS, as well as practical tools with *multifunctional* task on realisations of a PS, illustrated with the wall light of Curt Fischer in Fig. 8.3.

In this sense, I have devoted to each aspect a corresponding chapter, respectively a corresponding approach. Together they provide very promising possibilities in future work to combine all the aspects in an heuristic manner which we have currently dealt with separately in pairs of aspects in Def. 3.1.2 and 7.1.1 .

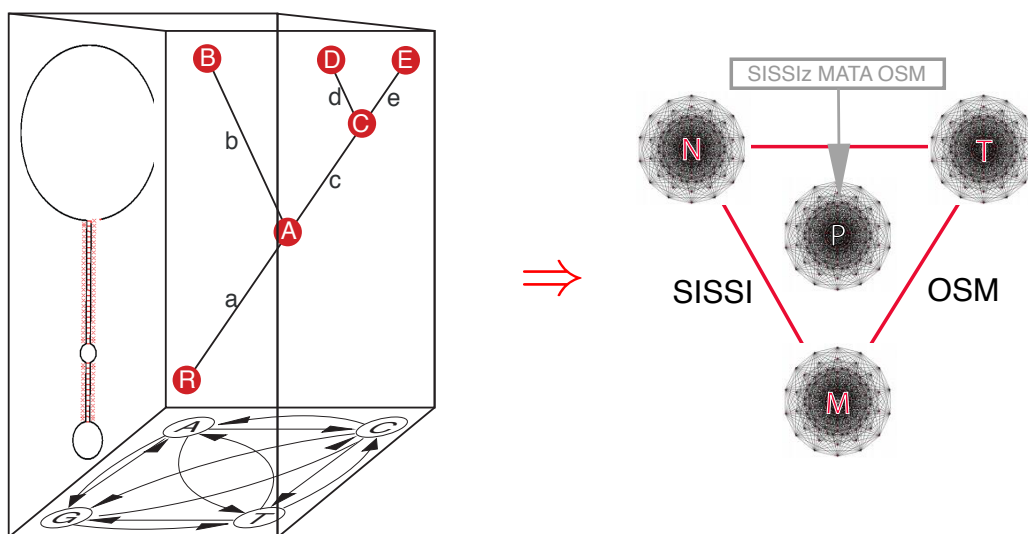


Figure 8.3: In future research, we need theoretical tools to understand the mechanism of the intertwined relationship between the three aspects of a PS (left), as well as practical tools (right) with multifunctional tasks and guidable light on realisations of a PS. In this thesis, we have developed three frameworks to consider observed or simulated sequences. Two approaches, **SISSI** and **OSM** were developed to generate sequences. They provide very promising possibilities to combine all the aspects which we have currently dealt with separately, see Fig. 8.1.

From SISSI to SISSIz to ...

SISSI simulated the evolution of a nucleotide sequence along a phylogenetic tree taking into account site-specific interactions. Thus, it is possible to mimic sequence evolution under structural constraints. SISSIz employs the SISSI framework to generate accurate background models for comparative genomic screens. We have combined the new “null” model with a consensus folding algorithm directly, resulting in a new variant of a thermodynamic structure-based RNA gene-finding program that is not influenced by the dinucleotide content.

Such a randomisation algorithm is not only of interest in the context of RNA gene prediction. It can be used for other comparative genomics applications whenever random alignments are needed as control. One could consider other applications in the context of RNA structures (e.g. prediction of conserved miRNA target sites) but also in different contexts (e.g. conserved sequence motifs). SISSI has the power to simulate a variety of evolutionary scenarios efficiently. So far, we use SISSI as an application for sequence evolution, where the states at the nodes are the nucleotides or amino acids. However, we are well aware that the mathematical and computational approaches are also applicable to the evolution of protein interaction networks, gene regulatory networks, and molecular networks in general.

SISSI & MATA

Moreover, we have also developed MATA that allows the inference of the neighbourhood system from a multiple sequence alignment. While SISSIz, SISSI serve as an adequate null hypothesis for gene prediction with a consensus algorithm, we want to combine SISSI with MATA under advanced covariance measures to directly improve structure prediction programs with an accurate threshold in an automatic manner.

Although, it is well known that a high amount of ncRNAs are based on conserved secondary structure with minimum free energy (following a series of algorithm discussed in Chap. 6), the phylogenetic definition of structure is independent of thermodynamic optimal structures. This was demonstrated in Chap. 4, where we generated sequences with a conserved neighbourhood system showing a low thermodynamically structure conservation index. Thus, other constraints could be important for this kind of structure. SISSI is general enough to generate sequences with any kind of constraints annotated by the neighbourhood system and MATA is flexible enough to include SISSI as a generator for an arbitrary (accurate) null-hypothesis.

MATA & OSM

Although MATA has a small amount of false positive predictions, the sensitivity is low. Presently, it measures the deviation of the frequency of the pair of nucleotides from its expected frequency given an estimated pair at the root sequence on the tree. An analytical formula of the pattern distributions does not seem feasible. Similarly, MATA’s second test is based on a parametric bootstrap approach to suggest positions “without ancestry”. The OSM framework allows the analytical computation of pattern distributions in an alignment. It counts the number of evolutionary changes on a tree, more precisely also on specific branches. Following Dutheil *et al.* (2005), a Pearson correlation coefficient between two corresponding substitution vectors in comparison to the expectation under the null hypothesis of independence can measure the amount of coevolution. Moreover, an accurate combination of OSM and MATA might improve the results of MATA and can feed into structure evolution, which is discussed below.

OSM & SISSI’s Simulacrum

In phylogenetics, the usages of Markov models is not only necessary for accurate reconstructions of phylogenetic trees, but also to check the model’s fidelity in reflecting the evolution of the sequences (see Sec. 2.3.3).

In the future, we are interested in the evaluation of different kinds of models and discussions about the simplicity versus the complexity of models. The use of supervised sequence evolution allows us to control and study the extent of structural and sequence conservation. Generated data using OSM in comparison to SISSI’s data can reveal the influence of the different aspects of a PS in future research. For example, additional work will be required to sort out which are the most important context effects, and whether simpler parameterizations or context-dependent rate matrices can be justified, e.g for phylogenetic inference under structural constraints and how the mutations are distributed on the tree.

Generating data including indels with site-specific interactions

While bioinformatics models of the insertion–deletion process are being developed without taking site-specific interactions into account, not much is known how site-specific interactions influence this process. So far we have suggested a first algorithm for SISSI to include an insertion deletion process (see Alg. 3.4.1). Such approaches will improve background models for comparative genomics, e.g. SISSIz. However, one important keypoint will be to obtain the empirical distribution for the length of insertion and deletions or the frequency of insertions and deletions with site-specific interactions. Getting simultaneous structural RNA sequence alignment, structure prediction and phylogenetic reconstruction together is still a problem. Generating data including indels with site-specific interactions might help to understand the underlying mechanism. For example, how the indel-process influences a neighbourhood system. So far, we have assumed that the neighbourhood system does not change during evolution. The indel process will lead to considerations of lineage specific structures.

8.2.1 PS:

PS Towards “PS Families”

A PS contributes to the understanding of divergence mechanisms of “families” through reconstruction of the evolutionary history. Long term evolution often includes dynamic changes, such as insertion, deletion and consequently a change of the neighbourhood system. Based on our phylogenetic definition of structure, we call these changes lineage specific neighbourhood systems (LSN). A phylogenetic definition of structure is a prerequisite for a definition of LSNs. In Chap. 4 we distinguished between ancestral and neighbourhood constraints. Thus, lineage specific evolution might be also possible under the same neighbourhood system. However, this might be an intertwined process. How can ancestral correlations influence the function of the molecule and thereby the evolutionary process itself, e.g from a selection viewpoint? Are only functional correlations highly conserved in evolution? Can ancestral correlations also lead to ”real” functional constraints? In future work we would like to extend test statistics that detect structural change points in a phylogenetic tree. From a practical viewpoint this approach is important for studies on the evolution of RNA genes. Structure prediction programs, including RNA gene finders, which are based mostly on structure conservation, can be improved by finding lineage specific and evolving structures. Based on the (in)complete phylogenetic definition of structure, it is not clear, if a phylogenetic structure conservation index (PSci) is definable. However, a heuristic framework for modeling a phylogenetic structure, as well as detecting changes of a PS, should combine the features of **SISSI**, **MATA** and **OSM**. From our viewpoint, future work should further connect the fields of structure prediction and phylogeny to address RNA and other gene families, in addition to the question of how the classical phenotypes and genotypes influence each other during evolution.

PS Towards “a Universal PS Framework”

A further step is the consideration of a PS *in vitro*. An example of a collaboration with experimentalists (Renée Schroeder Lab, MFPL) is ongoing, where we evaluate the evolution of RNA sequences under SELEX *constraints*. Genomic SELEX (Lorenz *et al.*, 2006) is a derivative of the widely used SELEX *in vitro* screen (Tuerk and Gold, 1990) used to identify functional RNAs. We observed in a genomic SELEX experiment of an *E.coli* genomic library to detect RNA sequences which bind Hfq with high affinity (Lorenz *et al.* unpubl.) less structurally stable sequences than one would expect. This raises the question how to define a PS for a SELEX experiment.

Furthermore, a phylogenetic definition of structure provides the theoretical framework for the modeling of the evolution of more complicated networks. A PS presents a universal description for arbitrary complex neighbourhood systems in a unifying framework, which is flexible enough in terms of simplicity and complexity. RNA sequences are one example to study the evolution of complex processes. However, many other sequence data are available to work with.

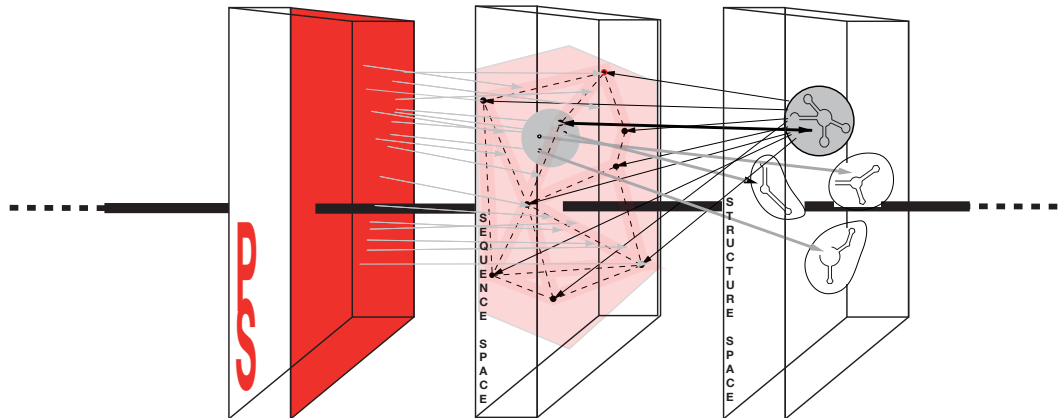


Figure 8.4: Towards an extension of the traditional sequence structure map. Left: this space represents the phylogenetic definition of structure given in this thesis; middle: classical sequence space, right: classical structure space. A neutral path (black) line is mapped from the structure space into the sequence space. The right part of the figure showing the schematic view of the RNA folding map is taken from Hofacker (1994). Here, the left space projects another neutral path at the sequence space, which could be described as an extension of the neutral path of the structure space. We call this path: *a neutral phylogenetic structure path*.

PS Towards “PS Evolving Structure Maps”

A PS can contribute to further considerations on how the classical phenotypes and genotypes influence each other during evolution. Fig. 8.4 shows the traditional sequence structure map for RNA secondary structure at the right side (e.g Fedoroff and Fontana, 2002; Fontana *et al.*, 1993b; Schuster and Stadler, 2007, and see Sec. 2.3).

A phylogenetic definition of structure could be interpreted as a third space that implies also a different concept of fitness. The PS allows neutral paths in the sequence space, which are different to the well known neutral paths projected from the structure space. Thus, we called it the phylogenetic neutral path. This path depends on the phylogenetic tree, the model and the neighbourhood system. Further extensions are possible, e.g coarse gained structure could be included as a fourth space behind the structure level. However, we classified that at the description level in Fig. 8.2.

On the other hand, where the neighbourhood system comes from, directly raises the question of

an Origin of Structure & an Origin of Life ?

8.3 WHAT IS A STRUCTURE

This thesis is only our starting point in collecting structure definitions for two current fields of computational biology closely related to RNA. Further research should clarify the meaning of structure in general and between scientists of different disciplines. This is a cross-disciplinary issue. As a follow-up, a short self-organized project at the CSSS08 (Complex System Summer School 2008 by the Santa Fe Institute, US) considered: molecular biology, philosophy, logic, architecture, computer science, software engineering, information theory, physics and psychology. The project started to elaborate shared features, attributes and themes of these definitions and it studied the history of this concept how it has evolved within these different disciplines. One result of this comparative analysis was that the definition of structure is continually evolving in each individual discipline.

For further research, and towards a general understanding of structure in the sense of an universal description of structure, the following fields might be interesting: Logic (cf. Dalen, 2004), Category Theory (cf. MacLane, 1971) and Complexity Theory (cf. Parisi, 1994; Atay *et al.*, 2008). In late nineteenth century advancements in mathematical logic proved to have an important impact on the concept of structure as it was applied within philosophy. First order logic offered philosophers a rich formal language in which to represent logic forms of various entities, ranging from sentences to scientific theories, thus allowing reasoning involving these entities to be represented within a logic (Russell, 1905; Frege, 1879, 1892). At the moment, concepts which suggest that the structures of a given sequence might be associated with Turing machines (TM) that are capable of producing the sequence as output (Wiener, 1994; Koppel, 1994) seem to offer one of the most appropriate concept of structure. To summarize Wiener's quote in Chap. 1: a structure is a sequence of a TM, which generates or accepts this sequence. This computational approach gives a useful foundation upon which to base a general understanding of structure.

In order to build upon this foundation, we shall attempt to specify the sequences/data and machines/programs that are present in those disciplines in which the concept of structure is used. Thus, in future research, the results of this thesis should be considered in these concepts. Wiener is using Turing Machines like Koppel for his structure concept, but in "*Thinking Turing Machines*". Thus, structures are 'attributed' and the object with the "attributed" structures only exists in organisms capable of thought.

In this thesis we have given only an example how RNA structures are attributed to concepts of two current fields of computational biology and their methodology at different abstract levels. We end with the answers of a survey among CSSS08 summer-school participants from different disciplines about the question of defining a structure.

WHAT IS A STRUCTURE ?

a reoccurring pattern of relationships – the arrangement and relations between the parts or elements of something complex – trying to find that out — an organizing concept – er, too many definitions! – it’s a class for variables in c.... it’s an organization of things... – it depends. in networks could be a collection of nodes and the connections between them. if in the context of a model is the relationship between variables . the idea is that a structure has a property of ‘permanence’, even if transitory... – it is a (any) departure from randomness. or, for instance if we are talking about the correlation structure among a set of variables, or the network structure (among, say, people), then it’s the general term describing what we’re going to look for—regardless of what we find (e.g., which links are actually present). – something consistent that aims at gathering people/logistics in a definite goal – a framework which controls how a system works – a set of relations between some collection of objects that a) clearly delimits this collection from the collection of all objects and b) organizes the interactions and dynamics within the collection of objects and between that collection and the extra-structural universe. – a structure is any system which defines the way in which objects are grouped or related. it is a specific reasoning framework. – good question – do not understand the question. – simulation – an object in some programming languages; an edifice; – a description for regularities in the world – describes how parts of a system are related/(inter)dependent and maybe how they fit together to add up to something that is greater than just the sum of the parts – no idea – something that allows you to describe a system in a more general or a more compact way – how am i supposed to answer this in one line ? – a coherent entity – the connections between individual units (that compose a larger whole) that determine the stability of the whole. – the result of the interlinking of nodes in a network, where dynamics can travel on. the parameters of these structures are from significant importance for the dynamics in such networks. – abstraction that allows us to model something in terms of pieces that are easier to understand. – a structure is a recognized pattern that serves some function. – structure can also refer to the general properties that a set of items have in common. for example, the structure of a tree would be roots, trunk, branches, leaves, etc. – a entity containing information – a framework to allocate and manage workload within a system – something that defines the nature of interaction and behaviour of “elements” in a system. for example, social structure. – any organizing macro-event. – that’s hard! – a collection of rules or constraints; a object that demonstrates a given set of rules or satisfies a set of constraints. – complex systems with certain elements... – where the interactions between entities in the system are different from some null hypothesis (exactly what depends on what questions you are asking - e.g. no interactions, random, every interaction exists) – word with different meanings. – if you know the answer, tell me. – a form, law, institution or norm that bounds activity or behavior – i’d wish you tell me :) – this question is way too ambiguous – a menhir – a type of formal interpretation which consists of a set, functions, and relations defined on the set. – a mathematical object or a data storage object in programming. – heterogenous parts that form patterns of interaction in some stable environment? – nothing – an underlying form that allows for the understanding of an object without explicit reference to its superficial qualities. WHAT IS A STRUCTURE ?

Appendix



Medusa, S. Boehl, 2004

PS

“nicht mehr als fragmente der dinge sind die dinge”

*The title, which when isolated may appear programmatic, is a quote from Léger’s writings and painting (Léger, 2005) and refers to the fragmentary structure of my work. The works can be understood as both complete in themselves and as imaginary, continuous – in the all over sense. And of course Léger, with his reference to other artists such as Cézanne, is also as an artistic position, as a station of conflict in my artistic stance which I perceive as monadic, which is both relevant basis and reference parameters for my own work . . . Given that everything is visible nobody can work artistically any more without being present in that which was before (Boehl, 2008). **TO SABINE***

The choice of works came about as a result of the possibilities which a scientific academic system offers and my search for a phylogenetic definition of structure (PS) towards my own point of interest (PI). I have to thank my supervisor *Arndt von Haeseler* for all the fights with and without words – with and without the other – to select and present the work in an adequate scientific way. In this sense I am also very happy about my two reviewers, *Ivo Hofacker* and *Nick Barton*, independent of the various results in the work - both place great importance on a well-founded knowledge of science.

Following a PS, my work is based on dialogues with scientists and artists, as well as persons from other fields with all the references to ancient work. Thus, each chapter is dedicated to a friend and an inspiration, which I will not recap here (beyond words). “My interest in these forms is non-historical and interpretive”. Furthermore in analogy to *Sabine Boehl*, my own intention lies in interweaving the signs to approach my point of interest (PI), although the definition of signs is (here somehow) difficult. Relating to RNA, I got my first infection through a paper of *Peter Schuster*, devoted to the birthday of *Manfred Eigen*, in a seminar about “signs” by *Oswald Wiener* at the art school in Duesseldorf.

Widespread

My work with a clear definition of (RNA) signs is based on the clear and helpful teaching of *Gerhard Steger* (University of Duesseldorf) and his references to other RNA people such as the *Vienna Group*. Coming to Vienna, these structures are (re)generated from individual modules of the whole *TBI* and seminars with the whole *Bompfünowerer Consortium*. In this warm environment my infection increased exponential. In particular, *Christoph Flamm* has taken care of me, and fancy RNA researchers such as *Dill, Rainer, Caro and Andrea* supported me any time. Last but not least I got a hard effect through a special and unique collaboration with *Stefan Washietl*.

Inspiring is not a matter of time, sometimes a short lineage specific neighbourhood can have a huge influence on the current diversity, e.g. short moments with: *Christian Reidys* stimulated to consider the connection between tree space and neighbourhood space and *John Mattick* established fatherly the complexity of RNA.

This onset was opposed by another one, the desire for phylogenetic relationships. Between my diploma and my PhD studies, just with less RNA occurrence, there was a risk to get an high phylogenetic infection in the Goldman Group (EBI, Hinxton, Uk). With the help of *Carolin Kosiol, Ari Loytynoya, Simon Whelan* and *Nick Goldman*, I implemented a simulation program for proteins with a pairing parameter as a function of the amino acid distance with a set of happy and unhappy combinations. With this analogy I got a huge phylogenetic stimulation in a nice group atmosphere. I searched through different phylogenetic conferences to satisfy my phylogenetic desire and finally, I have given my best at the institutes of *Arndt von Haeseler* in Duesseldorf and later in his CIBIV institute in Vienna. There, I cooperated mainly with *Thomas Schlegel* and *Steffen Klaere* supervised by *Arndt von Haeseler*, with *Matthias Dehmer* and *Martin Grabner* we build our own rules and, finally, with delicious *vietnamese food*, a *portuguese cure* and *H. Schmidt's* computer skills, I survived all infections.

My sincere thanks goes to all my fellow labmates from the CIBIV institute and *everybody* at the whole Biocenter: the good, the bad and the ugly, as well as the fairylike one. Moreover, from Duesseldorf my oldest fellow students, especially *Cynthia Sharma* and *Simone Linz* for their solidarity and friendship, *Andreas Wilm* and *Ingo Paulsen* for giving advice on the usage of computers, as well as the new head of the bioinformatic institute, *Martin Lercher*, for many sheet anchors in my home waters. Finally, I warmly thank *Alison Flint* for being more than a normal English teacher and *Korinna Thielen* and *Andrea Ulrich* for sharing visual perceptions and graphical skills. My special thanks goes to *Caroline Kosiol* and *Roland Fleissner*, which have read carefully parts of my thesis and shared all their knowledge with me independent of each infection and environment.

A PS to *in vitro* Aspects

The chapter *in vitro* aspects is an example, which disappear for various reasons, although it is no less important. A major aspect of my thesis is to focus and clarify the theoretical framework of a PS. However, as already described in the outlook, a further step is the consideration of a PS *in vitro* and a lot of work is already ongoing from different perspectives, for instance toward evolution of RNA sequences under SELEX constraints. Generally, I am sure that these experiments will help me to extend my theoretical framework to a PS *in vitro* in the near future. Thus, I want to thank the people involved of the *Schroeder Group* (MFPL): *Ivana Bilusic, Doris Chen, Christina Lorenz, Ursula Schöberl, Christina Waldsich, Robert Zimmermann and all the others.*

I have to thank *Renée Schroeder* for this lab story during the last years of my PhD and all her youthful enthusiasm. She introduced me not only to her lab work and her group, moreover, she has given me the possibility to present my work to a broader interdisciplinary audience. Although, all realisations might be appear highly diverse and different as intended (e.g. dependent on materials, environments and neighbourhoods), I would like to express my sincere gratitude to all the contacts and the experience for my future.

PS to Degeneration and Regeneration

Works are created which find themselves in permanent degeneration and regeneration.

For instance, during a project about a random walk in sequence space, I lost my way and it disappeared. However, it might be useful for the evolution of small RNAs in future research. Moreover, I am very eager to awesome projects, e.g about dynamic landscapes with a model of context and contingency in evolution², and discussions, e.g about structure, with the great summer school friends from Santa Fe such as *QiQi, Kathleen, David, Christopher, Jacob and Molly.*

For me, fragments are components in that they have to be related to my own work, which are mostly based on a PS and the PI³ at the moment. Thus, my work is predefined by the other and vice versa. Everybody has to define his or her place in which the work is fixed in themselves and “*structures*” are (re)generated from individual (PS) modules.....

I have tried to my best knowledge to describe a phylogenetic definition of structure and my point of interest. However, please contact me in case of any confusion.

²http://www.santafe.edu/events/workshops/images/b/b7/Evopaper_update.pdf

³point of interest, see Fig. 8.2.

Material and Methods

Software

For implementation of our **SISSI** software we would like to thank Andrew Rambaut for allowing us to use some code from Seq-Gen <http://tree.bio.ed.ac.uk/software/seqgen>. **SISSI** with energy is implemented using the C code library of the Vienna RNA package <http://www.tbi.univie.ac.at/~ivo/RNA>. In Chap. 4 we analysed structure predictions under the mutual information context and thermodynamic consensus matrix using **ConStruct** version 3.2.4 <http://www.biophys.uni-duesseldorf.de/construct3/>. We would like to thank Andreas Wilm for a fast version of **ConStruct** for our simulation studies and several scripts such as `comparect.pl` by Gardner and Giegerich (2004) and `scif.pl` for a fast computation of the structure conservation index. For the simulations in Chap. 6, Fig. 6.2 we used **SISSI** version 1.0 and `seq-gen` version 1.3.2 <http://tree.bio.ed.ac.uk/software/seqgen>. Mononucleotide shuffling was carried out using `shuffle-aln.pl` with option “`--mode conservative2`”. Together with `alifoldz.pl` it is available online <http://www.tbi.univie.ac.at/~ivo/RNA>. For the tests in Figs. 6.2 and 6.8 we used **RNAalifold** from the Vienna RNA package <http://www.tbi.univie.ac.at/~ivo/RNA> version 1.6.1. with options “`-nc 0 -cv 0`”) and **RNAz** <http://www.tbi.univie.ac.at/~wash/RNAz> version 1.0 with standard parameters. For implementation of our **SISSIz** software we used a series of third party C-code that is available as open source: **levmar** <http://www.ics.forth.gr/~lourakis/levmar> by Manolis Lourakis for least squares fitting, **BIONJ** <http://www.lirmm.fr/~w3ifa/MAAS/BIONJ/> (Gascuel, 1997) by Olivier Gascuel, **PHYML** <http://atgc.lirmm.fr/phyml> by Stéphane Guindon and Olivier Gascuel for maximum likelihood estimation of the transition/transversion rate, Vienna RNA package <http://www.tbi.univie.ac.at/~ivo/RNA> by Ivo L. Hofacker and others for consensus folding in **SISSIz**.

Sequence Data

For the simulation studies of Chap. 4 the archeabacteria 5S RNA alignment and the corresponding structure was taken from the 5S ribosomal RNA data bank (Szymanski *et al.*, 2000). The performance of **MATA** was tested on a secondary structure of the *Bacillus sub-*

tilis riboswitch (Batey *et al.*, 2004) and a sequence alignment of 111 bacterial sequences provided by Gerhard Steger (Gräf *et al.*, 2005). For the benchmark of **SISSIZ** we used sequences from the following eight Rfam families: RF00001 (5S rRNA), RF00004 (U2 snRNA), RF00005 (tRNA), RF00008 (Hammerhead ribozyme), RF00012 (U3 snRNA), RF00020 (U5 snRNA), RF00029 (Group II intron), RF00104 (mir-10 precursor). From these sequences, a set of non-redundant alignments between 3 and 6 sequences per alignment and mean pairwise identity between appr. 50% and 100% was created as described (Washietl *et al.*, 2005c; Washietl and Hofacker, 2004). The families were chosen because they represent different structural families and contain enough sequences to create sets of reasonable sample size.

Genomic alignments were extracted from Multiz 17-way vertebrate alignments available at the UCSC genome browser <http://genome.ucsc.edu>, (Karolchik *et al.*, 2007). For creating the set of 1000 alignments used for Figs. 6.6 and 6.8, we used the `rnazWindow.pl` script from the **RNAz** software package www.tbi.univie.ac.at/~wash/RNAz (Washietl, 2007) to get typical alignment blocks as used previously in genomic ncRNA screens (e.g. Washietl *et al.* (2007b) or Rose *et al.* (2007)). For the benchmark we selected for each positive example of the structural RNA set a negative example from the genomic alignments. Subsets of sequences were chosen to get the same number of sequences and the same mean pairwise identity (± 0.05) as the structural RNA counterpart. Also the alignment length was adjusted accordingly (limited to a maximum length of 150).

Bibliography

- A. F. Bompfünewerer Consortium, Backofen, R., Bernhart, S. H., Flamm, C., Fried, C., Fritzsche, G., Hackermüller, J., Hertel, J., Hofacker, I. L., Missal, K., Mosig, A., Prohaska, S. J., Rose, D., Stadler, P. F., Tanzer, A., Washietl, S. and Will, S. (2007) RNAs everywhere: genome-wide annotation of structured RNAs. *J Exp Zool B Mol Dev Evol*, **308**, 1–25.
- Akmaev, V., Kelley, S. and Stormo, G. (1999) A phylogenetic approach to RNA structure prediction. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **7**, 10–17.
- Akmaev, V. R., Scott, K. T. and Stormo, G. D. (2000) Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics*, **16**, 501–512.
- Altschul, S. F. and Erickson, B. W. (1985) Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol*, **2**, 526–538.
- Amaral, P. P., Dinger, M. E., Mercer, T. R. and Mattick, J. S. (2008) The eukaryotic genome as an RNA machine. *Science*, **319**, 1787–1789.
- Ancel, L. W. and Fontana, W. (2000) Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.*, **288**, 242–283.
- Andersen, E. S., Lind-Thomsen, A., Knudsen, B., Kristensen, S. E., Havgaard, J. H., Torarinsson, E., Larsen, N., Zwieb, C., Sestoft, P., Kjems, J. and Gorodkin, J. (2007) Semiautomated improvement of RNA alignments. *RNA*, **11**, 1850–1859.
- Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H. and Murphy, K. P. (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, **23**, 19–28.
- Anisimova, M. and Kosiol, C. (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol. Biol. Evol.*, **26**, 255–271.
- Aris-Brosou, S. and Excoffier, L. (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol. Biol. Evol.*, **13**, 494–504.

- Arndt, P. F., Burge, C. B. and Hwa, T. (2003) DNA sequence evolution with neighbor-dependent mutation. *J.Comput.Biol.*, **10**, 313–322.
- Atay, F., Jalan, S. and Jost, J. (2008) Randomness, chaos, and structure. *Chaotic Dynamics*, pages 1–11.
- Axmann, I. M., Kensche, P., Vogel, J., Kohl, S., Herzog, H. and Hess, W. R. (2005) Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol*, **6**.
- Babak, T., Blencowe, B. J. and Hughes, T. R. (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, **8**, 33.
- Baele, G. (2009) *Detecting complex substitution patterns in non-coding sequences*. Ph.D. thesis, University of Gent.
- Baele, G., Van de Peer, Y. and Vansteelandt, S. (2008) A model-based approach to study nearest-neighbor influences reveals complex substitution patterns in non-coding sequences. *Syst. Biol.*, **57**, 675–692.
- Bartel, D. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–97.
- Batey, R. T., Gilbert, S. D. and Montange, R. K. (2004) Structure of a natural guanine-responsive riboswitch complex with the metabolite hypoxanthine. *Nature*, **432**, 411.
- Bennington, G. (1997) *Politics and Friendship. A Discussion with Jacques Derrida*. Centre for Modern French Thought,, University of Sussex,.
- Bérard, J., Gouéré, J. B. and Piau, D. (2008) Solvable models of neighbor-dependent substitution processes. *Math Biosci*, **211**, 56–88.
- Bernhart, S., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. and Hofacker, I. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, **1**, 3.
- Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R. and Stadler, P. F. (2008) RNAali-fold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
- Bindewald, E. and Shapiro, B. A. (2006) RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, **12**, 342–352.
- Bishop, M. J. and Thompson, E. A. (1986) Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.*, **190**, 159–165.

- Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D. and Miller, W. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, **14**, 708–715.
- Boehl, S. (2008) *Nicht mehr als Fragmente der Dinge sind die Dinge*. Kunstverein Arnsberg, Gemany.
- Bona, M. (2004) *Combinatorics of Permutations*. Chapman Hall-CRC, Halle.
- Brachet, J. (1941) La localisation des acides pentosenucliques dans les tissus animaux es les oeligufs d’amphibiens en voie de l’eveloppment. *Arch. Biol.*, **53**, 1941.
- Brown, J. W. (1999) The ribonuclease P database. *Nucleic Acids Research*, **27**, 314.
- Bruno, W. J. (1996) Modeling residues usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.*, **13**, 1368–1374.
- Bulmer, M. (1986) Neighboring base effects on substitution rates in pseudogenes. *Mol. Biol. Evol.*, **3**, 322–329.
- Busch, A. and Backofen, R. (2007) INFO-RNA—a server for fast inverse RNA folding satisfying sequence constraints. *Nucleic Acids Res.*, **35**, W310–313.
- Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–63.
- Cartwright, R. A. (2005) DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics*, **21 Suppl 3**, i31–38.
- Cate, J., Gooding, A., Podell, E., Zhou, K., Golden, B., Kundrot, C., Cech, T. and Doudna, J. (1996) Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science*, **273**, 2–33.
- Cavender, J. A. (1978) Taxonomy with confidence. *Math. Biosci.*, **40**, 271–280.
- Chiu, D. K. and Kolodziejczak, T. (1991) Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.*, **7**, 347–352.
- Chomsky, N. (1959) On certain formal properties of grammars. *Inform. Cont.*, **2**, 137–167.
- Choudhur, S. (2003) The path from nuclein to human genome: A brief history of DNA with a note on human gen sequencing and its impact on future research in biology. *Bulletin of Science Technology Society*, **23**, 360–367.
- Christensen, O. F. (2006) Pseudo-likelihood for non-reversible nucleotide substitution models with neighbor dependent rates. *Stat. Appl. Genet. Mol. Biol.*, **5**, 1–29.

- Christensen, O. F., Hobolth, A. and Jensen, J. L. (2005) Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *J. Comput. Biol.*, **12**, 1166–1182.
- Clote, P., Ferré, F., Kranakis, E. and Krizanc, D. (2005) Structural rna has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578–591.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2001) *Introduction to Algorithms*. MIT Press and McGraw-Hill, Second edn..
- Coventry, A., Kleitman, D. J. and Berger, B. (2004) MSARi: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc Natl Acad Sci U S A*, **101**, 12102–12107.
- Cox, D. R. (1962) Further results on tests of separate families of hypotheses. . *J. Roy. Statistics. Soc. B*, **24**, 406–424.
- Crick, F. H. (1958) On protein synthesis. *Symp. Soc. Exp. Biol.*, **XII**, 139–163.
- Crick, F. H. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.
- Dalen, D. (2004) *Logic and Structure*. Kindle Edition, Springer, Fourth edn..
- Dalli, D., Wilm, A., Mainz, I. and Steger, G. (2006) Stral: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, **22**, 1593–9.
- Darwin, C. (1859) *The Origin of Species*. Murray edition, London.
- Dehmer, M., Emmert-Streib, F. and Gesell, T. (2008) A comparative analysis of multidimensional features of objects resembling sets of graphs. *Applied Mathematics and Computation*, **196**, 221 – 235.
- Deleuze, G. (1973) *Woran erkennt man den Strukturalismus?* Merve Verlag, Berlin.
- Deufjhard, D., Huisinga, W., Fischer, A. and Schütte, C. (2000) Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. *Linear Algebra Appl.*, **315**, 39.59.
- Do, C. B., Woods, D. A. and Batzoglou, S. (2006) Contrafold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–8.
- Dobzhansky, T. (1973) Nothing Make Sense Except in the Light of Evolution. *Am Biol Teach*, **35**, 125–129.
- Dowell, R. D. and Eddy, S. R. (2004) Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction. *BMC Bioinformatics*, **5**, 71.

- Duerr, H. P. (1985) *Dreamtime: Concerning the Boundary Between Wilderness and Civilization*. Basil Blackwell, NY.
- Duret, L. and Galtier, N. (2000) The covariation between tpa deficiency, cpg deficiency, and g+c content of human isochores is due to a mathematical artifact. *Mol Biol Evol*, **17**, 1620–1625.
- Dutheil, J., Pupko, T., Jean-Marie, A. and Galtier, N. (2005) A model-based approach for detecting coevolving positions in a molecule. *Mol. Biol. Evol.*, **22**, 1919–1928.
- Eddy, S. R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*, **22**, 2079–88.
- Eigen, M., Winkler-Oswatitsch, R. and Dress, A. (1988) Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 5913–5917.
- Faith, D. P. (1992) Conservation evaluation and phylogenetic diversity. *Biol. Conservat.*, **61**, 1–10.
- Farris, J. (1973) A probability model for inferring evolutionary trees. *Syst. Zool*, **22**, 250–256.
- Fedoroff, N. and Fontana, W. (2002) Genetic networks. Small numbers of big molecules. *Science*, **297**, 1129–1131.
- Felsenstein, J. (1978) Cases in Which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Zoology*, **27**, 401–410.
- Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
- Felsenstein, J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.
- Felsenstein, J. and Churchill, G. A. (1996) A hidden markov model approach to variation among sites in rate of evolution. *J. Mol. Evol.*, **13**, 92–104.
- Fitch, W. M. (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. . *Syst. Zool*, **20**, 406–416.
- Fitch, W. M. and Ayala, F. J. (1994) The superoxide dismutase molecular clock revisited. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 6802–6807.
- Flamm, C. and Hofacker, I. (2008) Beyond energy minimization: approaches to the kinetic folding of RNA. *Chemical Monthly*, **139**, 447–457.

- Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F. and Zehl, M. (2001) Design of multistable RNA molecules. *RNA*, **7**, 254–65.
- Fleißner, R. (2004) *Sequence alignment and phylogenetic inference*. Logos Verlag, Berlin, PhD Thesis.
- Fleißner, R., Metzler, D. and von Haeseler, A. (2005) Simultaneous statistical alignment and phylogeny reconstruction. *Syst Biol*, **54**, 548–561.
- Fontana, W. (2002) Modelling 'evo-devo' with RNA. *Bioessays*, **24**, 1164–1177.
- Fontana, W., Konings, D., Stadler, P. and Schuster, P. (1993a) Statistics of RNA secondary structures. *Biopolymers*, **33**, 1389–1404.
- Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Tarazona, P., Weinberger, E. D. and Schuster, P. (1993b) RNA folding landscapes and combinatorial landscapes. *Phys. Rev. E*, **47**, 2083–2099.
- Forsdyke, D. R. (2007) Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues. *J Theor Biol*, **248**, 745–753.
- Franklin, R. E. and Gosling, R. G. (1953) Molecular configuration in sodium thymonucleate. *Nature*, **171**, 740–741.
- Frege, G. (1879) *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Verlag von Louis Nebert, Halle.
- Frege, G. (1892) Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, **100**, 25–50.
- Freyhult, E., Gardner, P. and Moulton, V. (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, **6**, 241.
- Gardner, P. P. and Giegerich, R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
- Gascuel, O. (1997) Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Mol Biol Evol*, **14**, 685–695.
- Gesell, T. and von Haeseler, A. (2006) In silico sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics*, **22**, 716–722.
- Giegerich, R., Voss, B. and Rehmsmeier, M. (2004) Abstract shapes of RNA. *Nucleic Acids Res*, **32**, 4843–4851.
- Goldman, N. (1993) Simple diagnostic statistical tests of models for DNA substitution. *J.Mol.Evol.*, **37**, 650–661.

- Goldman, N., Thorne, J. L. and Jones, D. T. (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458.
- Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.
- Gorodkin, J., Staerfeldt, H. H., Lund, O. and Brunak, S. (1999) MatrixPlot: visualizing sequence constraints. *Bioinformatics*, **15**, 769–770.
- Gräf, S., Teune, J. H., Strothmann, D., Kurtz, S. and Steger, G. (2005) A computational approach to search for non-coding RNAs in large genomic data. In Hammann, C. and Nellen, W. (eds.), *Nucleic Acids and Molecular Biology, Vol. 17*, Springer-Verlag.
- Grassly, N., Adachi, J. and Rambaut, A. (1997) PSeq-Gen: An application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 559–560.
- Griffiths, P. (2008) In what sense does 'nothing make sense except in the light of evolution'? *Acta Biotheoretica*, *in press*.
- Griffiths-Jones, S. (2007) Annotating noncoding RNA genes. *Annu Rev Genomics Hum Genet*, **8**, 279–298.
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucl. Acids Res.*, **33**, D121–124.
- Gruber, A. R., Bernhart, S. H., Hofacker, I. L. and Washietl, S. (2008) Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, **9**, 122.
- Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**, 849–857.
- Gutell, R., Power, A., Hertz, G., Putz, E. and Stormo, G. (1992a) Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acid Research*, **20**, 5785–5795.
- Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J. and Stormo, G. D. (1992b) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.*, **20**, 5785–5795.
- von Haeseler, A. and Churchill, G. A. (1993) Network models for sequence evolution. *Journal of Molecular Evolution*, **37**, 77–85.
- von Haeseler, A. and Schöniger, M. (1998) Evolution of DNA or amino acid sequences with dependent sites. *J. Comput. Biol.*, **5**, 149–164.

- Hall, B. G. (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol. Biol. Evol.*, **22**, 792–802.
- Halpern, A. L. and Bruno, W. J. (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, **15**, 910–917.
- Hamada, M., Kiryu, H., Sato, K., Mituyama, T. and Asai, K. (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.
- Harding, E. F. (1971) The probabilities of rooted tree shapes generated by random bifurcation. *Advances in Applied Probability*, **3**, 44–77.
- Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.
- Hoagland, M., Stephenson, M. L., Scott, J. F., Hecht, L. I. and Zamecnik, P. (1958) A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem*, **231**, 241–57.
- Hoagland, M., Zamecnik, P., and Stephenson, M. L. (1957) Intermediate reactions in protein biosynthesis. *Biochim Biophys Acta*, **24**, 215–6.
- Hofacker, I., Fekete, M. and Stadler, P. (2002a) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
- Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M. and Schuster, P. (1994a) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
- Hofacker, I., Schuster, P. and Stadler, P. (1998) Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, **89**, 177–207.
- Hofacker, I. and Stadler, P. (2007) RNA secondary structures. In Lengauer, T. (ed.), *Bioinformatics: From Genomes to Therapies*, vol. 1, pages 439–489, Wiley-VCH, Weinheim, Germany.
- Hofacker, I. L. (1994) *A statistical characterization of the sequence to structure mapping in RNA*. University of Vienna, Institute für Theoretische Chemie, PhD Thesis.
- Hofacker, I. L. (2007) How microRNAs choose their targets. *Nat. Genet.*, **39**, 1191–1192.
- Hofacker, I. L., Bernhart, S. H. and Stadler, P. F. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.
- Hofacker, I. L., Fekete, M. and Stadler, P. F. (2002b) Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, **319**, 1059–1066.

- Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. and Schuster, P. (1994b) Fast folding and comparison of RNA secondary structures. *Monatsh Chem*, **125**, 167–188.
- Hogeweg, P. and Hesper, B. (1984) Energy directed folding of RNA sequences. *Nucleic Acids Res.*, **12**, 67–74.
- Holmes, I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, **6**, 73 [epub].
- Hudelot, C., Gowri-Shankar, H., Rattray, M. and Higgs, P. (2003) RNA-based phylogenetic methods: Application to mammalian mitochondrial RNA sequences. *Mol. Phylogenet. Evol.*
- Huelsenbeck (1995) Performance of phylogenetic methods in simulation. *System. Biol.*, **44**, 17–48.
- Huelsenbeck, J. P. and Bollback, J. P. (2001) Empirical and Hierarchical Bayesian Estimation of Ancestral States. *Systematic Biology*, **50**, 351–366.
- Huelsenbeck, J. P., Nielsen, R. and Bollback, J. P. (2003) Stochastic mapping of morphological characters. *Syst. Biol.*, **52**, 131–158.
- Hull Havgaard, J., Lyngsø, R., Stormo, G. and Gorodkin, J. (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.
- Hwang, D. G. and Green, P. (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 13994–14001.
- Innan, H. and Stephan, W. (2001) Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. *Genetics*, **159**, 389–399.
- Jensen, J. and Pedersen, A.-M. (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv Appl Prob*, **32**, 499–517.
- Ji, Y., Xu, X. and Stormo, G. D. (2004) A graph theoretical approach for prediction common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, **20**, 1591–1602.
- Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. In Munro, H. N. (ed.), *Mammalian Protein Metabolism*, vol. 3, pages 21–123, Academic Press, New York.
- Kaern, M., Elston, T. C., Blake, W. J. and Collins, J. J. (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, **6**, 451–464.

- Karlin, K. S. and Taylor, H. M. (1975) *A first course in stochastic processes*. Academic Press Inc., London, Second edn..
- Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., Kober, K. M., Miller, W., Pedersen, J. S., Pohl, A., Raney, B. J., Rhead, B., Rosenbloom, K. R., Smith, K. E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A. S., Haussler, D. and Kent, W. J. (2007) The UCSC genome browser database: 2008 update. *Nucleic Acids Res*, **35**.
- Kelchner, S. A. and Thomas, M. A. (2007) Model use in phylogenetics: nine key questions. *Trends Ecol. Evol. (Amst.)*, **22**, 87–94.
- Kemp, C. D. (1967) On a contagious distribution suggested for accident data. *Biometrics*, **23**, 241–255.
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. and Segal, E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
- Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.
- Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci.*, **78**, 454–458.
- Klingler, T. and Brutlag, D. (1993) Detection of correlations in tRNA sequences with structural implications. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1**, 225–233.
- Knudsen, B. and Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucl. Acids Res.*, **31**, 3423–3428.
- Knudsen, B. and Hein, J. J. (1999) Using stochastic context free grammars and molecular evolution to predict RNA secondary structure. *Bioinformatics*, **15**, 446–454.
- Koppel, M. (1994) Structure. In Herken, R. (ed.), *The Universal Turing Machine, A Half-Century Survey*, pages 403–419, Springer-Verlag, New York.
- Kosakovskiy, S. L., Frost, S. D. W. and Muse, S. (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21**, 676–679.
- Koshi, J. M. and Goldstein, R. A. (1995) Context dependent optimal substitution matrices. *Protein Eng.*, **8**, 641–645.
- Koshi, J. M. and Goldstein, R. A. (1997) Mutation matrices and physical-chemical properties: correlations and implications. *Proteins*, **27**, 336–344.

- Kruger, K., Grabowski, P., Zaug, A., Sands, J., Gottschling, D. and Cech, T. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell*, **31**, 147–157.
- Lapedes, A. S., Giraud, B. G., Liu, L. C. and Stormo, G. D. (1999) Correlated mutations in protein sequences: phylogenetic and structural effects. *Proceedings of the IMS/AMS Int. Conf. Stat Comp. Mol. Biol. Monograph Series of the Institute for Mathematical Statistics, Hayward, CA.*, **33**, 236–256.
- Lartillot, N. and Philippe, H. (2004) A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, **21**, 1095–1109.
- Léger, F. (2005) *Funzioni della pittura*. Milan, Italy.
- Leontis, N. B., Lescoute, A. and Westhof, E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16**, 279–287.
- Lescoute, A., Leontis, N. B., Massire, C. and Westhof, E. (2005) Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res.*, **33**, 2395–2409.
- Lindgreen, S., Gardner, P. P. and Krogh, A. (2006) Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics*, **22**, 2988–2995.
- Lockhart, P., Novis, P., Milligan, B., Riden, J., Rambaut, A. and Larkum, A. (2006) Heterotachy and tree building: A case study with plastids and eubacteria. *Mol. Biol. Evol.*, **23**, 40–45.
- Lopez, P., Casane, D. and Philippe, H. (2002) Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.*, **19**, 1–7.
- Lorenz, C., Pelchrzim, F. V. and Schroeder, R. (2006) Genomic systematic evolution of ligands by exponential enrichment (genomic selex) for the identification of protein-binding RNAs independent of their expression levels. *Nature Protocols*, **1**, 2204–12.
- Lück, R., Steger, G. and Riesner, D. (1996) Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV, and the mRNA of prion protein. *J. Mol. Biol.*, **258**, 813–826.
- Lunter, G. and Hein, J. (2004) A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, **20**, i216–i223.
- MacLane, S. (1971) *Categories for the Working Mathematician*. Springer-Verlag, Berlin.
- Maeda, J. (2006) *The Laws of Simplicity*. MIT Press, Cambridge, Massachusetts.

- Martin, L. C., Gloor, G. B., Dunn, S. D. and Wahl, L. M. (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.
- Mathews, D. H., Sabina, J., Zuker, M. and Turner, H. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
- Mathews, D. H. and Turner, D. H. (2002) Dynalign: An algorithm for finding secondary structures common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.
- Mathews, D. H. and Turner, D. H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol*, **16**, 270–8.
- McCaskill, J. (1990a) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
- McCaskill, J. S. (1990b) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
- Metzler, D. (2003) Statistical alignment based on fragment insertion and deletion models. *Bioinformatics*, **19**, 490–499.
- Meyer, I. M. and Miklos, I. (2007) SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.*, **3**, e149.
- Meyer, S. and von Haeseler, A. (2003) Identifying site-specific substitution rates. *Mol. Biol. Evol.*, **20**, 182–189.
- Meyers, L. A., Lee, J. F., Cowperthwaite, M. and Ellington, A. D. (2004) The robustness of naturally and artificially selected nucleic acid secondary structures. *J. Mol. Evol.*, **58**, 681–691.
- Miescher, J. F. (1969) Über die chemische Zusammensetzung der Eiterzelle. *Medizinisch-Chemische Untersuchungen*, **4**, 441–460.
- Miklós, I., Lunter, G. and Holmes, I. (2004) A "long indel" model for evolutionary sequence alignment. *Mol Biol Evol*, **21**, 529–540.
- Miklós, I. and Toroczka, Z. (2001) An improved model for statistical alignment. In Gascuel, O. and Moret, B. M. E. (eds.), *Algorithms in bioinformatics*, pages 1–10, Berlin, Springer.
- Minin, V. N. and Suchard, M. A. (2008) Counting labeled transitions in continuous-time Markov models of evolution. *J Math Biol*, **56**, 391–412.

- Missal, K., Rose, D. and Stadler, P. F. (2005) Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics*, **21 Suppl 2**, 77–78.
- Missal, K., Zhu, X., Rose, D., Deng, W., Skogerbo, G., Chen, R. and Stadler, P. F. (2006) Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J Exp Zool B Mol Dev Evol*, **306**, 379–392.
- Morton, B. R. (1995) Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 9717–9721.
- Moulton, V., Zuker, M., Steel, M., Pointon, R. and Penny, D. (2000) Metrics on RNA secondary structures. *J Comput Biol*, **7**, 277–292.
- Mourier, T., Carret, C., Kyes, S., Christodoulou, Z., Gardner, P. P., Jeffares, D. C., Pinches, R., Barrell, B., Berriman, M., Griffiths-Jones, S., Ivens, A., Newbold, C. and Pain, A. (2007) Genome-wide discovery and verification of novel structured RNAs in *plasmodium falciparum*. *Genome Res*.
- Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S., Stadler, P. and Hofacker, I. (2006) Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.
- Muse, S. V. (1995) Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics*, **139**, 1429–1439.
- Muse, S. V. and Gaut, B. S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, **11**, 715–724.
- Navidi, W. C., Churchill, G. A. and von Haeseler, A. (1991) Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Mol. Biol. Evol.*, **8**, 128–143.
- Nefes, J. M., van de Peer, Y., de Rijk, P., Chapelle, S. and de Wachter, R. (1993) Compilation of small ribosomal subunit rna structures. *Nuc. Acid Research*, **21**, 3025–3049.
- Nelson, D. and Cox, M. (2004) *Lehninger Principles of Biochemistry, Fourth Edition*. W. H. Freeman.
- Neyman, J. (1971) Molecular studies of evolution: A source of novel statistical problems. In S. S. Gupta, J. Y. (ed.), *Statistical Decision Theory and related Problems*, pages 1–27, S. S. Gupta, J. Yackel.
- Nicholas, J. S., Hoyle, D. C. and Higgs, P. G. (2000) RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum-likelihood methods. *Genetics*, **157**, 399–411.

- Nielsen, R. (2002) Mapping mutations on phylogenies. *Syst. Biol.*, **51**, 729–739.
- Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F. and O’Neal, C. (1965) RNA codewords and protein synthesis, VII. on the general nature of the RNA code. *Proc Natl Acad Sci*, **53**, 1161–8.
- Norris, J. R. (1997) *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
- Nussinov, R. and Jacobson, A. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, **77**, 6309–13.
- Nussinov, R., Pieczenik, G., Griggs, J. and Kleitman, D. (1978) Algorithms for loop matching. *SIAM J. Appl. Math.*, **35**, 68–82.
- Pagel, M. and Meade, A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, **53**, 571–581.
- Pang, A., Smith, A. D., Nuin, P. A. and Tillier, E. R. (2005) SIMPROT: using an empirically determined indel distribution in simulations of protein evolution. *BMC Bioinformatics*, **6**, 236.
- Paradis, E., Strimmer, K., Claude, J., Jobb, G., Opgen-Rhein, R., Dutheil, J., Noel, Y. and Bolker, B. (2004) *ape: Analysis of Phylogenetics and Evolution*. R package version 1.4.
- Parisi, G. (1994) Complexity in biology: The point of view of a physicist.
- Parisi, G. and Echave, J. (2001) Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.*, **18**, 750–756.
- Parisi, G. and Echave, J. (2005) Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. *Gene*, **345**, 45–53.
- Pedersen, A.-M. and Jensen, J. (2001) A dependent rates model and MCMC based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.*, **18**, 763–776.
- Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W. and Haussler, D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*, **2**.
- Pedersen, J. S., Meyer, I. M., Forsberg, R. and Hein, J. (2004a) An evolutionary model for protein-coding regions with conserved RNA structure. *Mol. Biol. Evol.*, **21**, 1913–1922.

- Pedersen, J. S., Meyer, I. M., Forsberg, R., Simmonds, P. and Hein, J. (2004b) A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucl. Acids Res.*, **32**, 4925–4936.
- Van de Peer, Y., Neefs, J. M., De Rijk, P. and De Wachter, R. (1993) Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. *J. Mol. Evol.*, **37**, 221–232.
- Philippe, H., Casane, D., Gribaldo, S., Lopez, P. and Meunier, J. (2003) Heterotachy and functional shift in protein evolution. *IUBMB Life*, **55**, 257–265.
- Pipas, J. and McMahon, J. (1975) Method for predicting RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 2017–2021.
- Pollock, D. D., Taylor, W. R. and Goldman, N. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol*, **287**, 187–98.
- Posada, D. and Buckley, T. (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.*, **53**, 793–808.
- R Development Core Team (2004) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- R. Durbin, S. Eddy, A. K. and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 2001st edn..
- Rambaut, A. and Grassly, N. C. (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.
- Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.*, **43**, 304–311.
- Rich, A. and Watson, J. D. (1954) Some relations between DNA and RNA. *Proc Natl Accad Sci*, **40**, 759–64.
- Rivas, E. and Eddy, S. R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
- Rivas, E. and Eddy, S. R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8–8.
- Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N. and Thorne, J. L. (2003) Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol.*, **20**, 1692–1704.

- Rodriguez, F., Oliver, J. L., Main, A. and Medina, J. R. (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.*, **142**, 485–501.
- Rose, D., Hackermueller, J., Washietl, S., Reiche, K., Hertel, J., Findeiss, S., Stadler, P. F. and Prohaska, S. J. (2007) Computational RNomics of drosophilids. *BMC Genomics*, **8**, 406.
- Rosenberg, M. (2005) MySSP: non-stationary evolutionary sequence simulation, including indels. *Evolutionary Bioinformatics Online*, **1**, 81–83.
- Ruan, J., Stormo, G. D. and Zhang, W. (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58–66.
- Russell, B. (1905) On denoting. *Mind*, **14**, 479–493.
- Rzhetsky, A. (1995) Estimating substitution rates in ribosomal RNA genes. *Genetics*, **141**, 771–783.
- Saitou, N. and Nei, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
- Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjolander, K., Underwood, R. C. and Haussler, D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, **22**, 5112–5120.
- Sandmann, T. and Cohen, S. M. (2007) Identification of novel drosophila melanogaster micrnas. *PLoS ONE*, **2**.
- Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
- Savill, N. J., Hoyle, D. C. and Higgs, P. G. (2000) RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum-likelihood methods. *Genetics*, **157**, 399–411.
- Schlegel, T. (2007) Inferring secondary structure from RNA alignments and their trees. *University Duesseldorf*, PhD Thesis.
- Schöniger, M. and von Haeseler, A. (1994) A stochastic model for the evolution of auto-correlated DNA sequences. *Mol. Phylogenet. Evol.*, **3**, 240–247.
- Schöniger, M. and von Haeseler, A. (1995a) Performance of the maximum likelihood, neighbor joining, and maximum parsimony methods when sequence site are not independent. *Syst. Biol.*, **4**, 533–547.
- Schöniger, M. and von Haeseler, A. (1995b) Simulating efficiently the evolution of DNA sequences. *Comput. Appl. Biosci.*, **11**, 111–115.

- Schöniger, M. and von Haeseler, A. (1999) Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. *J. Mol. Evol.*, **49**, 691–698.
- Schuster, P. and Stadler, P. F. (2007) Modeling conformational flexibility and evolution of structure: RNA as an example. In *Structural Approaches to Sequence Evolution*, Springer Berlin Heidelberg.
- Semple, C. and Steel, M. (2003) *Phylogenetics*. Oxford Lectures Series in Mathematics and its Applications, J. Ball and D. Welsh (eds.), Oxford University Press.
- Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423.
- Shapiro, B. (1988) An algorithm for comparing multiple RNA secondary structures. *CABIOS*, **4**, 387–393.
- Shapiro, B. and Zhang, K. (1990a) Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*, **6**, 309–318.
- Shapiro, B. A. and Zhang, K. Z. (1990b) Comparing multiple RNA secondary structures using tree comparisons. *Comput Appl Biosci*, **6**, 309–318.
- Siepel, A. and Haussler, D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, **21**, 468–488.
- Smit, S., Widmann, J. and Knight, R. (2007) Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Res.*, **35**, 3339–3354.
- Smith, A. D., Lui, T. W. H. and Tillier, E. R. M. (2004) Empirical models for substitution in ribosomal RNA. *Mol. Biol. Evol.*, **21**, 419–427.
- Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
- Steel, M. (2005) Should phylogenetic models be trying to "fit an elephant"? *Trends Genet.*, **21**, 307–309.
- Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J. and Giegerich, R. (2006) RNASHAPES: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
- Steger, G. (2003) *Bioinformatik*. Birkhäuser Verlag, Basel, Switzerland.
- Stephan, W. (1996) The rate of compensatory evolution. *Genetics*, **144**, 419–426.
- Stoye, J., Evers, D. and Meyer, F. (1998) Rose: generating sequence families. *Bioinformatics*, **14**, 157–163.

- Strope, C. L., Scott, S. D. and Moriyama, E. N. (2007) indel-Seq-Gen: a new protein family simulator incorporating domains, motifs, and indels. *Mol. Biol. Evol.*, **24**, 640–649.
- Sullivan, J. and Swofford, D. L. (1997) Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. *Evolution*, **4**, 77–86.
- Szymanski, M., Barciszewska, M. Z., Barciszewski, J. and Erdmann, V. A. (2000) 5S ribosomal RNA database Y2K. *Nucleic Acids Res.*, **28**, 166–167.
- Tabaska, J. E., Cary, R. B., Gabow, H. N. and Stormo, G. D. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, **14**, 691–699.
- Tafer, H., Ameres, S. L., Obernosterer, G., Gebeshuber, C. A., Schroeder, R., Martinez, J. and Hofacker, I. L. (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.
- Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**, 512–526.
- Tarkovsky, A. (1972) Solaris. *Visual Programme Systems*, **1972**.
- Tarkovsky, A. (1979) Stalker. *Visual Programme Systems*, **1979**.
- Tavaré, S. (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. *Lec. Math. Life Sci.*, **17**, 57–86.
- Theweleit, K. (1994) *Buch der Könige 2y. Recording Angel's Mysteries*. Stroemfeld.
- Thorne, J., Kishino, H. and Felsenstein, J. (1991) An evolutionary model for maximum likelihood alignment of dna sequences. *J Mol Evol*, **33**, 114–124.
- Thorne, J., Kishino, H. and Felsenstein, J. (1992) Inching toward reality: An improved likelihood model of sequence evolution. *J Mol Evol*, **34**, 3–16.
- Thorne, J. L., Goldman, N. and Jones, D. T. (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, **13**, 666–673.
- Tillier, E. (1994) Maximum likelihood with multi-parameter models of substitution. *J. Mol. Evol.*, **39**, 409–417.
- Tillier, E. and Collins, R. (1998) High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics.*, **148**, 1993–2002.
- Tillier, E. R. M. and Collins, R. A. (1995) Neighbor joining and maximum likelihood with RNA sequences: Addressing the interdependence of sites. *Mol. Biol. Evol.*, **12**, 7–15.

- Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M. and Gorodkin, J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res*, **16**, 885–9.
- Tuerk, C. and Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
- Tufféry, P. (2002) CS-PSeq-Gen: Simulating the evolution of protein sequence under constraints. *Bioinformatics*, **18**, 1015–1016.
- Turner, D. H. and Sugimoto, N. (1988) RNA structure prediction. *Annu Rev Biophys Biophys Chem*, **17**, 167–192.
- Uzilov, A. V., Keegan, J. M. and Mathews, D. H. (2006) Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, **7**, 173.
- Uzzell, T. and Corbin, K. W. (1971) Fitting Discrete Probability Distributions to Evolutionary Events. *Science*, **172**, 1089–1096.
- del Val, C., Rivas, E., Torres-Quesada, O., Toro, N. and Jiménez-Zurdo, J. I. (2007) Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *sinorhizobium meliloti* by comparative genomics. *Mol Microbiol*, **66**, 1080–1091.
- Vinh, L. S. and von Haeseler, A. (2004) IQPNNI: Moving fast through tree space and stopping in time. *Mol. Biol. Evol.*, **21**, 1565–1571.
- Voss, B., Giegerich, R. and Rehmsmeier, M. (2006) Complete probabilistic analysis of RNA shapes. *BMC Biol*, **4**, 5–5.
- Washietl, S. (2007) Prediction of structural noncoding RNAs with RNAz. *Methods Mol Biol*, **395**, 503–526.
- Washietl, S. *et al.* (2007a) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res*, **17**, 852–64.
- Washietl, S., Hofacker, I., Lukasser, M., Hüttenhofer, A. and Stadler, P. (2005a) Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. *Nature Biotech.*, **23**, 1383–1390.
- Washietl, S. and Hofacker, I. L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol*, **342**, 19–30.
- Washietl, S., Hofacker, I. L., Lukasser, M., Hüttenhofer, A. and Stadler, P. F. (2005b) Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. *Nat Biotechnol*, **23**, 1383–1390.

- Washietl, S., Hofacker, I. L. and Stadler, P. F. (2005c) Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, **102**, 2454–2459.
- Washietl, S., Pedersen, J. S., Korbelt, J. O., Stocsits, C., Gruber, A. R., Hackermüller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., Ucla, C., Wyss, C., Antonarakis, S. E., Denoeud, F., Lagarde, J., Drenkow, J., Kapranov, P., Gingeras, T. R., Guigó, R., Snyder, M., Gerstein, M. B., Reymond, A., Hofacker, I. L. and Stadler, P. F. (2007b) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res*, **17**, 852–864.
- Waterman, M. (1978) Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.*, **1**, 167 – 212.
- Waterman, M. and Smith, T. (1978) RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, **42**, 257–266.
- Watson, J. (1968) *he Double Helix: A Personal Account of the Discoverers of the Structure of DNA*. Atheneum, NewYork.
- Watson, J. D. and Crick, F. (1953) The structure of dna. *Cold Spring Harb Symb Quantl Biol*, **18**, 123–31.
- Weile, C., Gardner, P. P., Hedegaard, M. M. and Vinther, J. (2007) Use of tiling array data and RNA secondary structure predictions to identify noncoding rna genes. *BMC Genomics*, **8**, 244–244.
- Whelan, S., Lio, P. and Goldman, N. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.*, **17**, 262–272.
- Wiener, O. (1994) Form and content in thinking turing machines. In Herken, R. (ed.), *The Universal Turing Machine, A Half-Century Survey*, pages 583–607, Springer-Verlag, New York.
- Wiener, O. (1998) *Eine elementare Einführung in die Theorie der Turing-Maschinen*. Springer, Wien , New York.
- Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. and Backofen, R. (2007) Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, **3**, e65.
- Wilm, A., Linnenbrink, K. and Steger, G. (2008) ConStruct: Improved construction of RNA consensus structures. *BMC Bioinformatics*, **9**, 219.
- Workman, C. and Krogh, A. (1999) No evidence that mrnas have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, **27**, 4816–4822.

- Xia, T., SantaLucia, J., Burkard, M., Kierzek, R., Schroeder, S., Jiao, X., Cox, C. and Turner, D. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with watson-crick base pairs. *Biochemistry*, **37**, 14719–35.
- Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *J. Mol. Evol.*, **42**, 587–596.
- Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximative methods. *J. Mol. Evol.*, **39**, 306–314.
- Yang, Z. (1996) Maximum-likelihood models for combined analyses of multiple sequences data. *J. Mol. Evol.*, **42**, 587–596.
- Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences*, **13**, 555–556.
- Yang, Z. (2000) Complexity of the Simplest Phylogenetic Estimation Problem. *Proc. R. Soc. London*, **267**, 109–119.
- Yang, Z., Goldman, N. and Friday, A. (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.*, **11**, 316–324.
- Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.
- Yao, Z., Weinberg, Z. and Ruzzo, W. L. (2006) CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
- Yu, J. and Thorne, J. L. (2006) Dependence among sites in RNA evolution. *Mol. Biol. Evol.*, **23**, 1525–1537.
- Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406–15.
- Zuker, M. and Sankoff, D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol*, **46**, 591–621.
- Zuker, M. and Stiegler, P. (1981a) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
- Zuker, M. and Stiegler, P. (1981b) Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, **9**, 133–148.

List of Chapter Figures & Quotes

- Self-portrait, T. Gesell, Photo, Kunstakademie Münster, 1995, all rights reserved by the artist.
- Theweleit (1994), Buch der Könige, Band 2y, Recording Angel's Mysteries, Stroemfeld Roter Stern, 1994, loose translation phrase, page 438,
- when horses die . . . max.E CD16
- “Man Darf Auch Weben Was Man Nicht Sieht” (You Can Also Weave What You Do Not See) The Tapestries Of Dieter Roth And Ingrid Wiener, 1981-1984, whole, 220 x 182 cm, collection Franz Wassmer, Ennetbaden, © Dieter Roth Estate, Ingrid Wiener und Valie Export.
- Max, Thomas Brinkmann, Photo, Cologne, 2006, all rights reserved by the artist.
- Antoin Artaud: Retrieved July 2, 2009, from:
http://www.leninimports.com/antonin_artaud_gallery_21.jpg
- Mata Hari Retrieved July 2, 2009, from:
http://commons.wikimedia.org/wiki/File:Mata-Hari_Paris_1910.jpg
- Kurt Gödel, Retrieved 2007, from:
http://www.univie.ac.at/bvi/photo-gallery/photo_gallery.htm,
BVI, University of Vienna; thereon: collage with SISSI , including a letter alphabet and the observed counted frequencies of t's diary note (07.01.2007) along a golden touch, detail, screen print on paper, all rights for the collage reserved by the artist.
- Rosalind Franklin, Retrieved July 2, 2009, from :
http://vietsciences.free.fr/biographie/biologists/images/Franklin_Rosalind-lab.jpg
- Solaris. Dir. Andrei Tarkovsky. Perfs. Donatas Banionis, Natalya Bondarchuk, Jüri Järvet, Vladislav Dvorzhetsky, Nikolai Grinko, Anatoly Solonitsyn. Visual Programme Systems, 1972.

- Black Square (Kazimir Malevich, 1915, Oil on Canvas, State Russian Museum, St.Petersburg), Retrieved July 22, 2009, from :
<http://de.wikipedia.org/w/index.php?title=Datei:Malewitsch.jpg&filetimestamp=20060801090524>
- The light of Curt Fischer (1920),
Andrea Ulrich and Sebastian Jacobi provided the graphic of the light.
- Medusa, Sabine Boehl, detail, 2003-2004, 84.5 x 114 cm, glass beads on canvas, all rights reserved by the artist, courtesy Galerie Nächst St. Stephan Rosemarie Schwarzwälder
- Romy Schneider, Retrieved January 2, 2008, from:
http://www.divasthesite.com/Acting_Divas/Romy_Schneider.htm
- At first, stimulated by Prof. Oswald Wiener in seminars at the Art School in Dueseldorf, I became interested in a general discussion about structure, e.g. in the context of Turing Machines (Wiener, 1998). Furthermore, I grasped the opportunity of attending the CSSS08 (Complex System Summer School 2008, Santa Fe, US) to share and discuss my understanding in a short self-organized project involving ten participants, a so-called *Structure.Passion* project:
http://www.santafe.edu/events/workshops/index.php/Tanja_Gesell
http://www.santafe.edu/events/workshops/index.php/The_structure_definitions_wiki_page

I have tried to my best knowledge not to infringe copyright protection. However, please contact me in case of any infringement ⁴.

⁴Ich habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zu Verwendung der Bilder in dieser Arbeit eingeholt. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.

Curriculum Vitae

0647738

Personal Data

Name Tanja M. Gesell
Date of Birth February 25, 1973
Nationality German

Education

2006/01 - **Graduate studies in Molecular Biology**, University of Vienna, Austria; Continued PhD thesis with Arndt von Haeseler.
A Phylogenetic Definition of Structure.

2004/08 - 2005/12 **Graduate studies in Computer Science**, University of Duesseldorf; PhD thesis with Arndt von Haeseler.

Undergraduate studies in Biology
with Minors: Bioinformatics and Computer Science, University of Duesseldorf and Muenster, Germany.
Degree Diploma: December 18, 2003.

1993/04 - 2004/02 **Undergraduate studies in Fine Art**
with Minors: Philosophy (Aesthetics) and History of Art at the Kunstakademie Duesseldorf and Muenster, Germany.
Degree Diploma: February 18, 2004.

1983/06 - 1992/06 **High school**, Duesseldorf, Germany
Degree Matura: June 12, 1992.

Grants & Awards

2008/06 Fellow Participant of the Complex System Summer School 2008, Santa Fe Institute, US.

2004/03 - 2004/07 EU Marie Curie Fellow at the European Bioinformatics Institute, Goldman Group, Hinxton, Cambridge, UK.

2003 Meisterschülerin of Prof. R. Trockel, Kunstakademie Duesseldorf, Germany

Acknowledgments

Some intellectuals, those writers and scientists. They don't believe in anything! They've lost their sense of hope! My God! What kind of people are they? ... Science? Nonsense! In this situation, mediocrity and genius are equally helpless. We don't want to conquer space at all. We want to expand Earth endlessly. We don't want other worlds; we want a mirror. We seek contact and will never achieve it. We are in the foolish position of a man striving for a goal he fears and doesn't want. Man needs man! (Tarkovsky, 1972, 1979).

THANKS TO ALL!

Thesis Mixture Acknowledgment [®]

PROPERTIES & EFFECTS powerful psychotropic agents		SIDE EFFECTS the following undesirable effects may occur	
● happiness	a state of well being	● sadness	a state of feeling unhappy
● inspiring	energy, drive	● imbalance	disproportion, inequality, asymmetry
● clear thinking	goals	● confusion	state of disorder, lack of clearness
● realistic thinking	rational accomplishments	● dreamy	visionary, vague
● realistic point of view	motivation	● unrealistic	not compatible with reality or fact
● passion	beyond words	● ocean	happiness
● artistic	sublime	● insomnia	sleeplessness
● biblical	fictional	● anxiety	a state of apprehension and psychic tension
● inhalation	aspiration	● tachycardia	rapid heart rate
● Stalker	guide	● HAL 9000	heuristic
● cleverness	smart	● humdrum	follow the routines
● effective	fast worker	● authority dependence	stop thinking
● copycat	using each possible information	● helpful	transfer of knowledge
● merciful	caring	● fairylike	...
● craziness	...	● untouchable	...
● resilience	power of resistance	● you have to swallow	after reconstitution
● angelic	healing power	● compulsive reader	...
● hero	everday hero	● anti-pode	Octopus
● Oedipus	oedipal	● patience	...
● honesty	integrity	● egocentricity	self-centredness
● enthusiasm	progress	● greed	Gollum
● humour	fun	● selfishness	retrotransposon
● sweetness	e.g $C_6H_{12}O_6$	● bitterness	unpleasant taste
● commitment	emotional confinement	● flakiness	Iago
● neutral	...	● thoughtlessness	lacking in consideration for others
● neutral	...	● unknown	..

Table 1: Instructions: Read carefully before use – Pack: 50’s tablets – Storage: Protect from light. Store in a dry place below 35 C – NOTE: Shake both oral suspension and pediatric drops well before using – Keep out of reach of children and animals – some effects and side effects should not occur in all people .

Did I forget something...?



SISSI



Wiener Wissenschafts-, Forschungs- und Technologiefonds



