Inferring Secondary Structure from RNA Alignments and their Trees

Inaugural-Dissertation

zur

Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Thomas Schlegel

aus Halle/Saale

Düsseldorf

2007

Aus dem Institut für Informatik
der Heinrich-Heine Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Arndt von Haeseler
Koreferent: Prof. Dr. Martin Lercher

Tag der mündlichen Prüfung: 22. Juni 2007

Danksagung

Vor allem danke ich meinem Betreuer Arndt von Haeseler für das Thema, interessante Diskussionen und die angenehme Arbeitsatmosphäre. Ich danke meinen Kollegen Tanja, Lutz, Stefan Z., Nicole, Jochen, Ingo P., Thomas L. und Michael für die Zusammenarbeit und Unterstützung. Martin Lercher danke dafür, dass er sich bereiterklärt hat, meine Arbeit zu begutachten. Gerhard Steger danke ich für die freundliche Bereitstellung des Riboswitch Alignments. Der Düsseldorf Entrepreneur Foundation danke ich für die finanzielle Unterstützung.

Nach der Pflicht die Kür:

Vielen Dank an die besten Freunde: Christian, Katja und Angela für Eure liebenswerten Eigenarten . . . die letzten elf Jahre lang . . . . . soviel Dank kann man gar nicht niederschreiben. Meinen lieben Eltern danke ich für einfach alles, genauso meinem Schwesterherz Kathrin.

Mein besonderer Dank gilt:

- Arndt, Uli und Jule – bei Euch fühlt man sich wie zu Hause und natürlich für den Rumtopf.
- Tobi, dem unerschöpflichen Quell an Zigaretten, für unterhaltsame Kaffeepausen und dem Versuch mir Fussball nahe zu bringen.
- Gunter und Judith für Paula, Wein, Zigaretten, Einblicke in Statistik sowie Soziologie und vielem mehr.
- Jochen, Roland, Nicole und Markus die mehr sind als nur Arbeitskollegen.
- Claudia und Anja – Mädels, bleibt so wie Ihr seid.

Weiterhin danke ich Enrico, Oliver, Lilian, Stefan K., Heike A. und Kerstin.

Contents

Introduction

After enunciating the central dogma of molecular biology in 1958 (Crick, 1958), the RNA was considered to be only an intermediate step that carries the information from DNA, that stores all genetic information, to proteins that catalyze the biochemical reactions within the cell. Over the years, it was recognized that RNA is essential in many biological processes (Meli *et al.*, 2001; Mattick and Makunin, 2006), where the function of the molecule is to a large degree determined by its structure.

Moreover, RNA plays an important role in phylogenetic analysis. Especially, the SSU rRNA is widely used for tree reconstruction, since it is available for many sequences, "sufficiently" long and it contains enough evolutionary information (Higgs, 2000). For the reconstruction of phylogenetic trees most methods assume that each site in a sequence evolves independently of each other. However, these approaches ignore that these molecules have complex three dimensional structures. To obtain a "good" phylogeny, evolutionary models have to incorporate such constraints.

The aim of structure prediction methods is to find these constraints from a sequence or a set of sequences. This is a quite challenging task since for a given sequence there are many possible structures. The number of possible secondary structures $S(l)$ of a RNA molecule with sequence length l can be

approximated by Waterman (1995):

$$S(l) \sim \sqrt{\frac{15+7\sqrt{5}}{8\pi}} l^{-3/2} \left(\frac{3+\sqrt{5}}{2}\right)^{l} \tag{1}$$

Beside experimental methods, there exists a broad variety of computational methods for structure prediction. Computational methods can be categorized in thermodynamic and comparative methods. Thermodynamic methods predict the secondary structure given a single nucleotide sequence, whereas comparative methods determine a consensus structure based on a set of aligned sequences (cf Zuker, 2000).

This thesis deals with the statistical inference of dependencies within a collection of biological sequences. These sequences may be either DNA, protein or RNA sequences. We will focus on RNA molecules. Dependencies of a RNA sequence are for example the secondary or tertiary structure.

A special focus of this work is the influence of the phylogeny in detecting dependencies. In chapter 1 we give a brief overview of RNA sequences, their structure and discuss models of sequence evolution. Then, we discuss the principles of thermodynamic and comparative structure prediction methods. Based on simulations, we investigate in chapter 2 how the phylogenetic relationship contributes to the ability in predicting the structure of RNA. Furthermore, we introduce two novel comparative methods for structure prediction in chapter 2 and 3. Finally, we apply these methods to synthetic data, sequences of tRNA and sequences containing a purine riboswitch and compare the results.

Chapter 1

Theoretical Background

This thesis deals with the development of tools to determine dependencies (a definition of dependencies is given in section 1.1.1) from related RNA sequences. RNA is a nucleic acid consisting of nucleotides. Nucleotides consists of three components: a base, a ribose sugar and a phosphate group. The bases of the RNA are adenine, guanine, cytosine and uracil, adenine and guanine being purines and cytosine and uracil being pyrimidines. For the purpose of this thesis we consider RNA molecules as strings from a four letter alphabet $\mathcal{A}$, where nucleotides are abbreviated by the first letter of their corresponding base, thus $\mathcal{A} = \{A, C, G, U\}$.

In this chapter, we will discuss the biological and mathematical requisites that are needed in chapter 2 and 3. We consider two aspects: the evolution of sequences and their structural elements. The evolution of sequences can be modeled by a Markov process as introduced in section 1.1.2. Then we will discuss structural elements in more detail. To extract structural information from RNA sequences we use statistical tests. The basics of such tests as well as classical structure prediction methods are reported in section 1.2. Finally, some problems relating structure prediction methods are discussed.

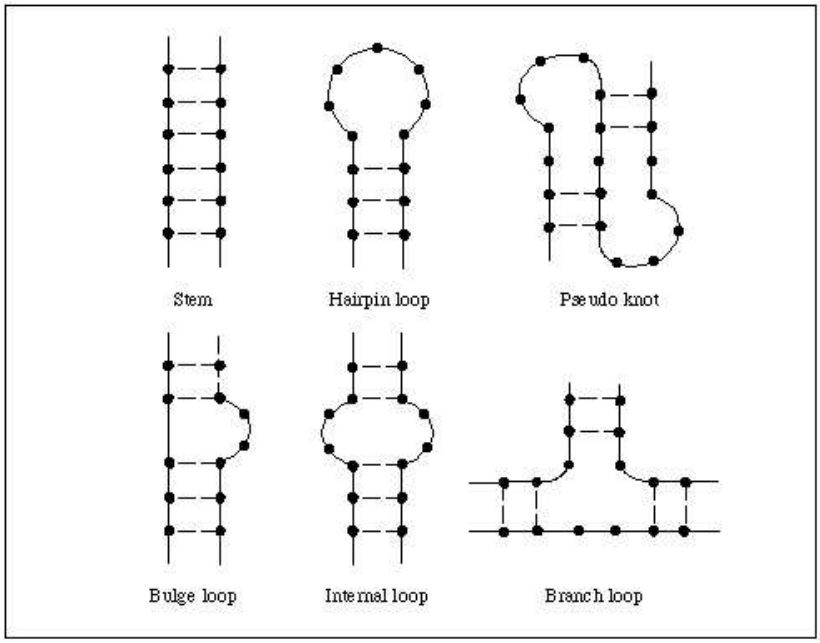

Figure 1.1: Different structural elements of RNA
Circles represent nucleotides and dashed lines represent base pairs (picture taken from *www.sacs.ucsf.edu/Training/rnastruc/RNA.gif*).

1.1 Biological Data and Molecular Evolution

1.1.1 RNA secondary and tertiary structure

The representation of RNA molecules as a linear sequence $\mathbf{a} = a_1, a_2, \ldots, a_l$ is denoted as primary structure. However, these molecules have in general a complex three dimensional structure. In the case of RNA, the basis of such structures is the ability of nucleotides to form hydrogen bonds to non neighboring bases to form base pairs. These base pairs occur between $A - U$ and $C - G$, also called Watson-Crick pairs and the wobble pair $G - U$.

The structural elements of the RNA can be distinguished in stems and loops. Stems are consecutive base pairs. They form a double helix as known from DNA. Loops are unpaired regions within RNA. Different combinations of loops and stems are summarized in Figure 1.1.

The secondary structure of a RNA sequence can be visualized as planar graph that satisfies the following condition: If a_j pairs with $a_{j'}$ and a_k is paired with $a_{k'}$ with $j < k < j'$, then $j < k' < j'$ (WATERMAN, 1995). As an example Figure 1.2A shows the secondary structure of a tRNA molecule. Note, that due to this definition of the secondary structure the pseudo knot shown in Figure 1.1 is not a secondary structural element.

The secondary structure, however, gives no information on the relative position of each nucleotide in three dimension. This can be exemplified by the tRNA shown in Figure 1.2. The secondary structure displays a clover leaf structure whereas the 3D representation, the so called tertiary structure, constitutes an L-shaped molecule (Figure 1.2C).

For a general description of dependencies within a RNA molecule containing l nucleotides, the definition of neighborhood systems $\mathcal{N} = (N_j)_{j=1,2,\ldots,l}$ is used. Each N_j contains the positions that interact with position j. It fulfills the following conditions (BREMAUD, 1999)

- $j \notin N_j$
- $j' \in N_j \Rightarrow j \in N_{j'}$.

In this thesis, we call two positions "correlated" or "dependent" when they are neighbors. A special case of dependencies is the secondary structure of RNA molecules. For illustration, consider the secondary structure of the tRNA molecule in Figure 1.2A. We can define two sites as dependent if they are base paired. For example, position 1 and position 72 are dependent, since $N_1 = \{72\}$ and $N_{72} = \{1\}$. Position 16 is located in a loop and has no neighbor, i.e. $N_{16} = \emptyset$.

A convenient method to display neighborhood systems are circle plots. A circle plot displaying the corresponding secondary structure of the tRNA

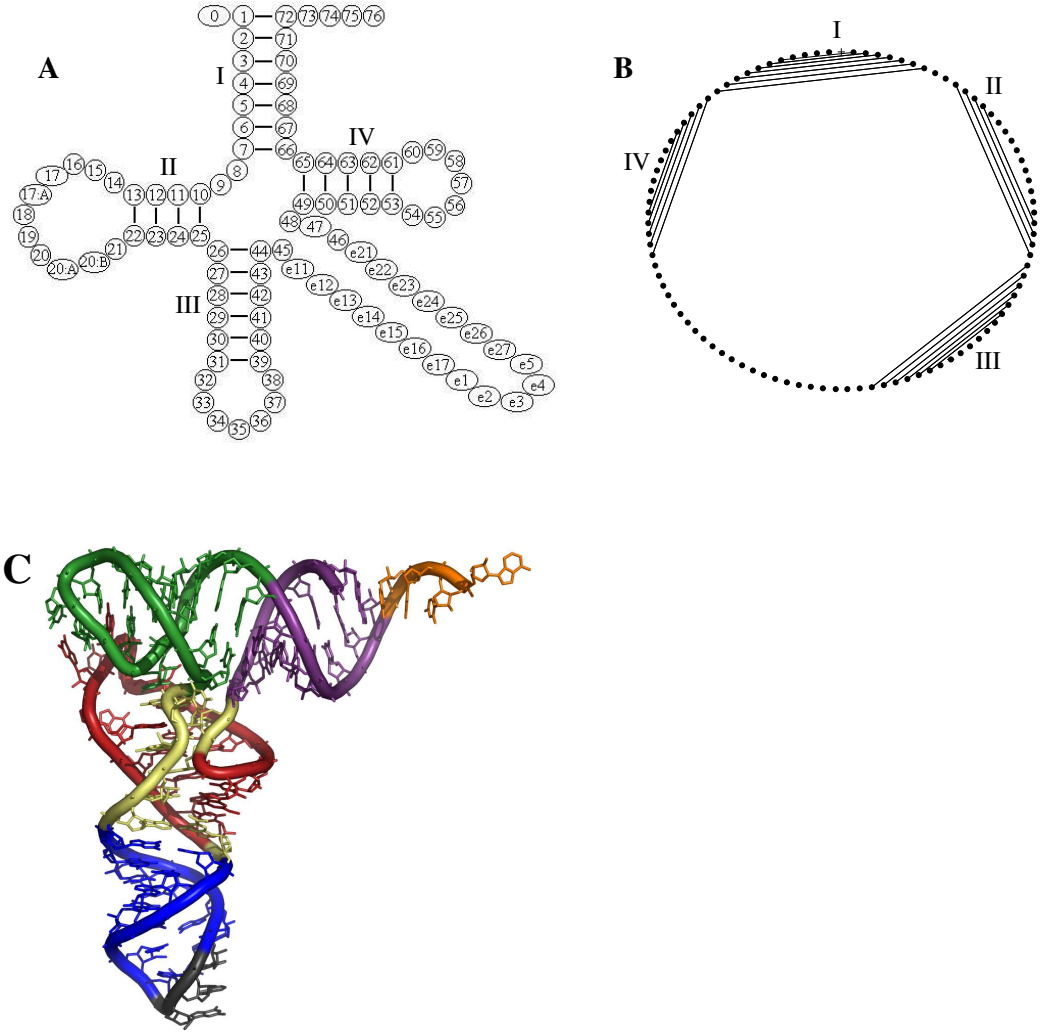

Figure 1.2: Three representations of a tRNA molecule containing four stem regions. **A**: Cloverleaf structure (secondary structure). **B**: A circle plot is another representation of the secondary structure. Circles represent nucleotides. Nucleotides connected by an edge are base pairs (Picture taken from *http://www.staff.uni-bayreuth.de/ btc914/search/index.html*). **C**: The 3d structure (tertiary structure).

molecule of Figure 1.2A is shown in Figure 1.2B. Each node represents a position in the molecule and each edge links two neighbors.

1.1.2 Sequence Alignment and Sequence Evolution

Alignments

To analyze a set of sequences we have to know which positions of the sequences are homologous. Sequences are related or homologous if they share one common ancestor. We will display the homology between bases of different sequences in form of a sequence alignment $\mathbf{D}$. An alignment is a data matrix where each row corresponds to a sequence and homologous nucleotides are written in a column. Since sequences are in general not of the same length the gap character "-" is introduced to account for inserted or deleted nucleotides. Thus, the alignment $\mathbf{D}$ is a $n \times l$ matrix with n sequences of length l. The entries D_{ij} denote the nucleotide at site j of sequence i. The column of an alignment is also called alignment site.

The nucleotides within an alignment site can differ. These differences can be explained by substitutions[1], i.e. a nucleotide is substituted by another one. Substitutions can be distinguished in transitions and transversions. Transitions are substitutions from a purine to a purine or from a pyrimidine to a pyrimidine. Transversions are substitutions from a purine to a pyrimidine or vice versa. Substitutions can occur due to replication errors of the DNA, as well as by mutagens like certain chemicals or UV light.

Sequence alignments are the basis of many molecular analysis. The final goal of inferring a "good" alignment from a collection of nowadays sequences is very challenging because these sequences differ in the nucleotide compo-

[1] A substitution is formally defined as a point mutation that is fixed in a population. In this thesis we will use "substitution" and "point mutation" exchangeable

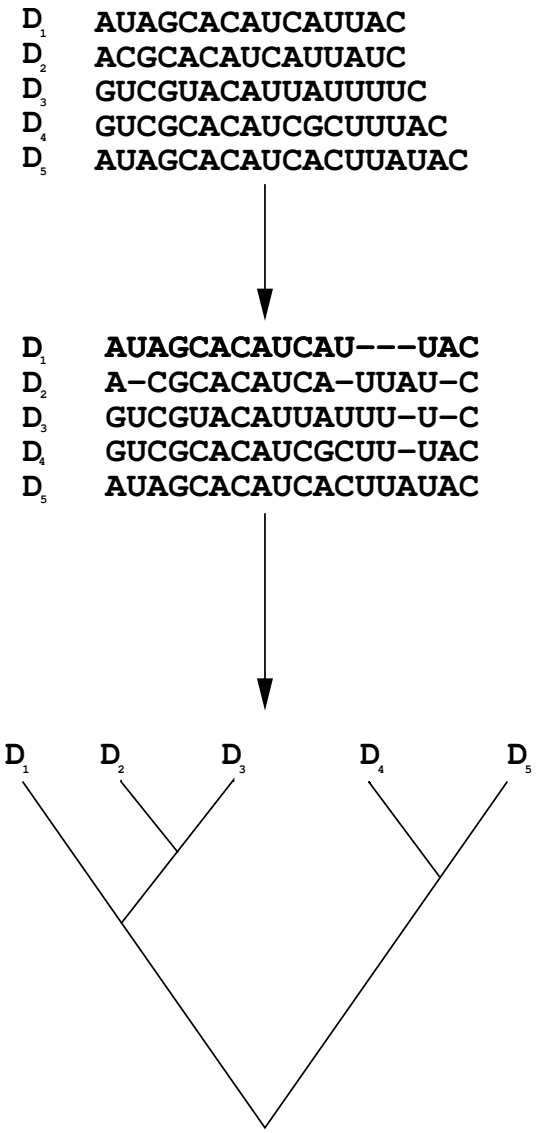

Figure 1.3: Top: RNA sequences from different organisms. Center: The sequence alignment displays the homology relation of nucleotides. Bottom: Reconstructed phylogenetic tree based on the sequence alignment

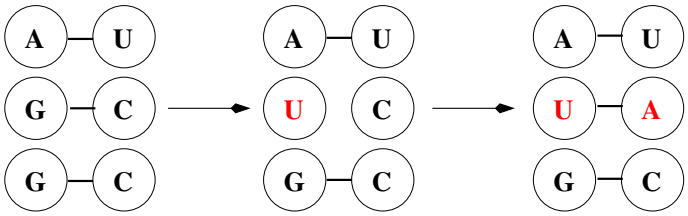

Figure 1.4: Example of a compensatory substitution.
After G is substituted a mispair is introduced. This is compensated by a substitution from C to A.

sition as well as sequence length (Figure 1.3). The different alignment algorithms will not be discussed here. For a summary see Wallace *et al.* (2005); Notredame (2002). In this thesis we assume the alignments as given.

Models of Sequence Evolution

If we consider tRNA sequences from different organisms we observe that the cloverleaf structure is to a high degree conserved (slight deviations exist, e.g. an additional base pair exists in the stem or a loop is missing (Steinberg and Cedergren, 1995)), although the nucleotide sequences differ.

In order to keep the structure, especially the base pairs in the stem, we have to model the evolution of dinucleotides. In more detail: if a nucleotide at a site j within a stem region is substituted then, the base paired site j' has to be substituted as well (cf Chen *et al.*, 1999). This substitution is called a compensatory substitutions. The mechanism is shown in Figure 1.4. Displayed is a part of a stem region. If G is substituted by an U, then a mispair is introduced and the stem is destabilized. To compensate this, there are two possibilities: First, the neighboring C is substituted to A to constitute a base pair, or second, a back mutation from U to G occurs.

To model compensatory substitutions we have to consider the evolution of dinucleotides. For clarity, single nucleotide substitution models are explained

first. Afterwards, these models can easily be extended to dinucleotide substitution models.

For a single nucleotide substitution model, we assume that a substitution at a position within a sequence occurs randomly and independently from any other position. Moreover, we assume that the nucleotide frequencies $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ do not change over time. Under these assumption a time-homogeneous stationary Markov process can model the substitution process (Tavaré, 1986). Each position in the sequence is then described by a discrete random variable. At the RNA level there are four possible states corresponding to the nucleotides A, C, G and U. The substitution from one nucleotide to another is then described by a four times four probability matrix $\mathbf{P}(t)$. The components $P_{jj'}(t)$ specify the probability of a substitution from nucleotide j to j' after a period of time $t > 0$.

The probability matrix is characterized by a rate matrix $\mathbf{Q}$ and is computed as:

$$\mathbf{P}(t) = \exp(\mathbf{Q}t). \tag{1.1}$$

Thus, it suffices to describe the substitution process by the rate matrix

$$\mathbf{Q} := Q_{j,j'} = \begin{cases} r_{jj'}\pi'_j & \text{if } j \neq j' \\ -\sum_{j \neq j'} Q_{jj'} & \text{if } j = j'. \end{cases} \tag{1.2}$$

with $j, j' \in \mathcal{A}$. $\mathbf{Q}$ provides an infinitesimal description of the substitution process. An entry $Q_{jj'}$ is the number of substitutions from nucleotide j to j' per unit time. The $r_{jj'} > 0$ are rate parameters, that account for transitions and transversions. Finally, parameters $\pi_A, \pi_C, \pi_G, \pi_T$ describe the frequencies of nucleotides A, C, G and T, respectively.

A collection of different rate matrices is given in Table 1.1. The most simple matrix is that of Jukes and Cantor (Jukes and Cantor, 1969) containing one parameter, i.e. each substitution occurs with the same rate α. A

	A	C	G	U	A	C	G	U
	JC69				K2P			
A	-	α	α	α	-	β	α	β
C	α	-	α	α	β	-	β	α
G	α	α	-	α	α	β	-	β
U	α	α	α	-	β	α	β	-
	HKY				TN93			
A	-	$\beta\pi_C$	$\alpha\pi_G$	$\beta\pi_U$	-	$\beta\pi_C$	$\alpha_1\pi_G$	$\beta\pi_U$
C	$\beta\pi_A$	-	$\beta\pi_G$	$\alpha\pi_U$	$\beta\pi_A$	-	$\beta\pi_G$	$\alpha_2\pi_U$
G	$\alpha\pi_A$	$\beta\pi_C$	-	$\beta\pi_U$	$\alpha_1\pi_A$	$\beta\pi_C$	-	$\beta\pi_U$
U	$\beta\pi_A$	$\alpha\pi_C$	$\beta\pi_G$	-	$\beta\pi_A$	$\alpha_2\pi_C$	$\beta\pi_G$	-
	F81				GTR			
A	-	π_C	π_G	π_U	-	$a\pi_C$	$b\pi_G$	$c\pi_U$
C	π_A	-	π_G	π_U	$a\pi_A$	-	$d\pi_G$	$e\pi_U$
G	π_A	π_C	-	π_U	$b\pi_A$	$d\pi_C$	-	$f\pi_U$
U	π_A	π_C	π_G	-	$c\pi_A$	$e\pi_C$	$f\pi_G$	-

Table 1.1: Rate matrices for different substitution models, JC69: Jukes-Cantor model (Jukes and Cantor, 1969), K2P: Kimura two parameter model (Kimura, 1980), HKY: Hasegawa-Kishino-Yano model (Hasegawa *et al.*, 1985), TN: Tamura-Nei model (Tamura and Nei, 1993), GTR: general time reversible model (Rodriguez *et al.*, 1990). The entries of the main diagonal equals the negative sum of the entries of the corresponding row.

more general model is the K2P-model of Kimura (KIMURA, 1980). It distinguishes between transitions and transversions. However, both models assume that each of the four nucleotides within the sequences is equally distributed with probability 0.25. More general single nucleotide substitution models (HASEGAWA *et al.*, 1985; TAMURA and NEI, 1993; RODRIGUEZ *et al.*, 1990) incorporate different base compositions. The parameters of each substitution model are estimated from the data.

A further assumption is that the substitution process is reversible; that is,

$$\pi_j P_{jj'}(t) = \pi_{j'} P_{j'j}(t). \tag{1.3}$$

This additional assumption implies that the substitution process has no preferred direction. From the reversibility assumption it follows that a stationary distribution $\boldsymbol{\pi}^S$ exists, where:

$$\boldsymbol{\pi}^S = \boldsymbol{\pi}^S \mathbf{P}(t). \tag{1.4}$$

This means that any initial nucleotide distribution $\boldsymbol{\pi}^i$ converges to the stationary distribution as $t \to \infty$ that is,

$$\boldsymbol{\pi}^i \mathbf{P}(t) \stackrel{t\to\infty}{\longrightarrow} \boldsymbol{\pi}^S, \tag{1.5}$$

where time t is measured in numbers of substitutions per unit time. Therefore the entries of the rate matrix $\mathbf{Q}$ have to be rescaled that the expected number of substitutions per unit time equals one, i.e. $-\sum_{i\in\mathcal{A}} Q_{ii}\pi_i^s = 1$ (STRIMMER and VON HAESELER, 2003).

As yet, we considered the case of independently evolving nucleotides that are represented by a four by four rate matrix $\mathbf{Q}$. The assumption of independently evolving sites is obviously violated in the stem regions of RNA sequences, due to compensatory substitution. To model compensatory substitutions we have to describe substitutions between dinucleotides.

The substitution model is then expressed by a Markov process characterized by a 16×16 rate matrix where the number of possible states are the nucleotide words of length two, that is $\mathcal{A} \times \mathcal{A} = \{AA, AC, \ldots, UU\}$. Thus, these models (SCHÖNIGER and VON HAESELER, 1994; TILLIER, 1994; TILLIER and COLLINS, 1998; MUSE, 1995; RZHETSKY, 1995; SAVILL *et al.*, 2001) describe the substitution of independently evolving dinucleotides and thus give generally a more realistic description of the sequence evolution. An example for a dinucleotide substitution model is the SH-model (SCHÖNIGER and VON HAESELER, 1994):

$$Q_{j,j'} = \begin{cases} \pi_{j'} & \text{if } \mathcal{H}(j,j') = 1 \\ 0 & \text{if } \mathcal{H}(j,j') = 2 \\ -\sum_{j \neq j'} Q_{jj'} & \text{if } j = j'. \end{cases} \tag{1.6}$$

with $j, j' \in \mathcal{A}^2$ and the Hamming distance $\mathcal{H}(j,j')$. That is, for this model a substitution occurs from one dinucleotide to another dinucleotide when they differ by one nucleotide.

More complex models can be obtained while extending the state space to $\mathcal{A}^k$. This corresponds to independently evolving sequence fragments of length k. A summary of different substitution models up to $k = 3$ is given in SIEPEL and HAUSSLER (2004); for a general description of the Markov process for any k see VON HAESELER and SCHONIGER (1998). Recently different substitution models were introduced that relax the assumption of independently evolving sequence fragments (e.g. JENSEN and PEDERSEN, 2000; GESELL and VON HAESELER, 2006; SIEPEL and HAUSSLER, 2004). These models account for context dependent substitutions, where a nucleotide is substituted depending on the nucleotides at other positions of the sequence.

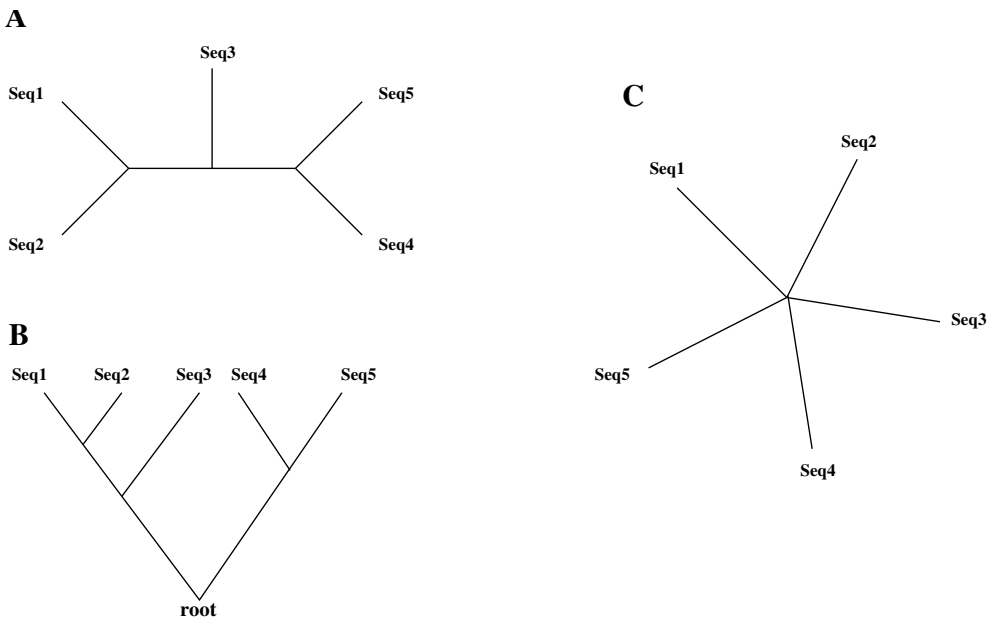

Figure 1.5: Phylogenetic trees of five sequences. **A**: unrooted tree, **B**: rooted tree, **C**: star tree.

Phylogenetic Trees

Sequence alignments are the basis to reconstruct phylogenetic trees (Figure 1.3). Phylogenetic trees are used to represent the evolutionary relationship among species. A tree is formally defined as a graph $\mathcal{T} = \mathcal{G}(E, V)$ with no cycles, where V is the set of vertices and E the set of edges connecting vertices (Semple and Steel, 2003). The branch length of a phylogeny is measured in numbers of substitutions per site. The distance between two vertices, say i and i' will be denoted with $t(i, i')$ and is called genetic distance. We distinguish between rooted and unrooted trees (Figure 1.5). In the case of the rooted tree an internal node is labeled as a root (Figure 1.5B). A special case of phylogenies are star trees, that is all external nodes of the tree have one common ancestor (see Figure 1.5C).

Since we have only information about contemporary sequences, the evolutionary history needs to be reconstructed. For the reconstruction of phy-

logenetic trees there exist four main methods: distance based methods like neighbor-joining (Saitou and Nei, 1987), methods based on the parsimonious principle, i.e. maximum parsimony (Fitch, 1971), statistical methods as maximum likelihood (Felsenstein, 1981) or Bayesian inference (Rannala and Yang, 1996). A detailed description of these methods and further tree reconstruction methods are given in Felsenstein (2004). In this thesis we use a maximum likelihood approach (Vinh and von Haeseler, 2004) to reconstruct the phylogeny of an alignment.

1.2 Structure Prediction Methods

A large number of computational methods have been developed for the prediction of secondary or tertiary structures of RNA sequences. Structure prediction methods try to determine a neighborhood system $\mathcal{N}$ from one sequence or a sequence alignment. These methods can be classified as thermodynamic methods and comparative methods.

1.2.1 Thermodynamic Methods

Thermodynamic approaches compute the secondary structure for a single RNA molecule (cf. Zuker, 2000), where the best structure is found by minimizing the free energy of the sequence. Moreover, the structure of the RNA has to obey the base pairing rules. For a sequence $\mathbf{D}_i$ and a structure $\mathbf{S}_k(\mathbf{D}_i)$ we can compute the corresponding free energy $\mathbf{E}_k$ for a given secondary structure k. Essential for the determination of the free energy is the use of thermodynamic parameters that are based on experimental data (cf. Mathews *et al.*, 1999). However, the fact that not all thermodynamic parameters are known with an appropriate accuracy could lead to a reduced accuracy in the predicted structure.

In addition, with thermodynamic methods the probability distribution of secondary structures for a given sequence can be computed (Zuker, 2000; Hofacker *et al.*, 2002; Lück *et al.*, 1999). The probability of a particular structure follows the Boltzmann distribution (cf. McCaskill, 1990), that is:

$$\mathbb{P}(S_k) = \frac{1}{Z} \exp\left(-\frac{\mathbf{E}_k}{RT}\right), \tag{1.7}$$

with the molecular gas constant R, the temperature T (measured in Kelvin) and the partition function $Z = \sum_k \exp(-\frac{\mathbf{E}_k}{RT})$. The structure with the highest probability is then the structure with the minimal free energy. Furthermore, we can obtain suboptimal structures with higher energies. Thus the probability distribution (Equation 1.7) allows co-occurrence of different structures in solution that are able to rearrange into each other (Steger, 2003).

As we noted already in Equation 1, the number of possible structures is enormous, since it grows exponentially with the sequence length. To find the structure with the minimum free energy within the set of possible structures different dynamic programming algorithms were suggested (for a review see Zuker (2000)).

1.2.2 Comparative Methods

In contrast to thermodynamic approaches, comparative methods are based on the analysis of a collection of RNA molecules, where sequences are represented in a multiple alignment. Comparative methods aim to determine if two sites in an alignment are correlated. They predict a consensus structure of all investigated sequences. Comparative methods detect not only base pairs in stem regions, but also so-called tertiary dependencies like pseudo-knots, or base triples (Gutell *et al.*, 1992; Tabaska *et al.*, 1998; Ji *et al.*, 2004; Dowell and Eddy, 2004). Furthermore, comparative methods are able to

suggest a *de novo* structure from an alignment.

In general, comparative methods are statistical significance tests to prove or disprove a certain statement. These statements are formulated as a null hypothesis $\mathbf{H}_0$ and an alternative hypothesis $\mathbf{H}_1$. For example, to test for compensatory substitutions between sites j and j' the null hypothesis can be formulated as: Sites j and j' evolve independently of each other, whereas $\mathbf{H}_1$ usually states the opposite (Sites j and j' do not evolve independently). To test whether $\mathbf{H}_0$ can not be rejected or if it should be rejected in favor of $\mathbf{H}_1$ an appropriate test statistic is computed. The choice of the test statistics depends on several aspects: Are the data continuous or discrete? How many parameters are necessary to describe the null hypothesis? Can the data be grouped? etc. (for a summary on how to select the test statistic see DYTHAM (2003)). After selecting the test statistics, the alternative hypothesis is then accepted with an significance value (or significance level) α, where α equals the probability of accepting $\mathbf{H}_1$ when $\mathbf{H}_0$ is true. The significance value is set before calculating the test statistic and has usually values of 0.05 or 0.01. Finally, to decide if the null hypothesis is rejected or not the probability of observing the data under the null hypothesis, the p-value, is computed. If the p-value is smaller than α, the alternative hypothesis is accepted.

Testing on a null hypothesis may lead to wrong decisions, the type I error and the type II error. A type I error occurs if we reject the null hypothesis although it is true and therefore it is also called false positive. The probability of commiting a type I error equals the significance level α. If the null hypothesis is not rejected although the alternative hypothesis is true then this is called a type II error. The probability of a type II error is generally denoted as β.

Comparative Methods for Structure Prediction

Comparative methods can be classified in methods that use only sequence data and methods that additionally incorporate phylogenetic information. Methods using only sequence data were proposed by Gutell *et al.* (1992), Chiu and Kolodziejczak (1991) and Klingler and Brutlag (1993). These methods investigate whether the number of nucleotide pairs at two sites in an alignment differ significantly from random expectation. If so, then both sites are called correlated, i.e. they are subject to structural constraints. For instance, they may be base paired as part of a helix or they may belong to other structural elements including pseudo-knots. The null hypothesis for these methods can be formulated as follows:

$$\mathbf{H}_0 : \mathbb{P}(X_j, X_{j'}) = \mathbb{P}(X_j)\mathbb{P}(X_{j'}) \quad X_j, X_{j'} \in \{A, C, G, U\}. \tag{1.8}$$

That is, the joint probability $\mathbb{P}(X_j, X_{j'})$ of observing the nucleotide pair $(X_j, X_{j'})$ at the alignment sites (j, j') equals the probability $\mathbb{P}(X_j)\mathbb{P}(X_{j'})$ to observe these pairs under independence. The alternative hypothesis is:

$$\mathbf{H}_1 : \mathbb{P}(X_j, X_{j'}) \neq \mathbb{P}(X_j)\mathbb{P}(X_{j'}). \tag{1.9}$$

In general, the probabilities $\mathbb{P}(X_j, X_{j'})$ are estimated by the frequencies of observing the nucleotide pair $(X_j, X_{j'})$ in the alignment, whereas $\mathbb{P}(X_j)$ are estimated by the frequencies of observing the nucleotide X_j at site j. As test statistics Chiu and Kolodziejczak (1991) and Gutell *et al.* (1992) use the mutual information score:

$$I(j, j') = \sum_{X_j \in \mathcal{A}} \sum_{X_{j'} \in \mathcal{A}} \mathbb{P}(X_j, X_{j'}) \log \frac{\mathbb{P}(X_j, X_{j'})}{\mathbb{P}(X_j)\mathbb{P}(X_{j'})} \tag{1.10}$$

If site j and j' are independent, then $I(j, j') = 0$. KLINGLER and BRUTLAG (1993) used as test statistics the χ^2-test on independence:

$$X^2(j, j') = n \sum_{X_j \in \mathcal{A}} \sum_{X_{j'} \in \mathcal{A}} \frac{\{\mathbb{P}(X_j, X_{j'}) - \mathbb{P}(X_j)\mathbb{P}(X_{j'})\}^2}{\mathbb{P}(X_j)\mathbb{P}(X_{j'})}. \tag{1.11}$$

If sites j and j' are independent then $X^2(j, j')$ and $2nI(j, j')$ follow a $\chi^2_{\alpha,dof}$-distribution. The degrees of freedom (dof) equal nine, i.e. for each site three (*number of parameters - number of restrictions*), where the probabilities $\mathbb{P}(x_i)$ are restricted by $\sum_{x_i \in \mathcal{A}} \mathbb{P}(x_i) = 1$ (EVANS and ROSENTHAL, 2003).

The null hypothesis is rejected in favor of $\mathbf{H}_1$ if:

- $2nI(j, j') \geq \chi^2_{\alpha,9}$ or
- $X^2(j, j') \geq \chi^2_{\alpha,9}$,

where $\chi^2_{\alpha,9}$ is the tabulated χ^2-value with significance value α.

However, these approaches are only valid if each sequence in the alignment can be viewed as an independent sample of the same evolutionary process. As sequences are generally 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 correlations (LAPEDES *et al.*, 1999).

In the phylogenetic literature methods abound that construct a phylogenetic tree assuming no structural constraints. The resulting tree is compared to a tree reconstructed under the assumption that structural constraints are known (SCHÖNIGER and VON HAESELER, 1994; MUSE, 1995; AKMAEV *et al.*, 1999; GULKO and HAUSSLER, 1996; POLLOCK *et al.*, 1999; KNUDSEN and HEIN, 1999). These methods determine whether the evolution of sequences on a phylogenetic tree is better described by a joint evolutionary model rather than independently evolving sites. Instead of comparing two alignment sites

as for the χ^2-test and the mutual information, these approaches compute the likelihood L, that is the probability of the alignment $\mathbf{D}$ given the phylogeny and an evolutionary model. The null and alternative hypothesis for a phylogeny T are:

$$\mathbf{H}_0 : L_0 = \mathbb{P}(\mathbf{D}|T, M_0, \mathcal{N}_0)$$
$$\mathbf{H}_1 : L_1 = \mathbb{P}(\mathbf{D}|T, M_1, \mathcal{N}_1)$$

with $\mathcal{N}_0$ and $\mathcal{N}_1$ being neighborhood systems and M_0 and M_1 are the substitution models for the different hypotheses, respectively.

A higher likelihood indicates that this model fits the data better. However, models with more parameters have in general a higher likelihood. To test if the increase in the likelihood is significantly different the Akaike information criterion (AIC) or likelihood ratio tests (LRT) are applied. AIC is defined as AIC=$\ln L + 2 * k$ (k = number of free parameters of the model). The model with the lowest AIC is then preferred. Thus, AIC penalizes models with many parameters. The likelihood ratio compares directly the likelihoods of the null and the alternative hypothesis and is computed as:

$$\delta = \log \frac{L_1}{L_0} \tag{1.12}$$

The likelihood ratio test can be applied if $\mathbf{H}_0$ is nested within $\mathbf{H}_1$, i.e. the null hypothesis is a special case of the alternative hypothesis. If $\mathbf{H}_0$ is true then 2δ is distributed according to a χ^2-distribution, with the degrees of freedom that equals the difference in the number of parameters of the two hypotheses. To apply AIC and LRT the sequences have to be "extremely" long (Goldman, 1993). Therefore, Cox's test should be applied (Cox, 1962). That is, the distribution of δ is simulated based on generated data under $\mathbf{H}_0$.

However, the likelihood ratio test cannot always be applied since many models are not nested (Savill *et al.*, 2001). But they can be compared to a

unconstrained model (Goldman, 1993; Navidi *et al.*, 1991). The likelihood for this model is computed as:

$$L = \prod_{C=1}^{v} (N_C/l)^{N_C},$$

where N_C is the number of sites in an alignment that are identical with nucleotide pattern C, l the alignment length and v the number of different nucleotide patterns in the alignment.

These tests revealed that dinucleotide substitution models describe the evolution of stem regions significantly better than single nucleotide substitution models. However, these tests can only be applied if the structure of the molecule is known. Only few approaches exist to improve secondary structure prediction based on the outcome of the tests (Schöniger and von Haeseler, 1999). But if there is no information about the secondary structure, these tests cannot be applied (Akmaev *et al.*, 2000).

1.2.3 False Positive Reduction

To employ significance tests as introduced in section 1.2.2, many sequences are required. By contrast, thermodynamic methods require only one sequence. To improve accuracy, some methods combine both approaches (Lück *et al.*, 1999; Hofacker *et al.*, 2002; Juan and Wilson, 1999). However, it is very difficult to determine the appropriate significance level α to reject the null hypothesis of independently sites. Therefore, more or less arbitrary significance levels are assigned (Akmaev *et al.*, 1999). Besides the standard statistical problems, especially too few sequences, that lead to false positive correlations (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. In the following, we use the χ^2-test,

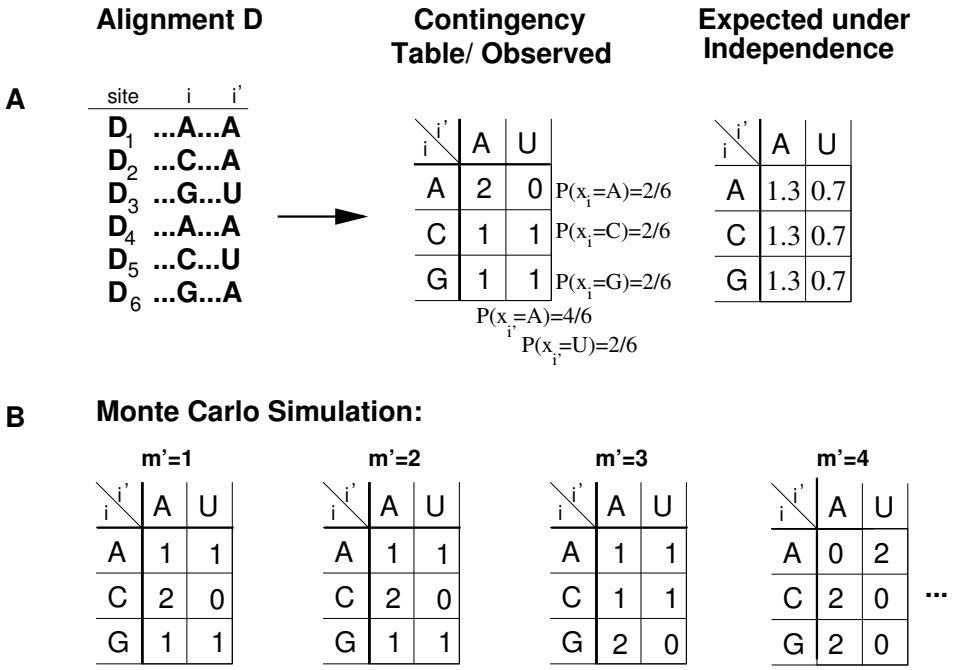

Figure 1.6: Monte Carlo Simulation exemplified for six sequences.
The contingency table contains the number of observed pairs of nucleotides for site j and j'. For the Monte Carlo Simulation m contingency tables are randomly generated based on the marginal probabilities $\mathbb{P}(x_j)$ and $\mathbb{P}(x_{j'})$ (see text for details).

that is widely used for a statistical test to discuss the influence of the number of sequences and the tree topology in detecting a neighborhood system.

Few Sequences

Statistical tests are approximately valid only for large sample size n (number of sequences). As a rule of thumb, the expected number of nucleotide pairs for the χ^2-test should be at least five for each of the 16 dinucleotides (Sachs, 1992). If the number of sequences in the alignment is small, then this is generally not the case. Moreover, an alignment position might be very conserved and therefore would not contain all of the four nucleotides.

We use a Monte Carlo Simulation to generate the χ^2-distribution based on the observed data. An example of such a simulation is displayed in Figure

1.6. We consider two sites j and j', with corresponding contingency table (Figure 1.6A). The entries equal the number of observed nucleotide pairs within the two sites. For instance, the frequency to observe the pair (A, A) equals 2. The expected number of the pair AA under independence (the null model) equals 4/3 ($n * \mathbb{P}(x_j, x_{j'}) = n * \mathbb{P}(x_j) * \mathbb{P}(x_{j'}) = 6 * 2/6 * 4/6 \approx 1.3$). The $X^2(j, j^{'})$ value of the observed base frequencies at sites $(j, j^{'})$ can be computed according to Equation 1.11. For the observed contingency table in Figure 1.6 we compute $X^2(j, j^{'}) = 1.5$.

The basis of the simulation are m contingency tables (Figure 1.6B). They are randomly generated with the condition that the frequencies $n * \mathbb{P}(x_j)$ and $n * \mathbb{P}(x_{j'})$ are the same for all tables $m' = 1, 2, \ldots, m$. Thus, the sum of dinucleotides within the rows and columns of the simulated tables has to be the same as for the observed contingency table. For each of the m contingency tables we compute a $X^2_{m'}$ value,e.g. in Figure 1.6 $X^2_{m'=1} = 1.5$. The p-value $p_{j,j'}$ of sites j and j' $(j, j' \in 1, 2, \ldots, l)$ is then estimated by the proportion of simulated $X^2_{m'}$ values greater than $X^2(j, j')$. That is:

$$p_{j,j'} = \frac{\#\{m' : X^2_{m'} \geq X^2(j, j')\}}{m} \tag{1.13}$$

If $p_{j,j'}$ is smaller than the significance level α then sites j and j' are considered to be correlated.

One should note that the state space of possible contingency tables can be small for small number of sequences. For example, in Figure 1.6 there exist only six possible contingency tables. Moreover, different tables can have the same X^2-value, e.g. for tables $m' = 1, 3$ and the observed table the X^2-value equals 1.5 for the remaining possible table X^2=6 (e.g. $m' = 4$). That is, for the above example there are only two possible X^2-values. The probability of observing $X^2 = 1.5$ is 0.8, whereas for $X^2 = 6$ the probability is 0.2 (see

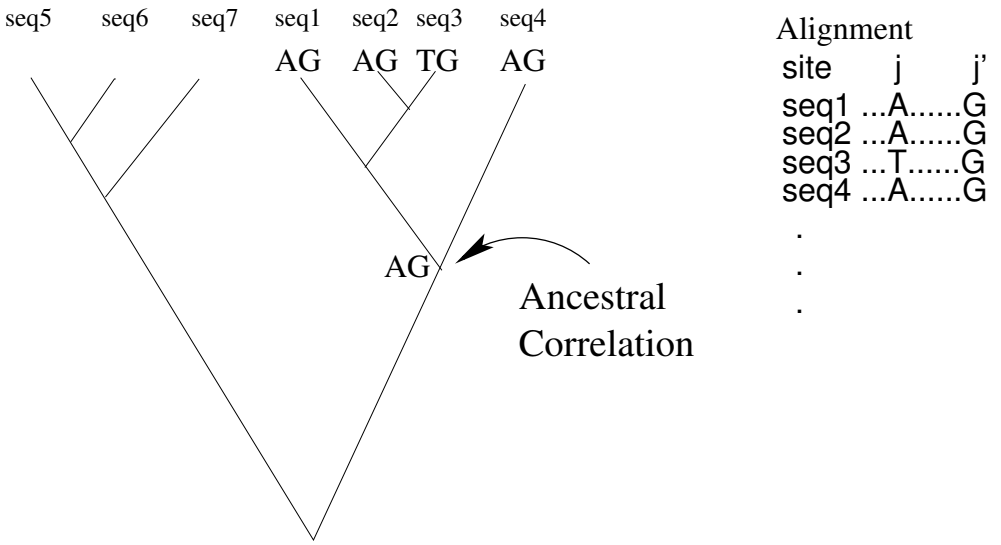

Figure 1.7: Ancestral Correlation
If the genetic distance between the internal node and the leaves of the tree is short, then they will share the same nucleotides. Therefore, sites j and j' from an alignment could be considered as correlated.

FISHER (1922)). In terms of the simulated tables, we will never reject the null hypothesis (assuming a significance value of one or five percent), since for a table with corresponding $X^2 = 1.5$ the p-value is 1 and for a contingency table with $X^2 = 6$ the p-value is 0.2.

Ancestral Correlation

Whenever we analyze a set of homologous sequences, they are related by a phylogenetic tree. That is, if we want to estimate a neighborhood system from an alignment, we have to take into account the evolutionary history (GOLDMAN *et al.*, 1996). The influence of the phylogeny when inferring correlated sites is illustrated in Figure 1.7. A phylogeny containing seven sequences is shown. If sequences are closely related (exemplarily the right part of the tree) then it is very likely that homologous nucleotides in these sequences share the

same nucleotide as their common ancestor. This is because the evolutionary distance between the sequences is too short for many substitutions to occur. For example, the common ancestor of site j carries an A and at site j' carries a G, then we will frequently observe nucleotide A at the external sequences of site j and nucleotide G at the external sequences of site j'. In a sequence alignment this would result in an over-representation of the pattern AG and could lead to the mis interpretation that site j and j' are correlated. The influence of ancestral nucleotides on the nucleotide distribution at an alignment site is called "Ancestral Correlation". To decide if sites j and j' are correlated, or if these sites show ancestral correlation, we have to investigate the evolution of nucleotides considering the ancestral states at the internal nodes of the phylogeny.

In a nutshell: To estimate dependencies from a sequence alignment, we need to distinguish between true dependencies and ancestral correlation. To do so, we require the sequence alignment as well as the evolutionary history of the sequences as represented by a phylogenetic tree.

Chapter 2

Estimating Dependencies using Subtrees

2.1 Introduction

To estimate a neighborhood system $\mathcal{N}$ from a sequence alignment, we will use the χ^2-test as test statistics. As discussed in section 1.2 such tests can strictly speaking only be applied when the sequences are related by a star phylogeny. Therefore, we will have a closer look on sequence alignments derived from star phylogenies. A further advantage of star phylogenies is that the influence of the tree topology is minimal.

To get reliable results all tests need reasonable amount of data and variation within the data (Higgs, 2000). Considering a sequence alignment, the fidelity of the obtained results depends therefore on the number of sequences and the variation within the alignment positions.

In section 2.2, we will investigate the outcome of the χ^2-test depending on these two quantities. Afterward we discuss the consequences when the χ^2-test is applied to non-star phylogenies. In section 2.3, we will introduce StarDep,

a method that predicts the consensus structure of a sequence alignment using only subtrees instead of the whole topology. We will demonstrate that under certain criteria these subtrees can be treated as star phylogenies. Thereafter, we will apply StarDep to synthetic and real data.

2.2 Simulation studies on star trees

We evaluate the ability of the χ^2-test to detect dependencies from a sequence alignment $\mathbf{D}$, where sequences evolved on a star phylogeny. We are interested in several questions: How many sequences are necessary to predict the secondary structure? Is there a relation of branch length to the number of detected correlated sites? How reliable are our estimates? We will use simulated data to answer these questions. Since we know the true dependency structure we can compare it to the outcome of the χ^2-test. For the simulations, we assumed a sequence containing 100 base pairs. The base pairs evolved according to the SH-model (Schöniger and von Haeseler, 1994) along a star phylogeny with branch length t_b. The alignments were generated using SISSI (Gesell and von Haeseler, 2006). The parameters that are used for the simulation are summarized in the appendix A.2.

If each site in the alignment evolved independently, then we expect that the nucleotide distribution $\boldsymbol{\pi}^i(t_b)$ at site i equals:

$$\boldsymbol{\pi}^i(t_b) = \boldsymbol{\pi}_{ri}\mathbf{P}(t_b), \tag{2.1}$$

with $\boldsymbol{\pi}_{ri}$ being the nucleotide distribution at the root r of site i and $\mathbf{P}(t_b)$ the transition probability matrix of a nucleotide substitution model (see Equation 1.1). We want to investigate if two sites evolve independently. We state as null hypothesis:

$$\mathbf{H}_0 : \pi(x_i, x_{i'}) = \pi(x_i)\pi(x_{i'}) \qquad \forall x_i, x_{i'} \in \mathcal{A} \tag{2.2}$$

That is, the joint probability of observing nucleotides x_i and $x_{i'}$ equals the product of observing nucleotide x_i and $x_{i'}$, independently of each other. In practice, $\pi(x_i)$ are estimated by the frequency of observing nucleotide $x_i \in \mathcal{A}$ at the alignment site i and $\pi(x_i, x_{i'})$ is approximated by the frequency of the observed dinucleotides at sites i and i'. As test statistic, we apply the χ^2-test on independence with nine degrees of freedom (Equation 1.11): The null hypothesis is rejected on a significance level α.

2.2.1 Influence of the Branch Length

First, we investigated the influence of the branch length t_b in detecting correlated sites. t_b ranges from 0.2–3.0. For each t_b we simulated 100 alignments, were each alignment contained 100 sequences and 200 sites. Thereafter, we applied the χ^2-test (Equation 1.11) and the Monte Carlo simulation described in section 1.2.3 to each alignment. That is, for an alignment containing 200 sites we analyzed all possible $\binom{200}{2}$ pairs of sites. Sites i and i' were considered to be correlated when the p-value $p_{i,i'}$ (Equation 1.13) is less equal the significance level α. For each alignment we counted the inferred number of true positive correlated sites and the number of inferred false positive correlated sites. The results are shown in Figure 2.1. Displayed are the mean numbers of true positive and false positive correlated sites for different significance values α (0.001, 0.01, 0.05).

For $\alpha = 0.05$ and $t_b = 0.2$ the average number of true positives equals 22. This number increases and equals 100 for $t_b = 1.2$. For $\alpha = 0.01$ and $\alpha = 0.001$ the number of true positives also increases up to 100 and is reached for 1.6 and 2.4, respectively. The average number of false positive base pairs is almost constant for each α. For $\alpha = 0.05$ it ranges between 3.0–5.2, for $\alpha = 0.1$ between 0.1–0.5 and for $\alpha = 0.001$ between 0.0–0.1. However, for a

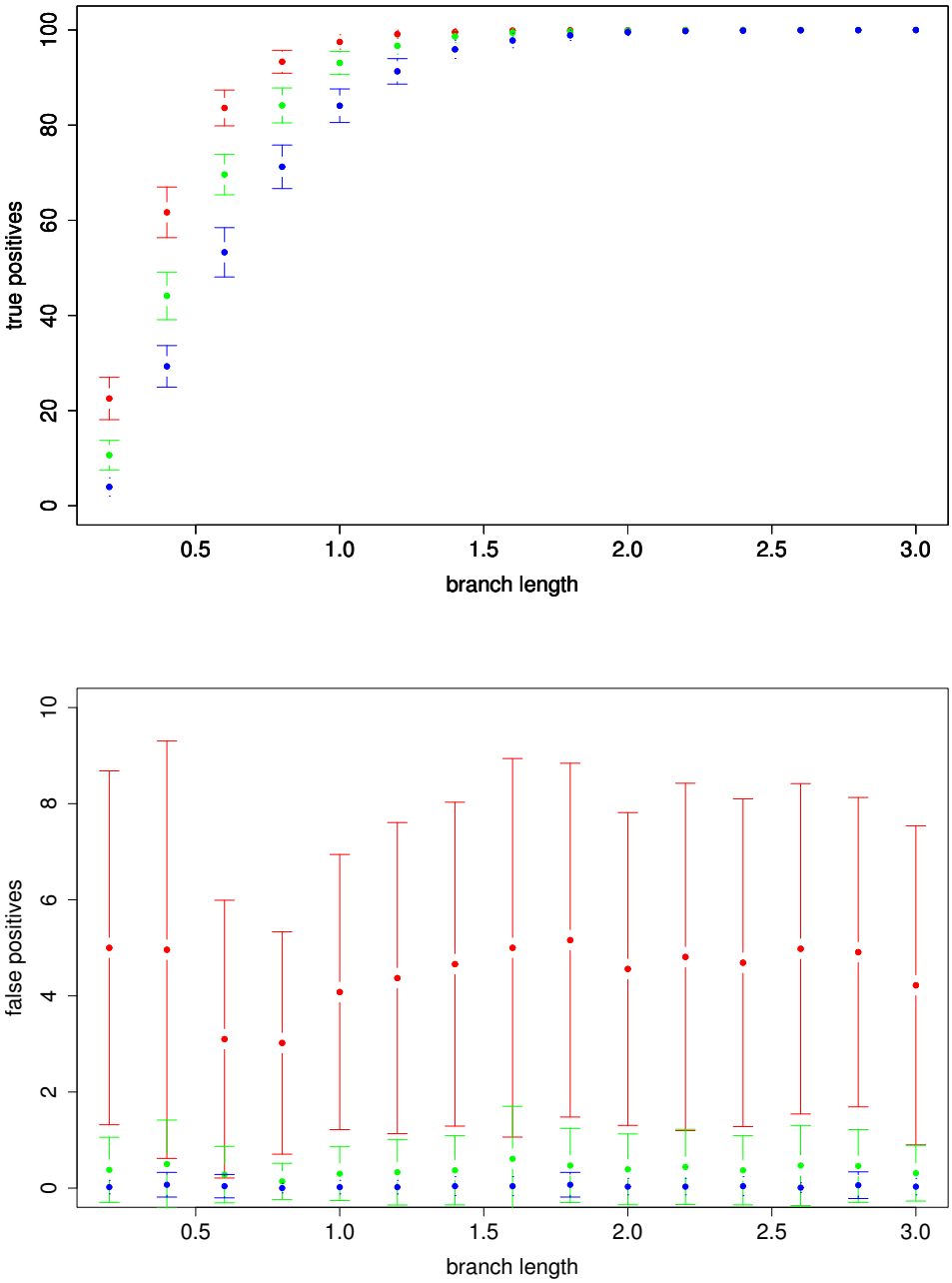

Figure 2.1: Number of detected true and false positive correlated sites depending on the branch length of the star tree for different significance levels α (red: $\alpha = 0.05$, green: $\alpha = 0.01$, blue: $\alpha = 0.001$. Error bars represent standard deviations.

significance level of five percent we expect for an alignment of 200 sites about 1000 false positives ($\binom{200}{2} \times 0.05$). Possibly, the low number of detected false positives is according to the sampling procedure of the contingency tables (see section 1.2.3).

Interestingly, in the region where the branches of the star tree are short (0.2–1.0) the χ^2-test missed many true positive correlations. This observation, however, is not surprising. If the branches have length zero, then all sequences in an alignment are identical and any test for correlation of pairs of sites is not applicable. Only if some variability at dependent sites is observed any test has the chance to suggest correlations.

2.2.2 Influence of the Number of Sequences

To determine the influence of the number of sequences n to detect correlated sites we analyzed sequence alignments with 10–1000 sequences. These alignments were generated on a "short" star tree with branch length 0.2 and a "long" star tree with branch length 1.0. We used the settings from section 2.2.1, i.e. the analysis is based on 100 alignments containing 100 sequences of length 200 comprising 100 base pairs. The results for the short tree are displayed in Figure 2.2 and for the large tree in Figure 2.3.

For the alignments derived from the short star tree, the number of detected true positives increases with increasing n. That is, for $n = 10$ we found no correlated site for all significance values ($\alpha = 0.001, 0.01, 0.05$; see Figure 2.2). For $n = 1000$ the mean number of true positives equals 11 for $\alpha = 0.001$, 22 for $\alpha = 0.01$ and 44 for $\alpha = 0.05$.

A different result emerges for the number of false positives. For alignments that ranges between 10–100 sequences this number is relative high compared to the number of detected true positives. For example, for $n = 100$ and

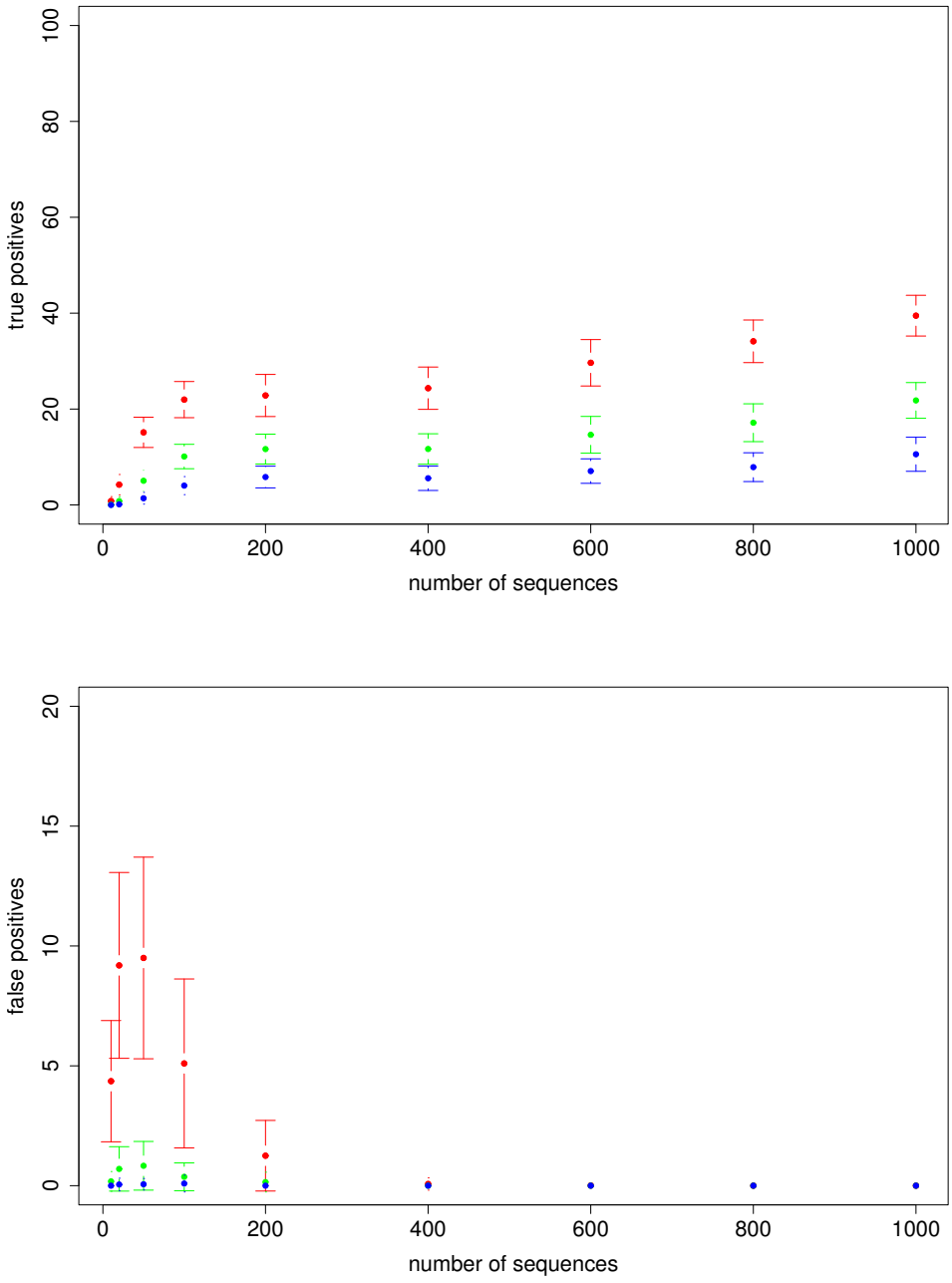

Figure 2.2: Number of detected true and false positive correlated sites depending on the number of sequences in the sequence alignment for different significance levels (red: $\alpha = 0.05$, green: $\alpha = 0.01$, blue: $\alpha = 0.001$). The length of the branches of the star tree equals 0.2 substitutions per site. Error bars represent standard deviations.

$\alpha = 0.05$ we observed on average 10 false positives. In comparison the average number of true positives equals 20. However, for increasing n the number of false positives decreases for all significance values.

We observed a distinct picture for alignments that were derived from long star trees with branch length $t_b = 1$.The results are displayed in Figure 2.3. For the investigated significance values α the mean number of detected true positives reaches 100 already for $n = 200$. The number of false positives is very large for small n. For example, for $n = 20$ and $\alpha = 0.05$ the false positive detected correlations exceeded the number of detected true positives (FP=120, TP=75). Nevertheless, for alignments with more than 100 sequences this number decreased.

The differences in the detection rates of true positives between short and long phylogenies can again be attributed to the low variability in the short star tree. Even if we investigate alignments with 1000 sequences, only 44 out of 100 base pairs ($\alpha = 0.05$) were detected for the short star tree. Consequently, we will not be able to detect the dependency structure for alignments even if we investigate many sequences. Furthermore, for alignments with only a few sequences the number of false positive correlations is very large.

2.2.3 Ancestral Correlation and χ^2-Test

As yet, we have analyzed the outcome of the χ^2-test applied to alignments that evolved on star phylogenies. Now we are interested in the performance of the χ^2-test in detecting correlated pairs when the tree topology is not a star tree. That is, we investigate alignments that are derived from bifurcating trees with 100–1000 sequences. The topologies were randomly generated and branch lengths of each topology were drawn from an uniform distribution.

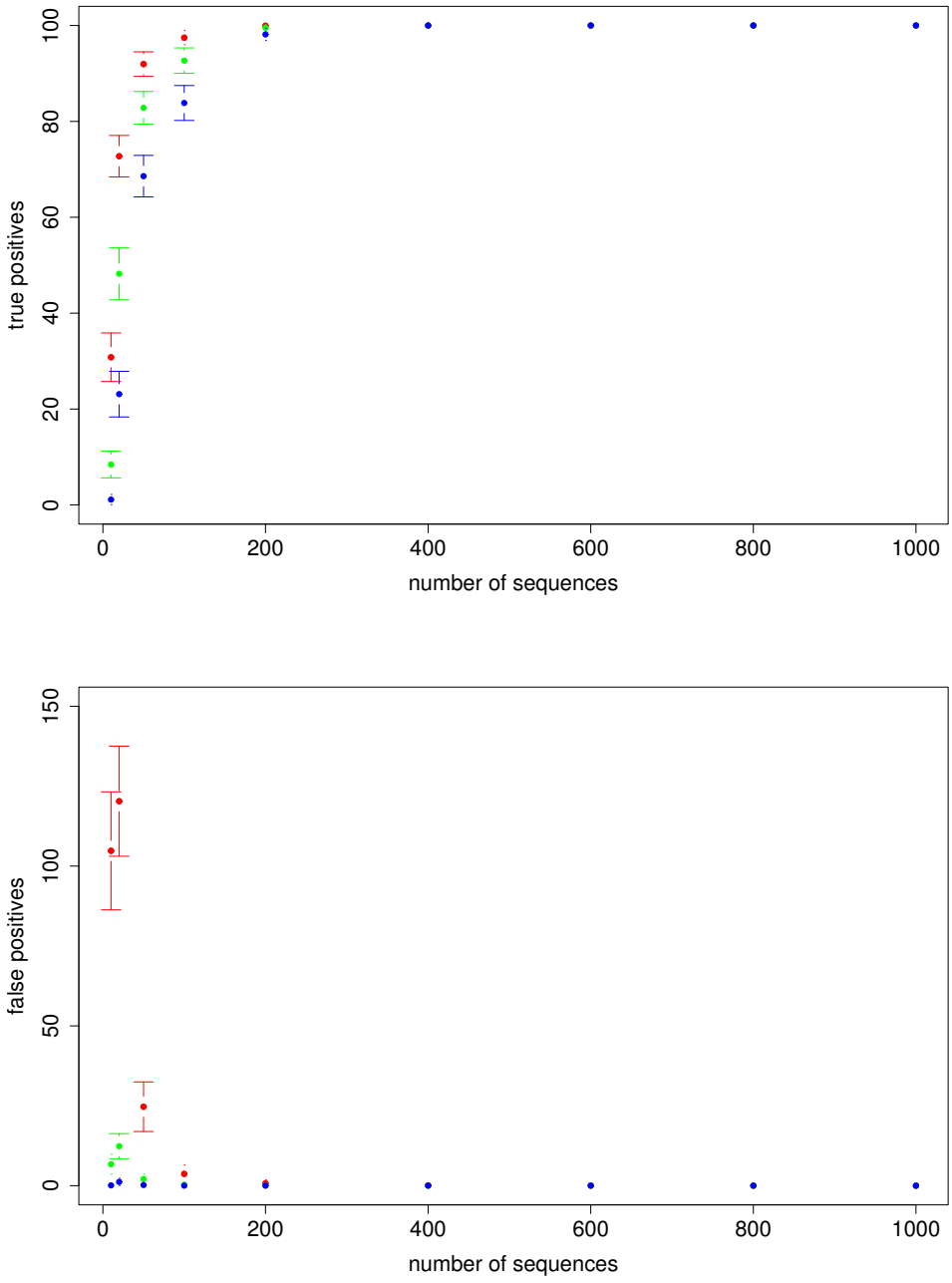

Figure 2.3: Number of detected true and false positive correlated sites depending on the number of sequences in the sequence alignment for different significance levels (red: $\alpha = 0.05$, green: $\alpha = 0.01$, blue: $\alpha = 0.001$). The length of the branches of the star tree equals 1.0 substitutions per site. Error bars represent standard deviations.

Thereafter, the branch length of the bifurcating trees were rescaled to compare the bifurcating trees to the star trees of section 2.2.2. That is, the total branch length of a bifurcating tree with n sequences equals the total branch length of the star tree with n sequences. For example: A star tree with 100 sequences has total branch length 100, the corresponding bifurcating tree has then also total branch length 100. Thus, the total number of substitutions that occurred on both trees is the same.

Our results are based on 100 simulations for each n and are summarized in Figure 2.4. Displayed are the mean numbers of detected true and false positive correlated sites, depending on the number of sequences. The number of true positives is 99 for the alignment containing 100 sequences. For alignments with a higher number of sequences this number equals 100 ($\alpha = 0.05$). A similar picture is obtained for a significance level of $\alpha = 0.01$ Thus, the prediction of true positives is comparable to that of star phylogenies (see Figure 2.3).

A different picture emerges for the number of false positive pairs. For the significance value $\alpha = 0.05$ the number of false positives exceeds the number of true positives for all n. Although for $\alpha = 0.001$ the number of false positives decreases to 90, it is still high compared to the detected true positives.

A comparison of the differences in detecting true and false positives for star and bifurcating trees is displayed in Table 2.1. The number of true positives are equal for both trees, whereas the number of false positives is for the bifurcating tree always considerably higher compared to the star tree.

In conclusion, if star phylogenies are investigated, then the ability of the χ^2-test in detecting correlated sites depends on the number of the investigates sequences and the branch length of the tree. The more sequences and the

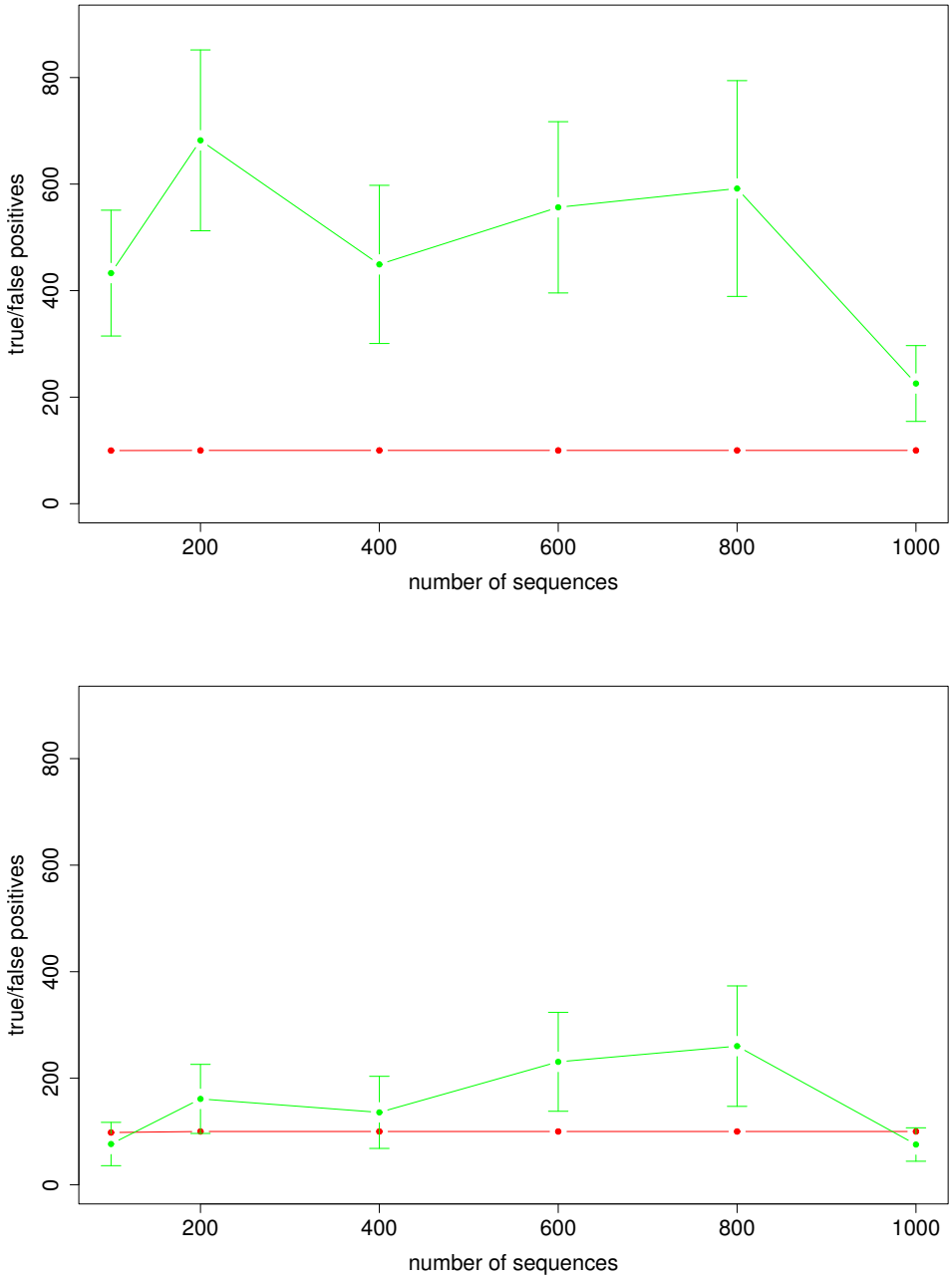

Figure 2.4: Number of detected true positive correlated sites (red line) and false positive correlated sites (green line) depending on the number of sequences in the sequence alignment for significance level $\alpha = 0.05$ (top) and $\alpha = 0.01$ (bottom). Error bars represent standard deviations.

nr.of seq.	TP_{star}	TP_{bf}	FP_{star}	FP_{bf}
100	98	99	5	420
200	100	100	0	680
600	100	100	0	570
1000	100	100	0	210

Table 2.1: Number of detected true positives (TP) and false positive (FP) for the star and bifurcating (bf) trees using the χ^2-test

longer the branch length the number of detected true positive dependencies increases. The number of false positives is small.

If the χ^2-test is applied to non-star phylogenies, a different result is obtained. Although the number of true positives is comparable to that of the star trees, we observe an inflation of false positives due to ancestral correlation. In the following section we introduce a method to detect dependencies from non-star phylogenies using the χ^2-test.

2.3 StarDep-Detecting Dependencies using Star Trees

The results from the previous section revealed that the application of the χ^2-test to non-star phylogenies may lead to a high number of false positive correlated pairs if it is applied to bifurcating trees. In this section we will introduce StarDep, a method that detects correlated sites from sequence alignments. As reported in section 1.2.3 it is often difficult to assign an appropriate significance value α. StarDep comprises a method that automatically determines a significance value. This method is based on minimum p-values (Ge *et al.*, 2003). StarDep analyzes subtrees of the phylogeny. We will show

that these subtrees can be considered as star trees (section 2.3.3). Before describing StarDep in detail, we give a brief motivation.

2.3.1 Motivation

We will explain our method by means of Figure 2.5. Displayed is a phylogeny (Figure 2.5A) that is based on an alignment of 20 sequences. The phylogeny can be subdivided into five groups ($T_1 - T_5$). The genetic distance between pairs of sequences within a group shall be "small" whereas the distance between sequences from different groups shall be "large". Since the sequences are closely related within a group, we expect high ancestral correlation. Moreover, the application of the χ^2-test would result in a high number of false positive correlated sites (see also Figure 2.4). To reduce the influence of ancestral correlation we will select sequences where the genetic distance between pairs of sequences exceeds a threshold t^S (a definition of t^S is given in section 2.3.2). Assuming that the distance between each group is "large enough",we can choose a sequence from each group resulting in a subtree of the original phylogeny. Obviously, there exist many possible subtrees that fulfill this condition. Two examples are displayed in Figure 2.5B. Moreover, we can assume for large t^S that no ancestral correlation is present. This allows the usage of the standard χ^2-test to detect correlated sites.

The analysis of subtrees may lead to the problem that the number of the selected sequences can be small. This may result in many false positive correlations (see also Figures 2.2 and 2.3).

In section 2.3.4 we will show that one subtree is not sufficient to obtain accurate results by means of detecting a high number of true positive and a low number of false positive correlated sites. Therefore, we will analyze alignments from many subtrees. That is, from each alignment derived from

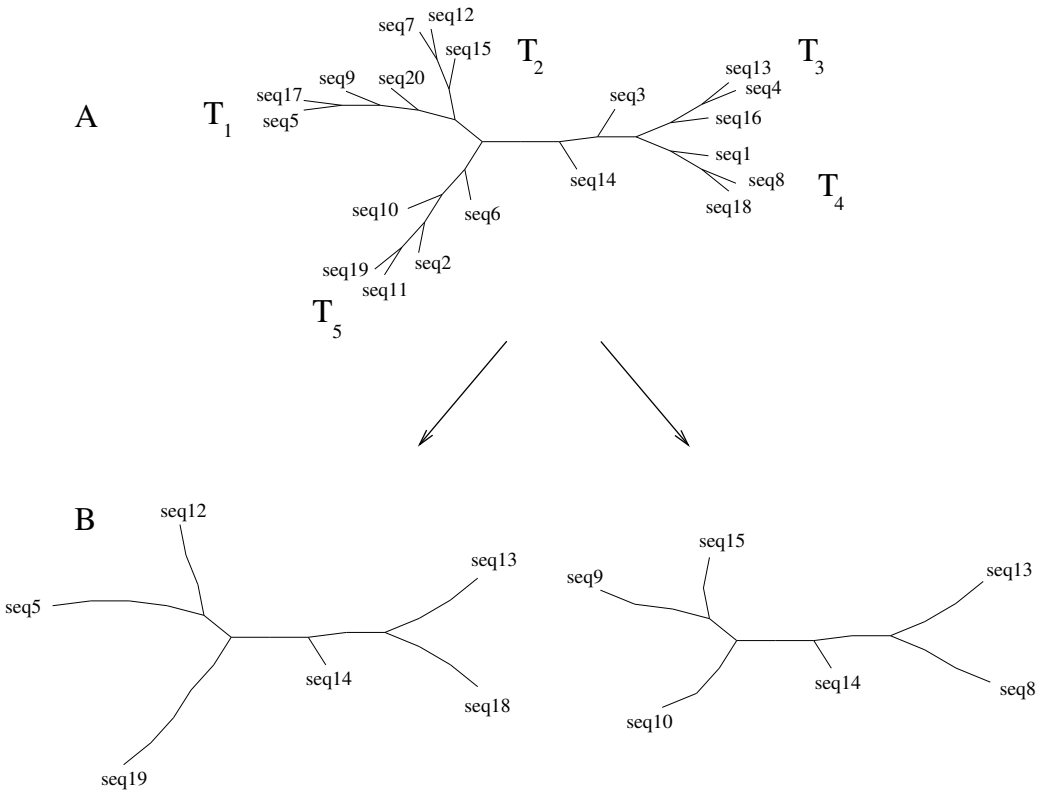

Figure 2.5: (A) The phylogeny of 20 sequences containing five subgroups ($T_1 - T_5$). (B) Two possible subtrees. The genetic distance between pairs of sequences of the subtrees has to be larger than t^S. The alignments derived from these subtrees are subject to a further analysis (see text for details).

the corresponding subtree we compute the number of true and false positive correlated pairs. Afterward, we use a summary statistics to display the results.

2.3.2 Estimating Time to Stationarity

In this section, we will define the meaning of "large" genetic distance. Therefore, we consider the sequences $\mathbf{D}^i$ and $\mathbf{D}^S$, where $\mathbf{D}^i$ is the ancestral sequence of $\mathbf{D}^S$. We are interested in the question: How large has the genetic distance

$t(\mathbf{D}^i, \mathbf{D}^S)$ between these two sequences to be, that $\mathbf{D}^S$ carries no information on the ancestral sequence? That is, when can we not reconstruct the ancestral sequences $\mathbf{D}^i$? For large genetic distances and high substitution rates it is shown that this reconstruction is impossible (cf MOSSEL, 2003). The article of MOSSEL (2003) also introduces a bound for the probability to determine the ancestral state.

Here we use a different approach. We assume that the ancestral sequence cannot be reconstructed when the base composition of $\mathbf{D}^S$ equals the stationary distribution. The nucleotide distribution after t time units equals: $\boldsymbol{\pi}(t) = \boldsymbol{\pi}^i\mathbf{P}(t)$ with $\boldsymbol{\pi}^i$ the initial distribution of sequence $\mathbf{D}^i$ and $\boldsymbol{\pi}(t)$ the nucleotide distribution of sequence $\mathbf{D}^S$ (see Equation 1.5). As we discussed in section 1.1.2 when $\boldsymbol{\pi}(t)$ reaches the stationary distribution $\boldsymbol{\pi}^S$, then all information about the initial distribution is lost. However, this case can only be obtained when t approaches infinity (see also Equation 1.5). Since this is not possible we will follow another strategy and ask for which time we can assume $\boldsymbol{\pi}(t)$ not to be significantly different from the stationary distribution. Thus, we state as null hypothesis:

$$\mathbf{H_0} : \boldsymbol{\pi}^S = \boldsymbol{\pi}(t) = \boldsymbol{\pi}^i\mathbf{P}(t). \tag{2.3}$$

The time for which we cannot reject the null hypothesis is then denoted by t^S, the time to stationarity.

In this context, the choice of the initial distribution $\boldsymbol{\pi}^i$ is problematic. For different initial distribution the estimated time t^S may differ dramatically. Therefore, it seems more appropriate to estimate for different initial distributions the corresponding t^S and then select the maximum to be the time to stationarity. We decided to choose as initial distribution the four cases where the initial sequence consists only of one nucleotide. The initial distributions are denoted by $\boldsymbol{\pi}^i_A$, $\boldsymbol{\pi}^i_C$, $\boldsymbol{\pi}^i_G$, $\boldsymbol{\pi}^i_U$, respectively. Exemplary, the

initial distribution $\boldsymbol{\pi}_A^i$ has then the form $\boldsymbol{\pi}_A^i = (1, 0, 0, 0)$. Intuitively, this four distributions consider the case that we start with a certain nucleotide. For these four cases we obtain:

$$\boldsymbol{\pi}(t, \rho) = \boldsymbol{\pi}_\rho^i \mathbf{P}(t), \tag{2.4}$$

with the nucleotide distribution $\boldsymbol{\pi}(t, \rho)$ that evolved for a time t starting with the root nucleotide $\rho \in \mathcal{A}$.

To test whether $\boldsymbol{\pi}(t, \rho)$ equals the stationary distribution we assign the following null hypothesis

$$\mathbf{H_0} : \boldsymbol{\pi}(t, \rho) = \boldsymbol{\pi}^S. \tag{2.5}$$

Since there are four initial distributions we have to reject four null-hypothesis. That is, we are looking for the times t_ρ where we cannot reject the null-hypothesis for a given significance level α. We choose the maximum of the four times to be the time to stationarity t^S, i.e.

$$t^S = \max\{t_A, t_C, t_G, t_U\} \tag{2.6}$$

As test statistic we use the χ^2-test with three degrees of freedom. For l nucleotides that is

$$X^2(t, \rho) = l \sum_{j \in \mathcal{A}} \frac{(\pi_j(t, \rho) - \pi_j^S)^2}{\pi_j^S}. \tag{2.7}$$

We obtain t_ρ if $X^2(t, \rho) \leq 7.8$ (Bronstein and Semendjajew, 1996), the critical value for a χ^2-distribution with three degrees of freedom and a significance level $\alpha = 0.05$. Finally, t^S is computed according to Equation 2.6. The time t^S is a measure of how long a sequence needs to evolve until it reaches stationarity. Thus ancestral correlation is not present for sequences i, i' when the genetic distance $t(D_i, D_{i'})$ is larger than t^S, that is

$$t(i, i') > t^S \tag{2.8}$$

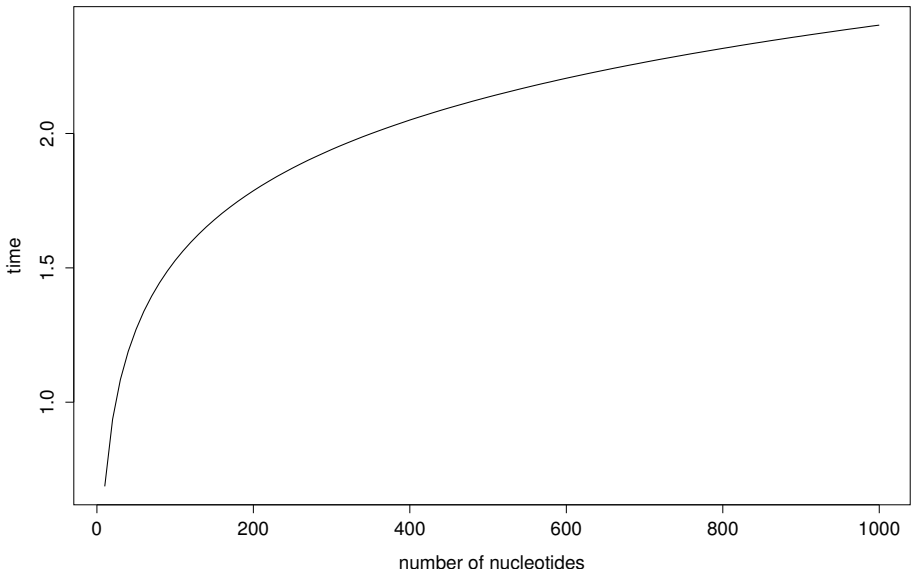

Figure 2.6: The computed time to stationarity t^S, measured in numbers per substitutions per site, depending on the number of nucleotides. t^S was estimated using the HKY substitution model (HASEGAWA *et al.*, 1985) (see text for details).

One should note, that Equation 2.7 depends on the sequence length l. Figure 2.6 visualizes this influence. Displayed is the estimated time to stationarity depending on the number of nucleotides. We used the HKY-model substitution model (HASEGAWA *et al.*, 1985) with the stationary distribution $\boldsymbol{\pi}^S = (0.2, 0.3, 0.3, 0, 2)$ and the transition transversion ratio of 1.2. With increasing l, t^S also increases. For example, if l equals 1000 then t^S is about 2.5. That is, the genetic distance between two sequence equals 2.5 substitutions per site. From the evolutionary point of view this is a relatively large number of substitutions. Therefore, it is unclear if the introduced method is an appropriate measure for t^S. Moreover, the estimation of t^S depends on the substitution model. The influence of these models on t^S is difficult to determine since the space of possible parameter compositions is infinite.

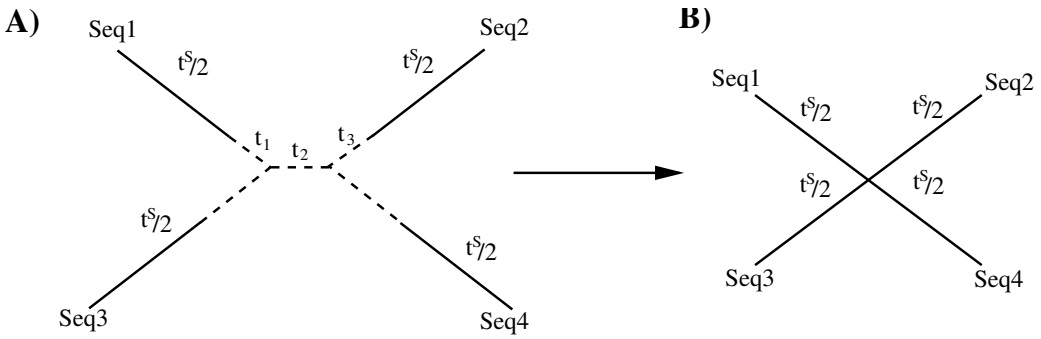

Figure 2.7: **A**: A bifurcating phylogeny containing four sequences.
B: If the genetic distance between pairs of sequences is greater than t^S then this phylogeny can be considered as star like.

2.3.3 Subtrees are equivalent to Star Trees

If we select a subtree, where the genetic distance between pairs of sequences is larger than t^s then this tree can be considered as a star tree with n leaves and branch length $t^S/2$. To see this, we will use the example shown in Figure 2.7. Displayed is a phylogeny containing four sequences, where the genetic distance between each pair of sequences is larger than t^S. Furthermore, we assume that the sequences evolved according to a Markov process with transition matrix $\mathbf{P}(t)$ (see section 1.1.2).

We will consider the evolution from sequence Seq1 to Seq2. The genetic distance between the two sequences is $t_{1,2} = t^S/2 + t_1 + t_2 + t_3 + t^S/2$, respectively. The nucleotide distribution of Seq1 is denoted by $\boldsymbol{\pi}^1$. Using the Chapman Kolmogorov equation $\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$ (Bremaud, 1999) and the stationarity assumption of the Markov process $\boldsymbol{\pi}^S = \boldsymbol{\pi}^S\mathbf{P}(t)$ (see Equation 1.4) we obtain:

$$\boldsymbol{\pi}^1\mathbf{P}(t_{1,2}) = \boldsymbol{\pi}^1\mathbf{P}(t^S/2)\mathbf{P}(t_i)\mathbf{P}(t^S/2) = \boldsymbol{\pi}^1\mathbf{P}(t^S)\mathbf{P}(t_i) \approx \boldsymbol{\pi}^S\mathbf{P}(t_i) = \boldsymbol{\pi}^S, \tag{2.9}$$

where $t_i = t_1 + t_2 + t_3$ is the sum of the length of the internal branches. The

result of Equation 2.3.3 is that the internal branches of the phylogeny do not need to be considered at all. The same conclusion holds for all other pairs of sequences. Moreover, this description leads to the star phylogeny in Figure 2.7 where the length of every branch equals $t^S/2$. Note: that Equation holds only if the multiplication of the transition matrices is commutative. For the Markov Process as introduced in section 1.1.2 this is true.

2.3.4 Reduction of false positive Correlations

Consider now a phylogenetic tree $\mathcal{T}$ with n sequences. Assuming we also know t^S. Thus, we can select a subtree $\mathcal{T}_1 \subseteq \mathcal{T}$ where the genetic distance between pairs of sequences is greater than t^S. Since $\mathcal{T}_1$ can be considered as star like we can apply the χ^2-test to the sequences derived from $\mathcal{T}_1$.

This approach can be applied only if pairs of sequences in $\mathcal{T}$ exist whose pairwise genetic distance is greater or equal than t^S otherwise $\mathcal{T}_1$ contains no sequences. Moreover, $\mathcal{T}_1$ should comprise many sequences since few sequences increase the number of false positives. Although we could apply the Monte Carlo simulation (section 1.2.3), many false positives will be detected.

To reduce false positives we will use many subtrees from the full phylogeny $\mathcal{T}$, resulting in $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_v$ subtrees (see also Figure 2.8). From each subtree we obtain the corresponding alignment $\mathbf{D}_k$ ($k = 1, 2, \ldots, v$; $\mathbf{D}_k \subseteq \mathbf{D}$). For each alignment $\mathbf{D}_k$, the p-value $p^k_{i,i'}$ for site i and i' is computed according to Equation 1.13. That is for, site i and i' we get v p-values. The average p-value for each pair of sites is given by:

$$\bar{p}_{i,i'}(\mathbf{D}) = \frac{1}{v} \sum_{k=1}^{v} p^k_{i,i'} \tag{2.10}$$

Intuitively, a small average p-values points to correlations that are present in all alignments. On the other hand, false positive pairs that are present in

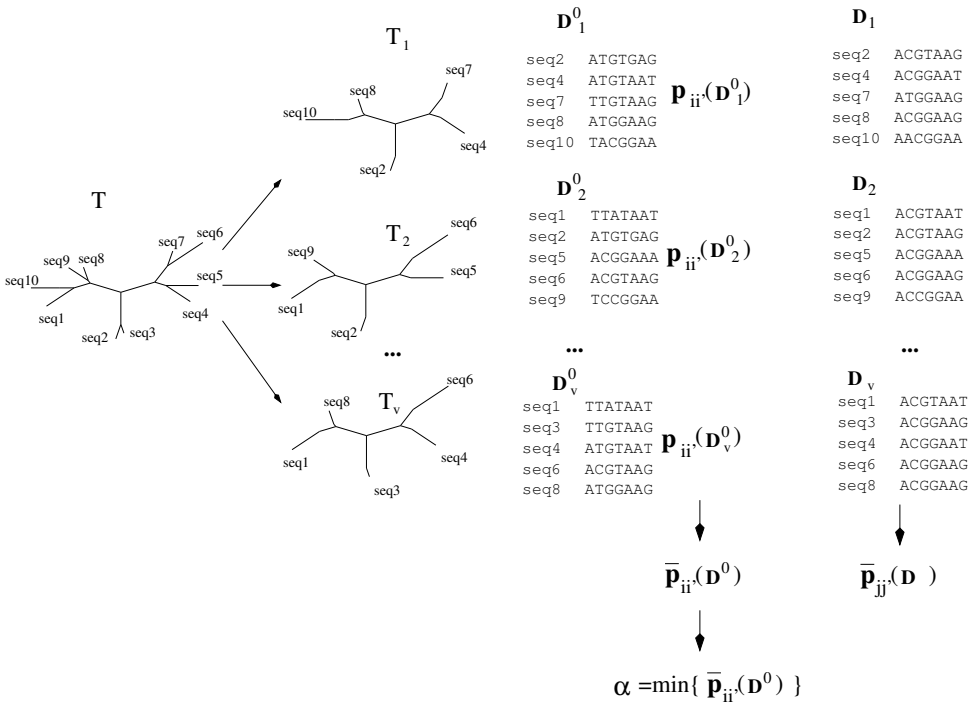

Figure 2.8: Assigning dependent pairs: From the phylogeny T the subtrees $T_1, T_2, \ldots, T_v$ are derived. The genetic distance between pairs of sequences in the corresponding subtree is greater or equal than t^S. For sites i and i' the p-value is computed for each alignment $\mathbf{D}_1^0, \mathbf{D}_2^0, \ldots, \mathbf{D}_v^0$ and their average $\bar{p}_{ii'}$. The minimum of the average p-values equals the significance level α. If the average p-value $\bar{p}_{jj'}(\mathbf{D})$ of sites j and j' is less equal α then they are considered to be correlated. See text for details.

one alignment should not be observed in another alignment. Thus having a high p-value in most subtrees. As discussed in section 1.2.3, the estimated p-values can be large and the average p-value can be large, too. Thus, we are not able to decide whether $\bar{p}_{i,i'}$ is significant.

To assign a significance value α, we generate an alignment $\mathbf{D}^0$ based on the substitution model $\mathcal{M}$ and the phylogeny $\mathcal{T}$ using Seq-Gen (Rambaut and Grassly, 1997). $\mathbf{D}^0$ constitutes an alignment of independently evolving sites. With $\mathbf{D}_k^0$ we denote the alignment derived from the subtree $\mathcal{T}_k$ $(k = 1, 2, \ldots, v)$.

As before, we apply the χ^2-test to each pair of sites of the alignment $\mathbf{D}_k^0$ and compute the average p-value $\bar{p}_{i,i'}(\mathbf{D}^0)$. We end up, with a collection of $l(l-1)/2$ average p-values. These average p-values characterize a distribution under the null hypothesis of independently evolving sites. Thus, the minimum of the average p-values describes therefore this pair of sites that can still be explained by independent evolution. We choose this value as the significance level α:

$$\alpha = \min_{i \neq j} \{\bar{p}_{i,j}(\mathbf{D}^0)\}. \tag{2.11}$$

Two sites in $\mathbf{D}$ are considered to be correlated if the average p-value of these sites is smaller than α, i.e. $\bar{p}_{i,i'}(\mathbf{D}) < \alpha$.

2.3.5 Estimating Dependencies on Star Like Trees: StarDep

Now we are ready to explain our strategy to detect correlated sites in more detail. The objective of StarDep is the estimation of a neighborhood system from a sequence alignment $\mathbf{D}$ (see also Figure 2.9). StarDep comprises several steps summarized in Figure 2.9. First, the phylogeny $\widehat{\mathcal{T}}$ and the parameters of the single nucleotide substitution model $\widehat{\mathcal{M}}$ are estimated from the sequence alignment $\mathbf{D}$ (Figure 2.9A) using IQPNNI (Vinh and von Haeseler, 2004). Based on $\widehat{\mathcal{T}}$ and $\widehat{\mathcal{M}}$, we generate a sequence alignment $\mathbf{D}^0$ with sequence length l (Figure 2.9B).

Using the parameters of the substitution model we can compute t^s (Section 2.3.2). t^s allows the selection of star like subtrees. The corresponding alignments are used for the inference of correlated sites. To obtain the subtrees, we create an $n \times n$ adjacency matrix $\mathbf{d}$, with entries

$$d_{ij} = \mathbb{1}(t(i,j) > t^S).$$

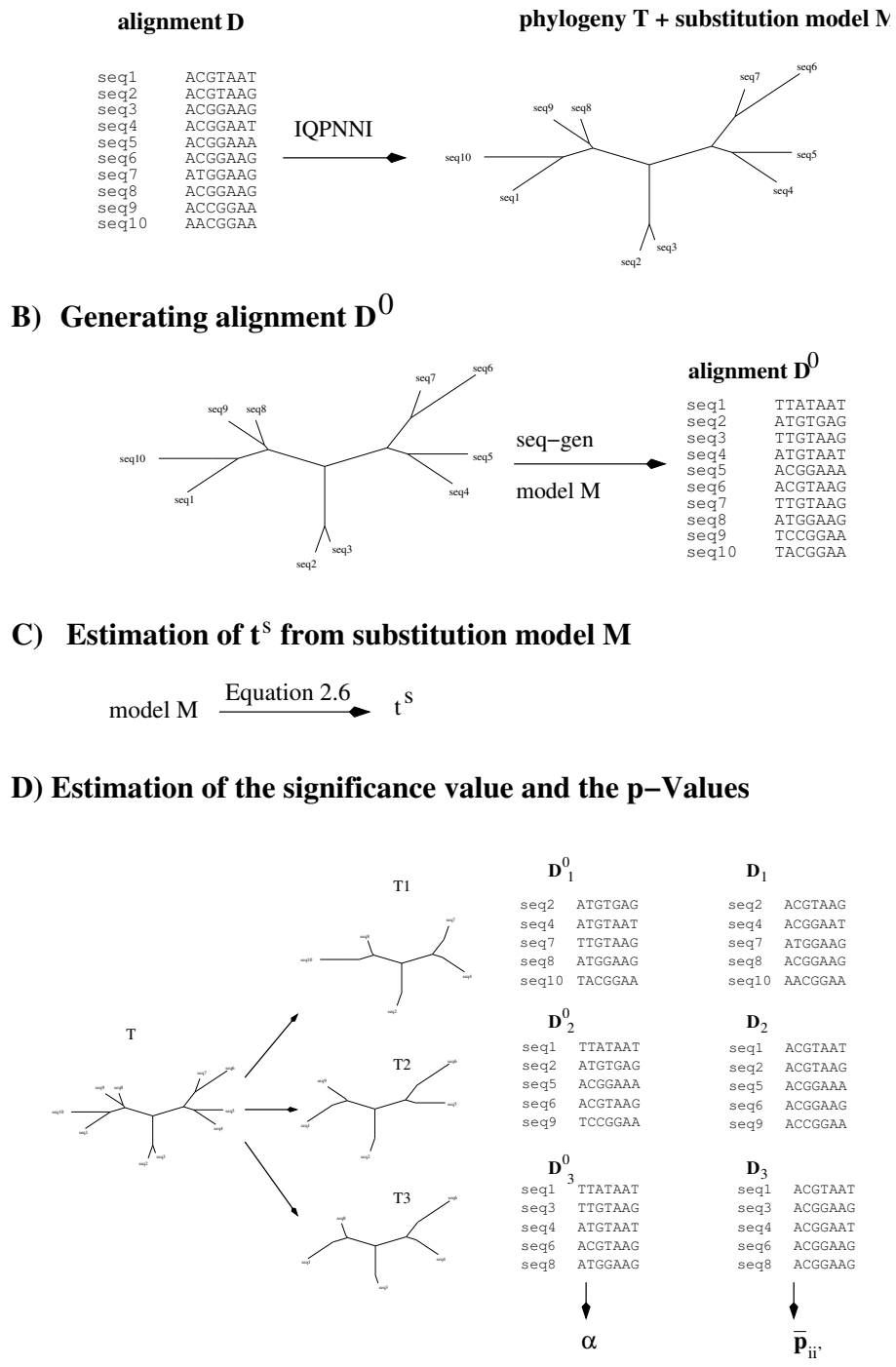

Figure 2.9: Summary of StarDep for an alignment of 10 sequences (see Text for details).

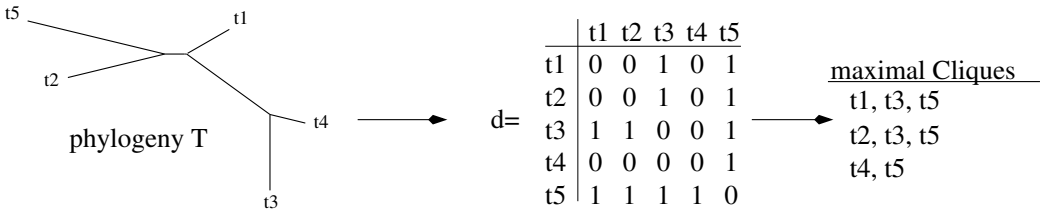

Figure 2.10: Finding subtrees: From the phylogeny T, the adjacency matrix **d** is derived. If the genetic distance between two sequences is greater than t^S then d_{ij} equals one, otherwise is is zero. From **d** maximal cliques are determined corresponding the subtrees that are used to a further analysis.

That is, if the genetic distance of two sequences is larger than t^S then d_{ij} equals one, otherwise it is zero. Finding the subtrees corresponds to the problem of finding maximal cliques of an undirected graph (Lauritzen, 1996). As a clique we define the set of sequences where the pairwise genetic distance of this sequences is greater t^s. A maximal clique is a clique that cannot be extended by an additional sequence. An example of maximal cliques for a phylogeny of five sequences is given in Figure 2.10.

From d_{ij} we find the maximal cliques using the *cliques* function of the *ggm* package as implemented in R (Marchetti and Drton, 2006). We end up with a collection of maximal cliques, where each clique corresponds to a subtree. We draw randomly p subtrees $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_p$ from the set of maximal cliques to a further analysis, where subtrees have to contain at least three sequences. To each alignment $\mathbf{D}_p$ derived from T_p we apply the χ^2-test to all pairs of sites. This results in the average p-values $\bar{p}_{ii'}$ (Figure 2.9D see also Section 2.3.4). If this value is below the significance value α, then these sites are considered to be correlated. The significance value α is estimated from $\mathbf{D}^0$ according to Equation 2.11.

2.4 Application

2.4.1 Performance on Synthetic Data

We evaluated the ability of StarDep to detect the neighborhood system of a RNA-molecule from a multiple sequence alignment. To this end, we carried out a simulation. We assumed the secondary structure of an artificial molecule as displayed in Figure 2.11. The molecule is 200 bases long and contains seven base paired regions (I-VII), where region VII represents a pseudo-knot. The base paired regions (54 base pairs) evolved according to the SH-model (SCHÖNIGER and VON HAESELER, 1994, see Equation 1.6) and the 92 remaining sites evolved according to the HKY model (HASEGAWA *et al.*, 1985). The parameter of the substitution models are summarized in appendix A.2. This molecule evolved on three different phylogenetic trees with 100 leaves using SISSI (GESELL and VON HAESELER, 2006). The trees were randomly generated, where the branch length were drawn from a uniform distribution with mean 0.1, 0.2 and 0.3. The result of such a simulation, $\mathbf{D}^1_{data}$, $\mathbf{D}^2_{data}$, $\mathbf{D}^3_{data}$ respectively, is then subject to a further analysis. We started with the estimation of the phylogenies $\mathcal{T}^g$ and the parameters of the substitution model $\mathcal{M}^g$ (HASEGAWA *et al.*, 1985) from the three alignments ($g = 1, 2, 3$). The total branch length of the estimated trees is 16.8, 31.1 and 56.8. Based on the substitution models we computed t_p^S using Equation 2.7 (see also Table 2.2). Figure 2.12 displays the graph $\chi^2(t, \rho)$ depending on t, exemplary for $\rho = A$ for alignment $\mathbf{D}_1$. For $t = 0$, χ^2 is about 500, with increasing t this number decreases. For all $t \geq 1.53$ the $X^2(t, \rho)$ is less than the critical value 7.8. Thus t_A equals 1.53. For t_C, t_G and t_U, we computed 1.4, 1.3 and 1.4, respectively. The maximum of these four values equals $t_1^S = 1.53$.

For the other two trees ($g = 2, 3$) we obtained $t_2^S = 1.5$ and $t_3^S = 1.51$.

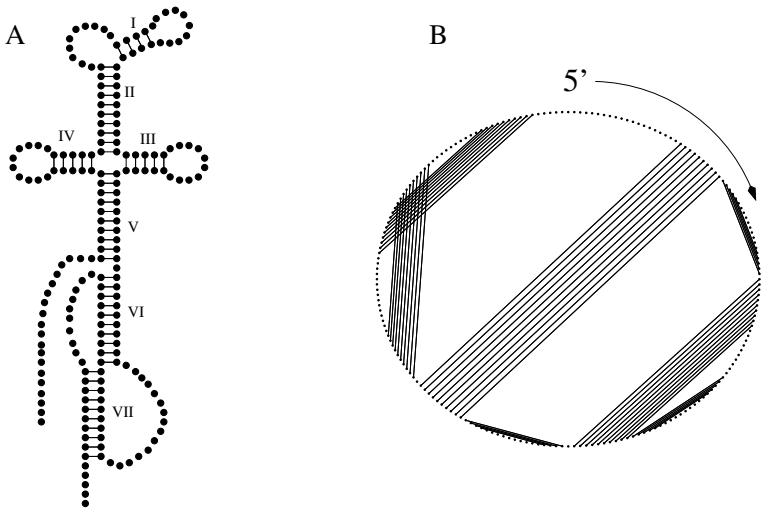

Figure 2.11: Two Representations of the Dependency Structure of $\mathbf{D}_{data}$.
A) schematic representation B) circle plot, bases are represented by vertices and correlated pairs by edges.

Since the sequences of the three alignments evolved according to the same substitution model these values should be identical. The differences within these values are due to slight differences that can be traced back to slight differences in the estimation of the parameters of the substitution model.

Using t_g^S, we draw randomly 100 maximum cliques (subtrees) from each phylogeny $\mathcal{T}_p^g$ $(p = 1, 2, \ldots, 100)$. The number of sequences of the subtrees derived from $\mathcal{T}^1$ ranges from 3 to 5, for $\mathcal{T}^2$ from 12-17 and for $\mathcal{T}^3$ from 25 to 29. We compute the significance values as explained in Section 2.3.5. The resulting estimates are $\alpha_1 = 0.46$, $\alpha_2 = 0.04$ and $\alpha_3 = 0.002$.

Finally, we compute the average p-values (Equation 2.10). Two sites within $\mathbf{D}^g$ are called correlated, when $\bar{p}_{i,i'}^g < \alpha_g$

Since we know the true dependency structure of the investigated molecule, we can compare it to the outcome of StarDep. The results are summarized in Table 2.2. For alignment $\mathbf{D}^1$ we detected two true positive correlated sites, for alignments $\mathbf{D}^2$ and $\mathbf{D}^3$, we obtain 23 and 43, respectively. For all

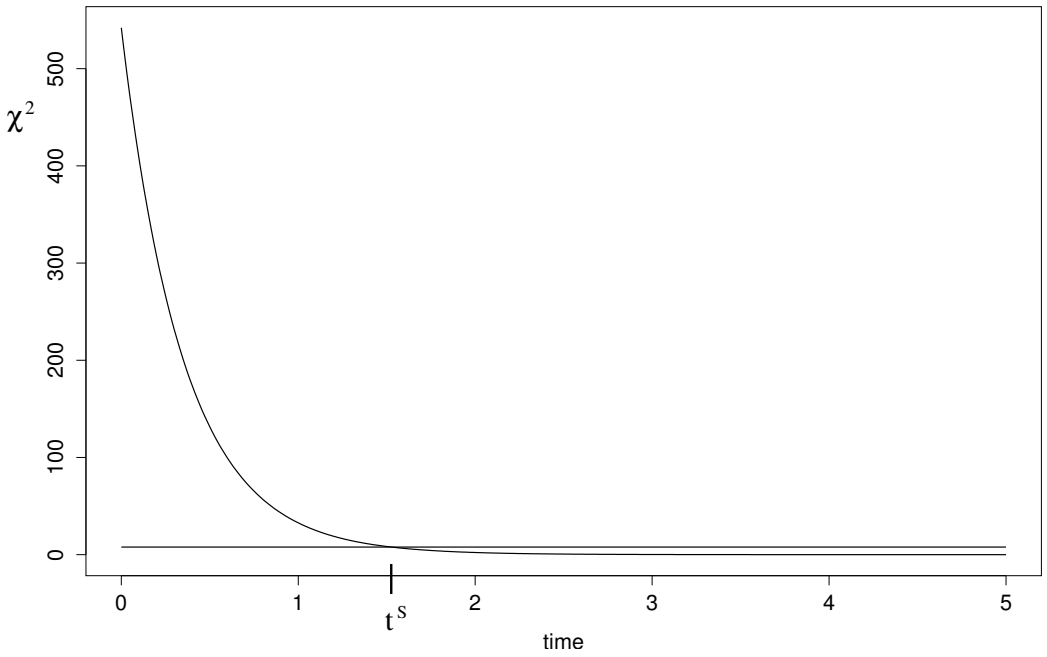

Figure 2.12: Graph of $X^2(t, \rho)$ vs t (see Equation 2.7) exemplary for $\rho = A$. For a significance level $\alpha = 0.05$ the critical value $\chi^2_{\alpha,3}$ of a χ^2-distribution with three degrees of freedom equals 7.8 (horizontal line). For $t > t^S$ the distributions $\boldsymbol{\pi}(t)$ is not significant different from the stationary distribution $\boldsymbol{\pi}^S$ (see text for details)

three alignments no false positive correlated site was detected. The increase of detected true positive correlations with increasing total branch length reflects the results from Figure 2.1, i.e if the total branch length are too small, then it is difficult to detect correlations. The influence of the number of used subtrees p for estimating correlated sited is shown in Figure 2.13, exemplary for $\mathbf{D}^2$. Displayed are the number of true positive (green line) and false positive (red line) correlated sites. The number of true positives is almost constant for all p. We detected 23 out of 54 true positive correlated sites for $p = 100$. The number of false positives decreases with increasing p, i.e. for $p = 1$ it is about 239, for $k \geq 100$ it is zero. We conclude that

tree	tbl	t^S	nr. of seq.	α	TP	FP	nr.of.bp.
$\mathcal{T}^1$	16.8	1.53	3-5	0.46	2	0	54
$\mathcal{T}^2$	31.1	1.5	12-17	0.04	23	0	54
$\mathcal{T}^3$	56.8	1.51	25-29	0.002	43	0	54

Table 2.2: Results of StarDep applied to alignments derived from three different phylogenies. 'tbl' is the total branch length of the phylogenies, 't^S' is the estimated time to stationarity, 'nr. of seq.' is the range of number of sequences in the subtrees, α is the estimated significance level obtained for 100 subtrees. TP and FP are the number of detected true- and false positive correlated sites and 'nr.of.bp.' is the number of base pairs.

the number of false positives can be reduced when we include many subtrees in our analysis. This observation is not surprising. If correlations are present than they should be verified in each alignment derived from the subtree. False positives correlations however that are present in one alignment are probably not present in another alignment (see Figure 2.13). Thus, the average of the p-values reflects the correlations that are present in all alignments.

2.4.2 Results of the tRNA Alignment

We applied StarDep to a sequence alignment of 135 eubacterial tRNA sequences (alignment length 99; see also appendix A.1). Transfer RNA are small molecules with a well-defined secondary structure. The cloverleaf structure (Sprinzl *et al.*, 1998) is displayed in Figure 2.14A (see also Figures 1.2). It contains four helical regions containing 22 base pairs represented as lines in the circle plot. To estimate the structure of the alignment, we performed all steps outlined in StarDep. Based on the alignment, we used IQPNNI (Vinh and von Haeseler, 2004) to reconstruct the phylogeny as well as the parameter of the substitution model $\mathcal{M}$ (base frequencies, transition transver-

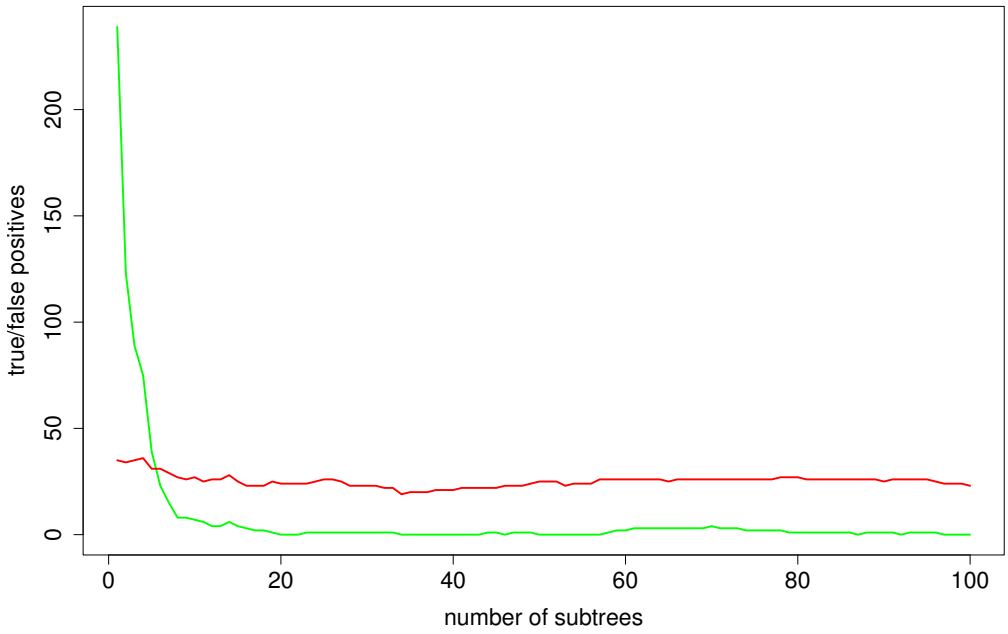

Figure 2.13: The number of detected true (red) and false (green) positive correlated sites dependent on the investigated subtrees (displayed for $\mathbf{D}^2$). The used significance value α_2 is based on 100 subtrees. The number of true positives remains relative constant, whereas the false positives decrease to zero.

sion ratio). We used the HKY-model (Hasegawa *et al.*, 1985). Using $\mathcal{M}$ we obtain for $t^S = 1.6$. We select randomly 100 subtrees from $\mathcal{T}$ as described. The number of sequences of the subtrees ranges from three to eight. For the significance level, we obtained $\alpha = 0.41$. Site i and i' are then called correlated if the p-value is less equal than α

The resulting estimates of StarDep are shown in Figure 2.14B. The detected dependencies are in good agreement with the expected secondary structure of the tRNA. We detected 15 from 22 base pairs from the expected secondary structure. Moreover, we detected two structural elements that are related to the three dimensional structure of the tRNA (between positions 16–71; and positions 27–48; see also Gutell *et al.* (1992)) .

However, seven base pairs of the secondary structure were not detected. Two base pairs were not detected since the corresponding positions were constant.

2.4.3 Results of the Purine Riboswitch

Additionally, we investigated an alignment of 111 bacterial sequences (Graef *et al.*, 2005) that include a purine riboswitch (see appendix A.1). The sequences comprise 106 nucleotides where 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 Figure 2.15. After estimating the parameters of the substitution model t^S was estimated to be 1.54. Using this value, we found only one maximal clique with three sequences. As shown in Figures 2.2 and 2.3 this is not a sufficient number to estimate a neighborhood system. Thus StarDep could not be applied to this data.

2.5 Discussion

In this chapter, the simulation studies showed some problems that one has to be aware of when estimating a neighborhood system from a sequence alignment. We investigated the ability of the χ^2-test in detecting correlated sites depending on the number of sequences n and the branch length t_b of the star tree. In general, we conclude that for increasing values of n and t_b the number of detected true positives also increases (see Figures 2.2, 2.3, 2.1). whereas the number of false positives is decreasing. However, if t_b is small,

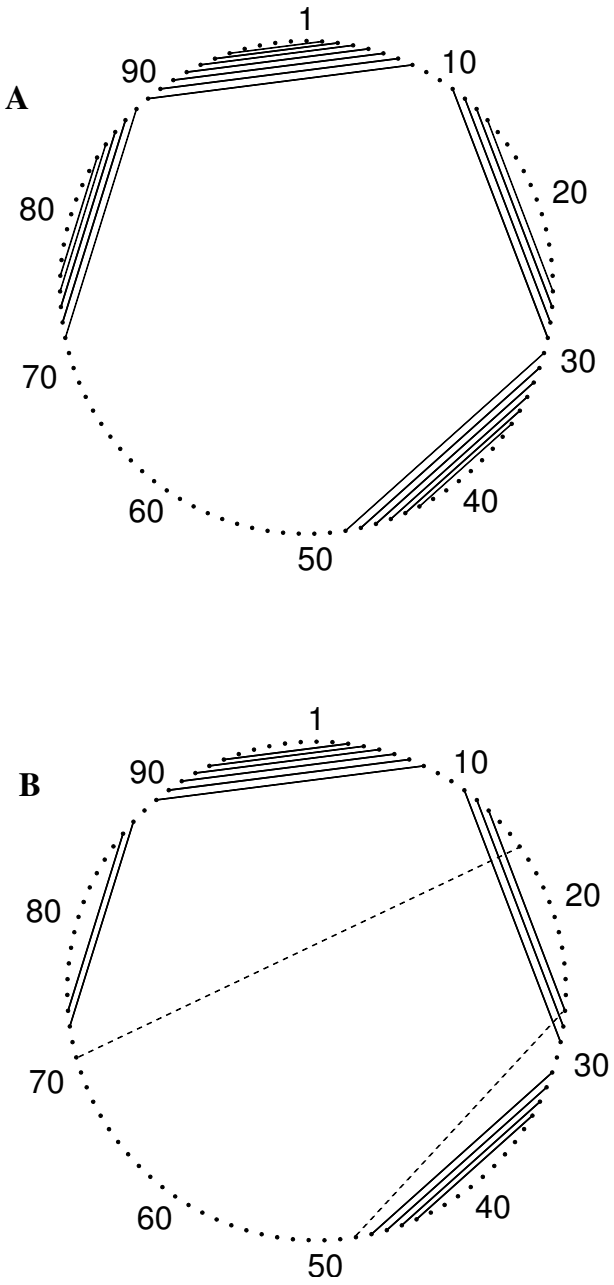

Figure 2.14: Circle plot of the tRNA
A: expected secondary structure of a tRNA sequence (Sprinzl *et al.*, 1998) **B**: estimated secondary structure using StarDep. Dashed lines represent tertiary structure elements (Gutell *et al.*, 1992).

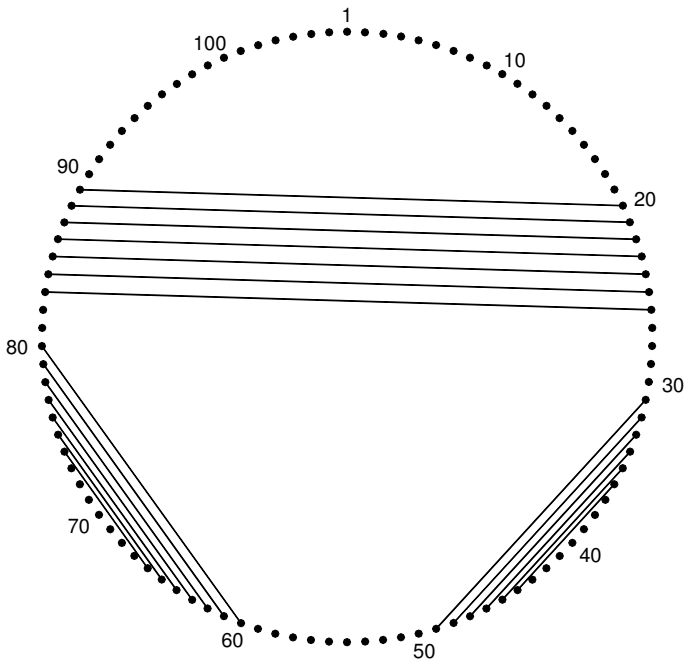

Figure 2.15: Secondary structure of the riboswitch alignment.

than it is difficult to detect dependencies even if n is large. For example, if $t_b = 0.2$ and $n = 1000$ only 40 percent of the true dependencies were detected. Although, our investigations are focused on star phylogenies, these conclusions are also true for non star phylogenies (see Table 2.2).

Moreover, we investigated the influence of ancestral correlations in detecting dependencies. We demonstrated that the disregard of the internal branching (ancestral correlation) of the phylogeny may lead to incorrect results by means of false positive correlated sites (Lapedes *et al.*, 1999, see also Figure 2.4).

In the second part of this chapter, we introduced StarDep, a method to predict a neighborhood system of a sequence alignment. For the analysis StarDep uses subtrees instead of the full phylogeny. We showed that sequences derived from these subtrees can be treated as independent sample and therefore the χ^2-test can be applied. Furthermore, we introduced

an heuristic to reduce false positive correlations. It is based on minimum p-values (Ge *et al.*, 2003). In simulation (Table 2.2) and the example of the tRNA (Figure 2.14), we showed that the accuracy can be improved by means of reducing false positive correlated sites.

The investigated subtrees rely on the estimation of t^S, the minimal genetic distance between pairs of sequences. If t^S is large compared to the pairwise genetic distances of the sequences StarDep cannot be applied as shown for the riboswitch alignment.

Chapter 3

Estimating Dependencies using Phylogenies

3.1 Introduction

In the previous chapter, we introduced StarDep. This method can be applied when genetic distances between pairs of sequences are large. Here, we introduce INFDEP (Inferring Dependencies) a method that allows statistical inference of correlated sites within a multiple sequence alignment where sequences evolved on a phylogeny. In contrast to StarDep, it includes the full phylogeny instead of subtrees in detecting the neighborhood system. INFDEP combines is a comparative method that includes an automated procedure to filter false positive correlations.

INFDEP is based on two summary statistics. The first statistics investigates pairs of sites and suggests potential correlations. The second statistics investigates the frequencies of nucleotides at a site and detects sites that cause false positive correlations. In section 3.2, we will explain the two test statistics. Subsequently, INFDEP is explained in more detail.

Based on simulated data we will evaluate the performance of the integrated approach. Finally, we apply the method to the alignment of the tRNA and the alignment comprising a purine riboswitch (Graef *et al.*, 2005).

3.2 INFDEP-Inferring Dependencies using phylogenetic Trees

First, we introduce some notations: With $\mathbf{D} = (\mathbf{D}_1, \ldots, \mathbf{D}_l)$ we denote a sequence alignment of length l with n sequences. That is, $\mathbf{D}_i$ $(i = 1, \ldots, l)$ denotes an n-dimensional pattern over the alphabet $\mathcal{A} = \{A, C, G, T\}$ of nucleotides. $\mathbf{D}_i$ represents the nucleotides at the ith site of the alignment for each of the n sequences. Thus, for n sequences 4^n patterns are possible.

With $\mathbf{D}_{ik}$ we denote the nucleotide at site i in sequence $k \quad (k = 1, \ldots, n)$. $\mathbf{D}$ constitutes the data we want to investigate.

We assume that the n sequences are related according to a phylogenetic tree T where the leaves represent the sequences in the alignment and the branch lengths of T reflect the amount of evolution. For the time being, we also assume, that this tree is rooted. The evolution of the nucleotides is then specified by a model of sequence evolution M (Tavaré, 1986; Rodriguez *et al.*, 1990) consisting of a rate matrix and a stationary distribution. The rate matrix typically belongs to the class of general time reversible models with stationary distribution $\boldsymbol{\pi} = (\pi_x)_{x \in \mathcal{A}}$. 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 $\boldsymbol{\theta}$ (generally scaled in expected substitutions per site) of the tree and the nucleotide $(R = u)$ at the root of the tree. Following standard computations and the assumption of independently

and identically distributed sites we can then compute the probability to observe alignment $\mathbf{D}$ (Felsenstein, 2004). To reduce the notational burden, we denote by

$$\mathbb{P}(\mathbf{p}|u) \equiv \mathbb{P}(\mathbf{p}, T, \boldsymbol{\theta}, M|R = u) \quad \text{for} \quad \mathbf{p} \in \mathcal{A}^n \tag{3.1}$$

the probability to observe pattern $\mathbf{p} = (\mathrm{p}_k)_{k=1,2,\ldots,n}$, if nucleotide u is present at the root of the tree. Assuming the independence of sites, it follows immediately that the joined probability to observe the pair of patterns $\mathbf{p}, \mathbf{q}$ is given by

$$\mathbb{P}(\mathbf{pq}|uv) = \mathbb{P}(\mathbf{p}|u)\mathbb{P}(\mathbf{q}|v). \tag{3.2}$$

Thus $(\mathbf{pq}) \in \mathcal{A}^n \times \mathcal{A}^n = \mathcal{A}^{2n}$, whereas $(uv) \in \mathcal{A}^2$. Furthermore, we denote with $\mathbf{n}_1(\mathbf{p}) = (n(x, \mathbf{p}))_{x\in\mathcal{A}}$ the base composition of pattern $\mathbf{p}$ and with $\mathbf{n}_2(\mathbf{pq}) = (n(xy, \mathbf{pq}))_{x,y\in\mathcal{A}}$ the contingency table of the patterns $\mathbf{p}$ and $\mathbf{q}$, where

$$n(x, \mathbf{p}) \equiv \sum_{k=1}^{n} \mathbb{1}(\mathrm{p}_k = x),\ x \in \mathcal{A} \tag{3.3}$$

$$n(xy, \mathbf{pq}) \equiv \sum_{k=1}^{n} \mathbb{1}(\mathrm{p}_k = x, \mathrm{q}_k = y),\ x, y \in \mathcal{A}. \tag{3.4}$$

The indicator function $\mathbb{1}(z)$ equals one if the argument z is true and is zero otherwise. That is to say, $\mathbf{n}_1(\mathbf{p})$ counts the number of times the letters A, C, G, T occur in pattern $\mathbf{p}$, while $\mathbf{n}_2(\mathbf{pq})$ counts the number of times a pair of nucleotides occurs. The expectation $\mathbb{N}_d(b)$ is given by

$$\mathbb{N}_d(b) = \sum_{\mathbf{a}\in\mathcal{A}^{nd}} \mathbb{P}(\mathbf{a}|b)\mathbf{n}_d(\mathbf{a}), \text{ where } \begin{cases} b \in \mathcal{A} & \text{if } d = 1 \\ b \in \mathcal{A}^2 & \text{if } d = 2\ . \end{cases} \tag{3.5}$$

$\mathbb{N}_1(b)$ is the nucleotide composition we expect conditional on the tree and the root, whereas $\mathbb{N}_2(b)$ is the expected composition of nucleotide pairs respectively. Thus, $\mathbb{N}_d(b)$ may be substantially different from the stationary

distribution. To measure the deviation, we define for an arbitrary pattern $\mathbf{a}$ either in $\mathcal{A}^n$ or $\mathcal{A}^{2n}$ and a fixed root assignment b a χ^2-type distance:

$$\Delta_d(\mathbf{a}|b) = \sum_{x \in \mathcal{A}^d} \frac{\left(\mathbb{N}_d(x|b) - n_d(x, \mathbf{a})\right)^2}{\mathbb{N}_d(x|b)} \qquad \text{for } d = 1, 2. \tag{3.6}$$

The collection of $\Delta_d(\mathbf{a}|b)$-values for every $\mathbf{a} \in \mathcal{A}^{nd}$ characterizes sequence evolution under independence. Therefore, we use functions $\Delta_d(\mathbf{a}|b)$ as a statistic to test the null-hypothesis of independently evolving sites. To this end, we need to determine the distribution of the $\Delta_d(\mathbf{a}|b)$ for each $b \in \mathcal{A}^d$. Since an analytical formula of the χ^2-type distributions seems not feasible, we use Monte Carlo simulations to approximate Δ_d. Thus, we simulate the evolution of m nucleotide patterns along the phylogeny T with respect to the root nucleotide. The expected nucleotide composition (Equation 3.5) is then approximated by $\mathbb{N}_d(b) \approx \frac{1}{m} \sum_{\mathbf{a}=1}^{m} \mathbf{n}_d(\mathbf{a})$ and the Δ_ds are computed according to Equation (3.6). Thus, we get an approximation of the null-distribution of $\Delta_d(\mathbf{a}|b)$ for each b. That is, if $d = 1$ we get four approximated distributions and for $d = 2$ we get 16 distributions. The p-value of the actually observed data $\Delta_d(\mathbf{D}_i|b)$ is then estimated by the proportion of simulated $\Delta_d(\mathbf{a}|b)$-values equal to or larger than $\Delta_d(\mathbf{D}_i|b)$ for any fixed b and $i = 1, 2, \ldots, l$. Thus, we obtain for the nucleotide pattern $\mathbf{D}_i$ at position i four p-values $\mathbb{P}(\mathbf{D}_i|R = u)$ one for each nucleotide at the root, and 16 p-values for the pair of positions $\mathbb{P}(\mathbf{D}_i\mathbf{D}_j|R = uv)$.

3.2.1 The EPWD test – Estimating Pairwise Dependencies

To classify alignment positions $\mathbf{D}_i$ and $\mathbf{D}_j$ as correlated, we require that the null-hypotheses of independently evolving sites is rejected for the 16 possible root assignments on significance level α. That is to say, if we assign at the

root of $\mathbf{D}_i$ the nucleotide $R_i = u$ and at the corresponding root of $\mathbf{D}_j$ the nucleotide $R_j = v$, then the p-values $\mathbb{P}(\mathbf{D}_i\mathbf{D}_j|R_{ij} = uv)$ have to be smaller than α for all assignments of root nucleotides $u, v \in \mathcal{A}$, in other words:

$$\max_{(u,v)\in\mathcal{A}^2}\{\mathbb{P}(\mathbf{D}_i\mathbf{D}_j|R_{ij} = uv)\} < \alpha. \tag{3.7}$$

We call $\mathbf{D}_i$ and $\mathbf{D}_j$ correlated if inequality 3.7 is true. Inequality 3.7 is based on the idea that only one $\mathbb{P}(\mathbf{D}_i\mathbf{D}_j|R_{ij} = uv) \geq \alpha$ suffices to retain the null-hypothesis, i.e. explains co-occurrence of both patterns by means of independent evolution. The collection of correlated sites for alignment $\mathbf{D}$ and a specified α is denoted by

$$\mathcal{C}_2^\alpha(\mathbf{D}) = \{(i, j)|\mathbf{D}_i, \mathbf{D}_j \text{ (fulfill inequality 3.7)}\}. \tag{3.8}$$

The set

$$\mathcal{C}_1^\alpha(\mathbf{D}) = \{i|\mathbf{D}_i \in \mathcal{C}_2^\alpha(\mathbf{D})\} \tag{3.9}$$

contains all alignment sites that appear to be correlated. We call this test EPWD (estimating pairwise dependencies). Note that $\mathcal{C}_1^\alpha(\mathbf{D})$ and $\mathcal{C}_2^\alpha(\mathbf{D})$ can be visualized in a circle plot graph, where $\mathcal{C}_1^\alpha(\mathbf{D})$ represents the nodes of the graph and $\mathcal{C}_2^\alpha(\mathbf{D})$ defines the edges (Figure 2.11).

In a nutshell: EPWD describes a contingency test taking the tree T and the branch lengths $\boldsymbol{\theta}$ 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 pairwise correlations. To this end, we introduce a second test.

3.2.2 The PWA test – Positions without Ancestry

Here we measure the base composition at an alignment position and ask for the probability that a given base composition deviates from the expected nu-

cleotide composition. With $\{\Delta_1(\mathbf{p}|u)\}_{\mathbf{p}\in\mathcal{A}^n}$ $u \in \mathcal{A}$, we denote the distribution of the χ^2-type distances (see Equation 3.6). Consider the four distributions $\Delta_1(\mathbf{p}|A)$, $\Delta_1(\mathbf{p}|C)$, $\Delta_1(\mathbf{p}|G)$, $\Delta_1(\mathbf{p}|T)$. For a pattern $\mathbf{D}_i$ from the data, we can compute the p-values for each distribution. That is, we compute the proportion of $\{\Delta_1(\mathbf{p}|u)\}$, $u \in \mathcal{A}$ that is larger than $\{\Delta_1(\mathbf{D}_i|u)\}$. Intuitively one would 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 $\sigma(\mathbf{p})$ for the four p-values $\mathbb{P}(\mathbf{D}_i|u)$. If $\sigma(\mathbf{D}_i)$ is small, then the information about the ancestral nucleotide state is lost or not present.

To estimate the p-value for $\sigma(\mathbf{D}_i)$ we determine the empirical distribution of $\sigma(\mathbf{p})$. That is, we draw m nucleotide pattern (typically $m = 1000 - 10,000$) from the distribution $\mathbb{P}(\mathbf{p}) = \sum_{u\in\mathcal{A}} \pi_u \mathbb{P}(\mathbf{p}|u)$. For each pattern $\mathbf{p}$ the four p-values are computed according to Equation 3.6 ($\Delta_1(\mathbf{p}|u)$ $u \in \mathcal{A}$). Subsequently the corresponding standard deviation is computed. Finally the p-value $\mathbb{P}(\sigma(\mathbf{D}_i))$ is estimated as

$$\mathbb{P}(\sigma(\mathbf{D}_i)) = \frac{|\{\sigma(\mathbf{p})|\sigma(\mathbf{p}) < \sigma(\mathbf{D}_i)\}|}{m}. \tag{3.10}$$

If $\mathbb{P}(\sigma(\mathbf{D}_i)) < \beta$ then the pattern $\mathbf{D}_i$ at site i is regarded as false positive site. Site i is then deleted from $\mathcal{C}_1^\alpha(\mathbf{D})$ and all pairs $(i,j) \in \mathcal{C}_2^\alpha(\mathbf{D})$ are called false positive correlations and thus are ignored. Finally, $\mathcal{C}_2^{\beta(\alpha)}(\mathbf{D})$ denotes the set of correlated pairs that are retained given α and β.

Thus, the PWA test detects sites that are rejected according to the null hypothesis of independently evolving nucleotides starting with a certain root nucleotide. We want to emphasize that the exclusion of a pattern strongly depends on the tree topology. For example on a phylogeny with generally long branches it is unlikely to observe a pattern where all sequences have

the same nucleotide. Consequently, the PWA test would reject this site. In contrast the same pattern would probably be kept when all sequences are closely related and consequently the probability of observing constant sites is higher.

3.2.3 The INFDEP Method (Inferring Dependencies)

Now we are ready to explain our strategy to determine the collection of correlated sites more precisely. We denote with correlated sites the outcome of INFDEP, whereas the true dependencies (from the multiple sequence alignment) are called dependent sites. The objective of INFDEP is that the number of correlated sites equals the number of dependent sites.

INFDEP starts with the estimation of the phylogenetic tree $\widehat{T}$ and the parameters for the nucleotide substitution model $\widehat{M}$ from the alignment $\mathbf{D}_{data}$. The substitution model $\widehat{M}$ comprises as parameters the base frequencies and the ratio of transitions and transversions. Based on $(\widehat{M}, \widehat{T})$, we generate an alignment $\mathbf{D}^0$ under independence using Seq-Gen (Rambaut and Grassly, 1997). From this alignment the distributions Δ_1 and Δ_2 are estimated according to Equation 3.6.

$\mathbf{D}^0$ constitutes an alignment of independently evolving sites, thus $\mathcal{C}_2^{\alpha}(\mathbf{D}^0)$ should be empty. However, the EPWD-test yields a set of (false positive) pairwise correlations $\mathcal{C}_2^{\alpha}(\mathbf{D}^0)$ for a given α.

Now, we apply the PWA-test to adjust $\beta(\alpha)$ such that $\mathcal{C}_2^{\beta(\alpha)}(\mathbf{D}^0) = \emptyset$. This value is denoted by $\beta_{\emptyset}(\alpha)$. This procedure is repeated for "every" α, $(0 < \alpha < 1)$.

Finally, we obtain a collection of $(\alpha, \beta_{\emptyset}(\alpha))$ pairs, that serve as "selector pairs" to determine correlated sites in biological data $\mathbf{D}_{data}$. The set $\mathcal{C}_2^{\beta_{\emptyset}(\alpha)}(\mathbf{D}_{data})$ comprises the collection of site-pairs (i, j) that could not be

rejected for the given "selector pair". Thus, $\mathcal{C}_2^{\beta_\emptyset(\alpha)}(\mathbf{D}_{data})$ contains the correlated sites. In a typical application, we start with small values of α, adjust $\beta_\emptyset(\alpha)$ accordingly and compute the number of correlated sites. Then we increase α gradually, adjust $\beta_\emptyset(\alpha)$, and again compute the number of correlated sites. This is repeated until no new correlated sites are found. The union $\mathcal{C}_2^{\beta_\emptyset(\alpha_1)}(\mathbf{D}_{data}) \cup \mathcal{C}_2^{\beta_\emptyset(\alpha_2)}(\mathbf{D}_{data}) \ldots$ then constitutes our collection of correlated sites.

The sensitivity to detect dependent sites can be further increased when correlated pairs are removed from the alignment. Subsequently, for the resulting shortened alignment a new phylogeny is reconstructed. Then INFDEP is again applied as explained. The renewed tree reconstruction is necessary since the removal of the correlated pairs may substantially change the topology as well as the estimated parameters of the substitution model.

3.3 Application

3.3.1 Performance of INFDEP on Synthetic Data

We evaluated the ability of INFDEP to detect the dependencies of a RNA-molecule given a multiple sequence alignment. To this end we carried out a simulation. We assumed the secondary structure of an artificial molecule as displayed in Figure 1. The molecule is 200 bases long and contains seven base paired regions (I-VII), where region VII represents a pseudo-knot. The base paired regions (54 base pairs) evolve according to the doublet substitution model (Schöniger and von Haeseler, 1994) and the 92 remaining sites evolve according to the HKY model (Hasegawa *et al.*, 1985). This molecule evolved on a phylogenetic tree with 100 leaves using SISSI (Gesell and von Haeseler, 2006).

The result of such a simulation, $\mathbf{D}_{data}$, is then subject to a further analysis. We performed all steps outlined in INFDEP. From $\mathbf{D}_{data}$ the phylogenetic tree $\widehat{T}$ and the parameter for the nucleotide substitution model $\widehat{M}$ (Hasegawa *et al.*, 1985) were inferred using IQPNNI (Vinh and von Haeseler, 2004). Based on $(\widehat{M}, \widehat{T})$, the alignment $\mathbf{D}^0$ (length 1000) and the distributions Δ_1, Δ_2 were generated under the assumption of independent sites using Seq-Gen (Rambaut and Grassly, 1997). Figure 3.1 displays the 16 $\Delta_2(\mathbf{pq}, uv)$ distributions according to Equation 3.6. These distributions result from the independent evolution of pairs of sites assuming that at the root of $\widehat{T}$ are the nucleotides u and v. Each distribution is based on 2,000 generated dinucleotide patterns that evolved on tree $\widehat{T}$. Then for $\alpha = 0.01, 0.05, 0.1, \ldots, 0.45$ the corresponding $\beta_\emptyset(\alpha)$ is determined, by first computing the set of potentially correlated sites $\mathcal{C}_2^\alpha(\mathbf{D}^0)$ and subsequently $\mathcal{C}_1^\alpha(\mathbf{D}^0)$. $\beta_\emptyset(\alpha)$ is then the minimal p-value $\mathbb{P}(\sigma(\mathbf{D}_i))$ (see Inequality 3.10) of site $i \in \mathcal{C}_1^\alpha(\mathbf{D}^0)$ where the set of correlated sites $\mathcal{C}_2^{\beta_\emptyset(\alpha)} = \emptyset$.

To estimate the number of correlated sites from $\mathbf{D}_{data}$ we compute the set of potentially correlated sites $\mathcal{C}_2^\alpha(\mathbf{D}_{data})$. In Equation 3.10 we set $\beta = \beta_\emptyset(\alpha)$ to detect sites that cause false positive correlations and end up with the set of correlated sites $\mathcal{C}_2^{\beta_\emptyset(\alpha)}(\mathbf{D}_{data})$.

The results of INFDEP are summarized in Table 3.1. The left part of Table 3.1 displays the analysis of $\mathbf{D}^0$ that leads to the estimation of the "selector pairs" $(\alpha, \beta_\emptyset(\alpha))$. Since $\mathbf{D}^0$ constitutes an alignment of independently evolving sites, the number of correlated sites $|\mathcal{C}_2^\alpha(\mathbf{D}^0)|$ (Table 3.1 column 2) are in fact artifacts. Furthermore, we compute the percentage of correlated sites (column 3) with respect to the total number of pairs, i.e $\binom{1000}{2}$. The parameter $\beta_\emptyset(\alpha)$ (Table 3.1 column 4) balances the effect of being too liberal with the acceptance of correlated sites as outcome of the EPWD-test. Recall,

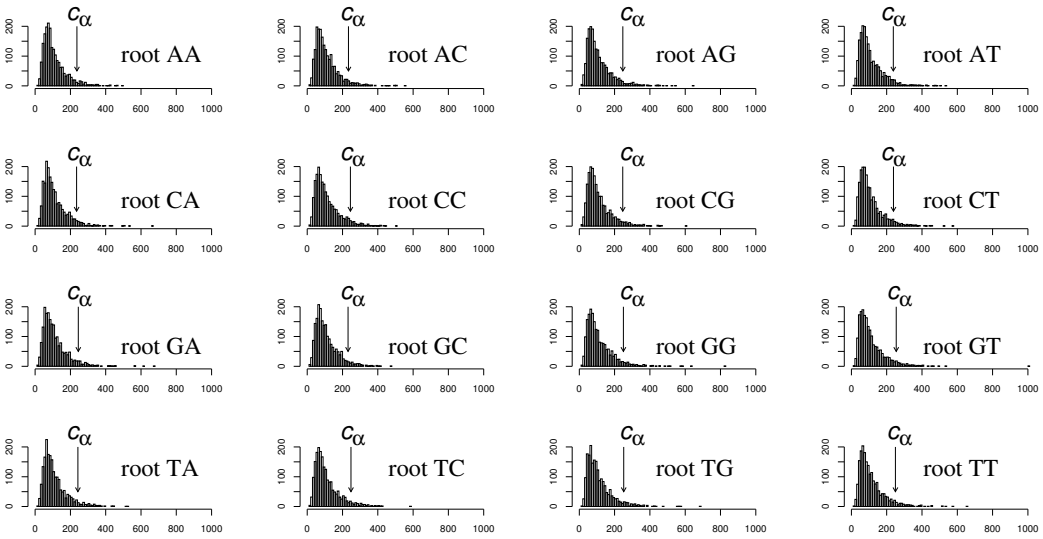

Figure 3.1: The Null Distributions of Δ_2 for all possible Root Nucleotides. Two positions $\mathbf{D}_i$ and $\mathbf{D}_j$ are called correlated if $\Delta_2(\mathbf{D}_i\mathbf{D}_j, uv)$ is larger than the critical value c_α for all Δ_2. The x-axis represents the Δ_2 values according to Equation 3.6

$\beta_\emptyset(\alpha)$ is the smallest value such that the number of "truly" correlated sites $\mathcal{C}_2^{\beta_\emptyset(\alpha)}(\mathbf{D}^0) = \emptyset$ (Table 3.1 column 4). To this end we notice that for increasing α the cardinality of potentially correlated sites and $\beta_\emptyset(\alpha)$ also increases.

The right part of Table 3.1 displays the analysis of the data $\mathbf{D}_{data}$. We also compute the number of potentially correlated sites (column 6) and subsequently the corresponding percentage with respect to the total number of possible correlations $\binom{200}{2}$ (column 7). We use the estimated selector pairs $(\alpha, \beta_\emptyset(\alpha))$ to detect the number of correlated sites $\mathcal{C}_2^{\beta_\emptyset(\alpha)}(\mathbf{D}_{data})$. The second last column in Table 3.1 illustrates that the number of correlated sites in $\mathcal{C}_2^{\beta_\emptyset(\alpha)}(\mathbf{D}_{data})$ increases with α up to a maximum of 22 true pairs for $\alpha = 0.1$. For larger α the number of elements declines until $\mathcal{C}_2^{\beta_\emptyset(\alpha)}(\mathbf{D}_{data})$ is empty. The last column shows the cumulative number of correlated sites as α increases.

We conclude the analysis of $\mathbf{D}_{data}$ with a total number of 46 correlated

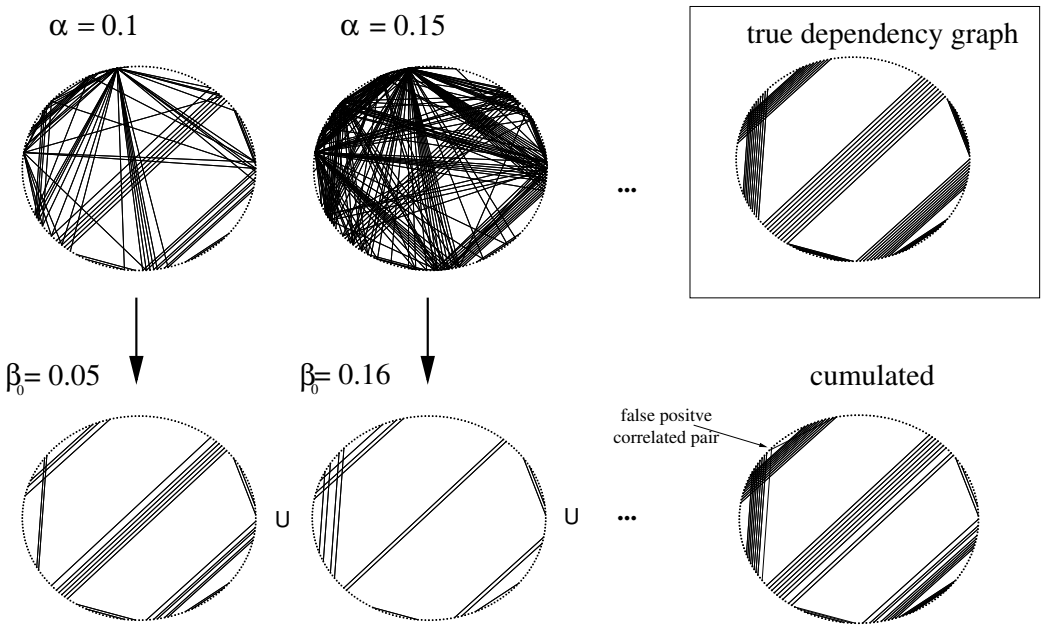

Figure 3.2: Dependency Graphs of Simulated Data.
In the first row are shown the potentially correlated pairs $\mathcal{C}_2^{\alpha}$ exemplary for two choices of $\alpha = 0.1, 0.15$ (EPWD-test). The second row displays the finally obtained dependencies $\mathcal{C}_2^{\beta_\emptyset(\alpha)}$ after applying the PWA test. The cumulated graph is obtained after superimposing all dependency graphs.

sites compared to 54 base pairs. Figure 3.2 visualizes the INFDEP method. The top row displays the potentially correlated sites exemplary for $\alpha = 0.1$ and $\alpha = 0.15$ and the bottom row shows the "surviving" correlations after the PWA test is applied. Superimposing the different dependency graphs lead to the cumulated circle plot (right part).

Since we know the true dependency structure, we can compare it to the result of our analysis. From 54 originally dependent pairs, we detected 45 correlated sites and one false positive correlation (arrow in Figure 3.2).

The accuracy of the method is improved when correlated sites, already detected as being correlated, are excluded from the alignment. Subsequently INFDEP is applied to the reduced alignment. We obtained four more corre-

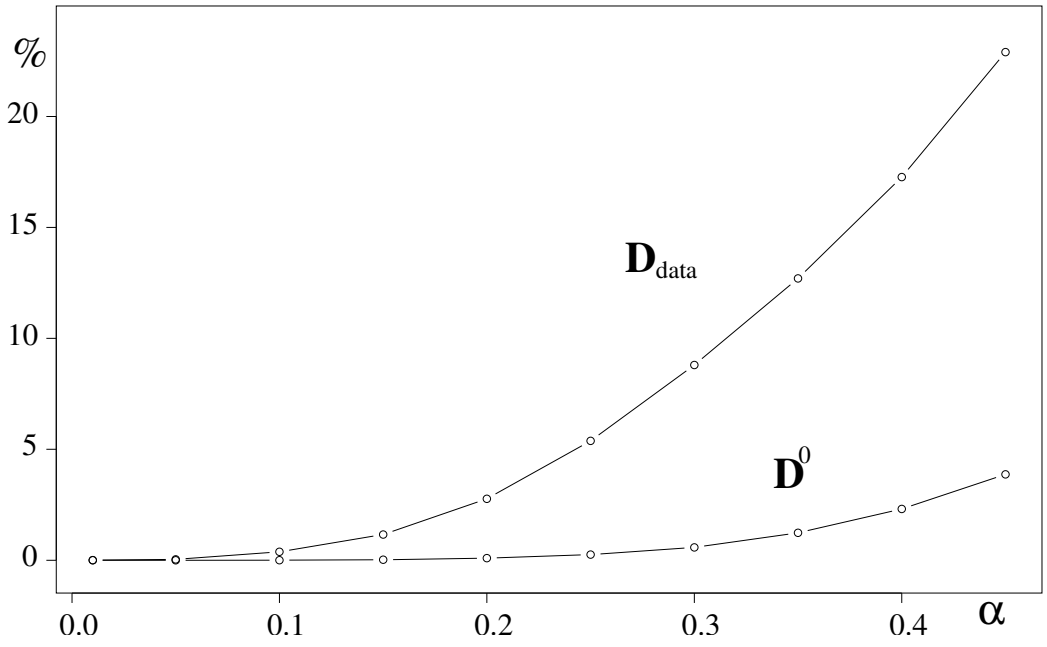

Figure 3.3: Percentage of the Potentially Correlated Pairs.
Displayed is the set $\mathcal{C}_2^\alpha$ with respect to the total number of pairs $\binom{l}{2}$ as function of α (for $\mathbf{D}^0$ ($l = 1000$) and $\mathbf{D}_{data}$ ($l = 200$))

lations in the $\mathbf{D}_{data}$ alignment, resulting in a total of 50 correlated sites (49 true positives and one false positive).

Interestingly, the number of potentially correlated sites $|\mathcal{C}_2^\alpha|$ shows with increasing α a much higher increase for $\mathbf{D}_{data}$ than $\mathbf{D}^0$ (Figure 3.3). This difference in accumulating potentially correlated sites may already indicate that dependencies between sites are present.

$\mathbf{D}^0$					data $\mathbf{D}_{data}$			
α	$\lvert\mathcal{C}_2^{\alpha}(\mathbf{D}^0)\rvert$	perc. $\mathbf{D}^0$	$\beta_{\emptyset}(\alpha)$	$\lvert\mathcal{C}_2^{\alpha,\beta_{\emptyset}}(\mathbf{D}^0)\rvert$	$\lvert\mathcal{C}_2^{\alpha}(\mathbf{D}_{data})\rvert$	perc. $\mathbf{D}_{data}$	$\lvert\mathcal{C}_2^{\alpha,\beta_{\emptyset}}(\mathbf{D}_{data})\rvert$	$\lvert\bigcup\mathcal{C}_2^{\alpha,\beta_{\emptyset}}(\mathbf{D}_{data})\rvert$
0.01	0	0	0.0	0	1	0.005	1	1
0.05	0	0	0.0	0	8	0.04	8	8
0.1	13	0.003	0.05	0	76	0.38	22	27
0.15	114	0.023	0.16	0	232	1.16	15	36
0.2	469	0.094	0.18	0	552	2.77	18	42
0.25	1278	0.256	0.37	0	1072	5.38	6	44
0.3	2887	0.578	0.48	0	1752	8.80	5	46
0.35	6180	1.237	0.67	0	2528	12.70	2	46
0.4	11544	2.311	0.71	0	3437	17.27	2	46
0.45	19338	3.871	0.8	0	4558	22.90	0	46

Table 3.1: Results of INFDEP obtained from the Alignments $\mathbf{D}^0$ (sites evolved independently) and $\mathbf{D}_{data}$ (containing dependencies).
$\lvert\mathcal{C}_2^{\alpha}(\mathbf{D}^0)\rvert$ and $\lvert\mathcal{C}_2^{\alpha}(\mathbf{D}_{data})\rvert$: are the number of potentially correlated pairs after applying the EPWD test for the corresponding significance value α. perc. $\mathbf{D}^0$ and perc. $\mathbf{D}_{data}$ gives the percentage of the potentially correlated pairs with respect to the total number of possible dependencies $\binom{l}{2}$ (for $\mathbf{D}_{data}$ $l = 200$; for $\mathbf{D}_0$ $l = 1000$). $\beta_{\emptyset}(\alpha)$ is adjusted that the number of true dependencies $\lvert\mathcal{C}_2^{\alpha,\beta_{\emptyset}}(\mathbf{D}_0)\rvert$ equals zero Finally $\lvert\mathcal{C}_2^{\alpha,\beta_{\emptyset}}(\mathbf{D}_{data})\rvert$ equals the number of true dependencies and $\lvert\bigcup\mathcal{C}_2^{\alpha,\beta_{\emptyset}}(\mathbf{D}_{data})\rvert$ is cumulated number of true dependencies.

3.3.2 Influence of Tree Topology

We test the influence of the underlying tree on our capability to detect the dependency structure of the alignment. To this end, we investigated the RNA molecule with the secondary structure from the previous section (Figure 2.11), i.e. 54 base pairs and 92 independently evolving sites.

Our analysis is based on six bifurcating trees with same topology but different mean branch length. The topology was generated using the ape package (Paradis *et al.*, 2004) that is included in the R environment (R Development Core Team, 2004). The branches of the tree were randomly drawn from a uniform distribution. Finally, the branches are rescaled resulting in six different trees $T_{0.05}, T_{0.1}, T_{0.2}, T_{0.3}, T_{0.4}, T_{0.5}$ with mean branch length $0.05, 0.1, 0.2, 0.3, 0.4, 0.5$ respectively.

To assess the sensitivity of INFDEP, we simulated 100 data sets for each tree using the doublet model (Schöniger and von Haeseler, 1994) for the base paired regions and the HKY model (Hasegawa *et al.*, 1985) for independently evolving sites. Each data set was then analyzed by INFDEP. Thereafter we count the number of true positive correlated sites and the number of false positive correlated sites.

The results are summarized in the box plot in Figure 3.4. For alignments derived from the tree with the shortest mean branch length 0.05 the number of inferred true positive correlated sites ranges from zero to three with median zero. For alignments derived from trees with larger mean branch length the number of true positives increases. Exemplary, for tree $T_{0.4}$ the median is 48 (range 41–54) and for $T_{0.5}$ the median is 47 (range 40–54).

A different result emerges for the number of detected false positive sites. Here we observe no obvious trend depending on the branch length, since the median of the false positives is for all of the trees between one and two (see also Figure 3.4).

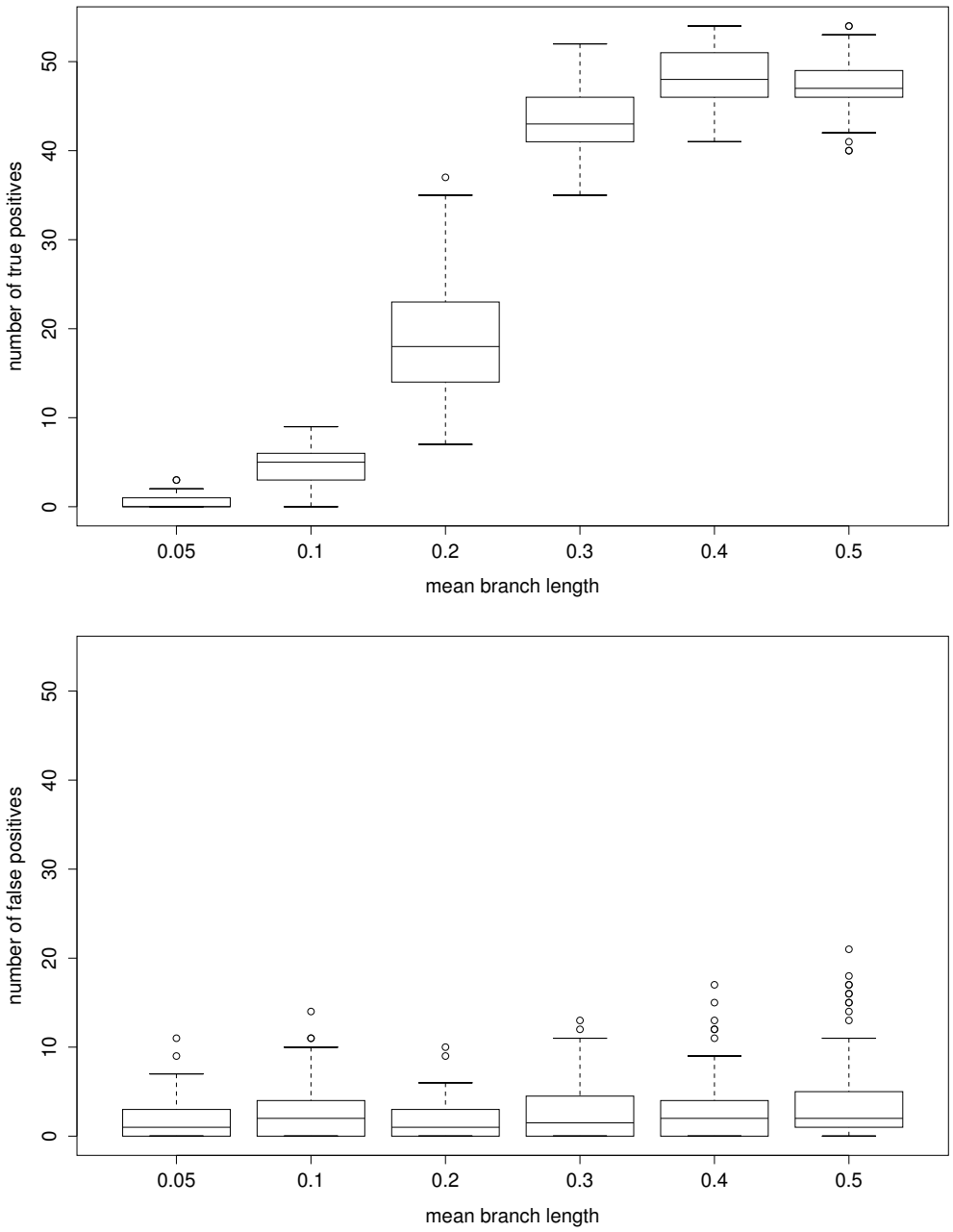

Figure 3.4: Number of Detected True Positive and False Positive Correlations vs Mean Branch Length.

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.

General conclusions are difficult because INFDEP strongly depends on the total branch length of the underlying tree. The sensitivity of INFDEP increases with increasing mean branch length from zero for tree $T_{0.05}$ to about 90 % for trees $T_{0.4}$ and $T_{0.5}$, whereas the number false positives is small for all trees. Moreover in 20%-41% of the simulated data no false positives were observed.

The relation between the number of detected false positives and the branch length can be explained by a lack of power of INFDEP. Assuming the case of a tree with zero branch length then no statistical method has the ability of detecting correlated sites, since no substitution occurred. With increasing branch length of the tree the number of substitution of independently evolving sites as well as substitutions between correlated sites accumulate. This accumulation of different substitution patterns allows a better detection of correlated sites.

3.3.3 Results of the tRNA Alignment

We apply INFDEP to the tRNA sequences alignment (Sprinzl *et al.*, 1998, see also section 2.4.2). For the analysis, sites that had more than 90 % gaps were excluded. The proposed secondary structure is shown in Figure 3.5A. The resulting estimates of the pairwise dependency structure is shown in Figure 3.5B. The dependencies finally obtained are in good agreement with the expected tRNA secondary structure. We obtain 13 base pairs from the secondary structure and additionally a tertiary structure element. However, compared to StarDep, INFDEP detects two base pairs less for the secondary structure and one base pair less for the tertiary structure. After excluding sites that were base paired, we repeat INFDEP with the reduced alignment, but no new dependencies were detected.

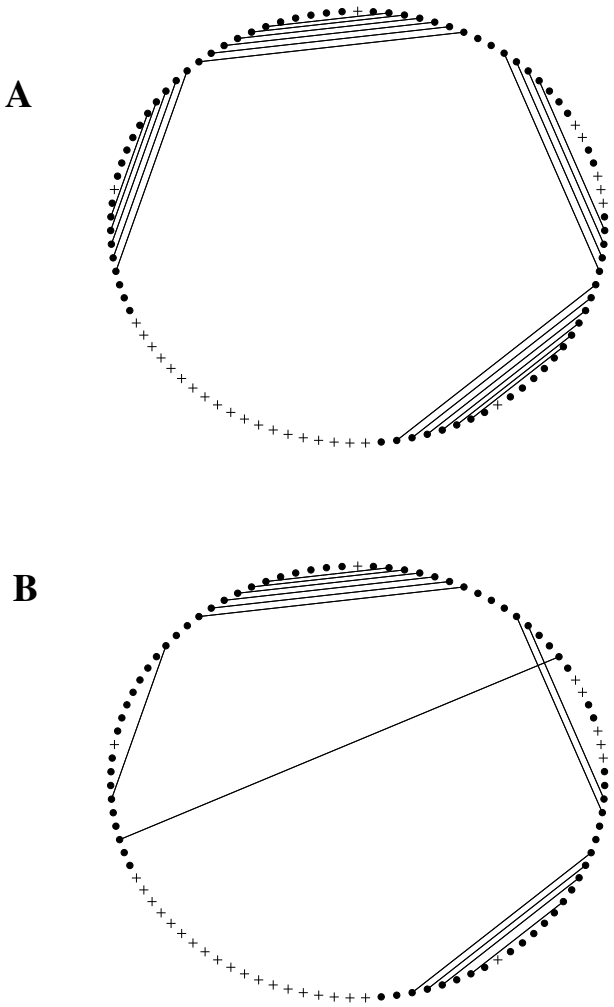

Figure 3.5: Circle plot of the tRNA
A: expected secondary structure **B**: estimated secondary structure using INFDEP. Lines in the circle plot represent base pairs of nucleotides. Excluded positions are marked with crosses.

3.3.4 Results of the Purine Riboswitch

We apply INFDEP to the sequence alignment that include a purine riboswitch (Graef *et al.*, 2005, see also section 2.4.3). 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 Figure 3.6A.

Based on the alignment, we performed all steps outlined in methods (section 3.2.3). For the full alignment, INFDEP detected nine correlated pairs that are shown as continuous lines in Figure 3.6B.

Subsequently, the corresponding 18 sites were excluded from the alignment. For the resulting reduced alignment we re-applied INFDEP. Recall that for the reduced alignment the tree is reconstructed again. Although the differences in the total branch length of the reconstructed trees $\widehat{T}_{full}$ and $\widehat{T}_{reduced}$ are relative small (21.2 and 20.7 resp.) there are relatively large differences in parameters of the substitution model. That is, the estimated parameters for the HKY model (base frequencies, transition transversion ratio) differ by 10–20 percent between the full and reduced alignment.

In the reduced alignment we detected four additional dependencies (dashed lines in Figure 3.6). Repeating INFDEP for an again reduced alignment did not lead to new correlated pairs.

The resulting circle plot of the estimated pairwise dependency structure is in good agreement with the secondary structure of the *Bacillus subtilis* (Batey *et al.*, 2004). We obtain 13 from 20 base pairs. No false positive base pair was suggested. However, seven base pairs were not detected. Four of the seven base pairs were not found because at least one of the two sites are conserved. For example the dependency between sites 25 and 84 is present in the secondary structure. These sites form a Watson Crick base pair where

at site 25 we observe in all 111 sequences an Uracil and at site 84 always an Adenine. However, a constant site is unlikely under the null hypothesis of independently evolving sites given the tree and the substitution model. These sites are rejected by the PWA test. The remaining three base pairs were not detected, since the p-value $\mathbb{P}(\sigma(\mathbf{D}_{full}))$ was below the corresponding $\beta_\emptyset(\alpha)$.

3.4 Discussion

We introduced INFDEP as a method to detect correlated sites from a sequence alignment. In contrast to (Gutell *et al.*, 1992; Klingler and Brutlag, 1993; Tabaska *et al.*, 1998) we also include the phylogeny of the sequences into the analysis. Moreover, no prior knowledge about the secondary structure of the molecule is needed.

INFDEP introduces the selector pairs $(\alpha, \beta_\emptyset(\alpha))$ that are derived from the EPWD test and the PWA test, respectively. That is, the EPWD test suggests correlated site pairs for a given significance value α, whereas the PWA test rejects sites with significance value $\beta_\emptyset(\alpha)$. Moreover, we vary α between zero and one and use not one fixed α as in classical test theory. The advantage of INFDEP is that it is self-consistent, i.e. no threshold is needed to be set in advance to assess significance.

The fundamental part of INFDEP is the PWA test. This heuristic enables the detection of sites that cause false positive correlations. However, the applied statistics especially the standard deviations of the p-values in the PWA test are not common use and need to be investigated in more detail.

Besides, the PWA test may be too liberal in rejecting sites, especially when dependent sites are constant. This is the case for some sites in the riboswitch alignment. This observation, however is not surprising. If a tree has

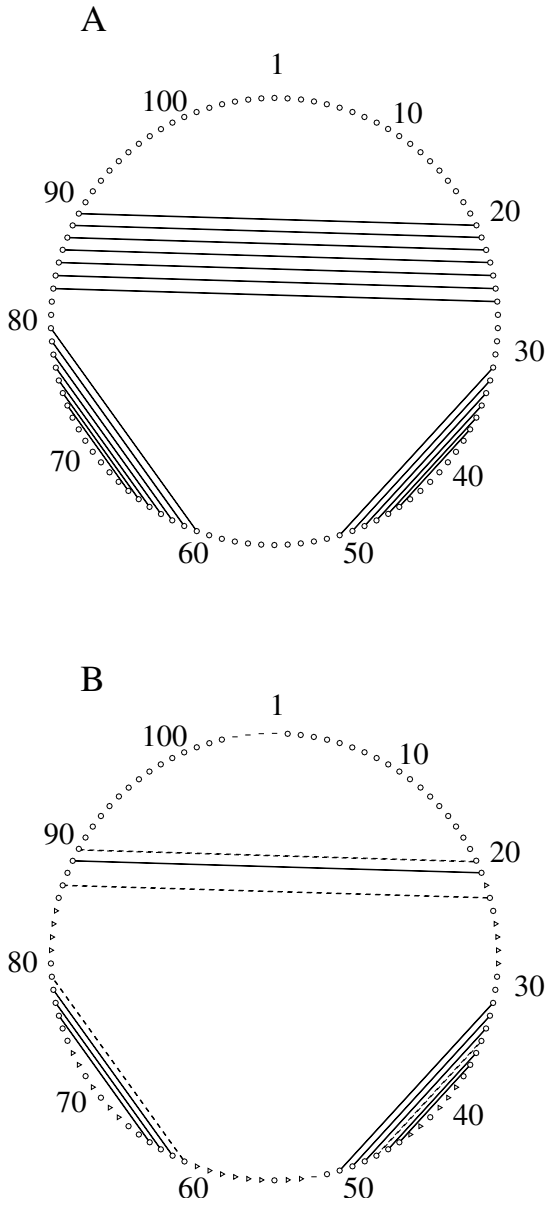

Figure 3.6: Dependency Graphs of Riboswitch Sequences.

A: The secondary structure of the riboswitch of *bacillus subtilis* (Batey *et al.*, 2004).

B: The estimated dependency structure. Sites indicated by a dash contain more than 50 gaps. Triangles display sites that are to 90 % conserved. Straight lines are detected correlated sites using the full alignment $\mathbf{D}_{full}$. Dashed lines are correlations using the reduced alignment $\mathbf{D}_{reduced}$ (see text for details).

total branch length zero, then all sequences in an alignment are identical and a test for correlations of pairs of sites is not applicable. Only if some variability of dependent sites is observed all comparative tests have the chance to suggest correlations. In such cases the dependency is easily detected by visual inspection. Hence, we recommend to investigate the potentially correlated pairs found by the EPWD test in more detail especially when they are situated in base paired regions.

However, the simulations and the analysis of the tRNA alignment and the riboswitch alignment showed that we are able to infer the underlying dependency structure of a sequence simply from an alignment.

It should be noted, that the ability of INFDEP to detect correlations depends on the total branch length of the underlying tree. For the trees with short total branch length the detection is harder than for trees with large total branch length. This can be explained by few substitutions of base pairs that occur on trees with short total branch length. Thus the differences in the sequences are not sufficient to distinguish between the evolution of base pairs or independent sites.

We could show that the number of detected true positive correlated sites can be increased when correlated sites are excluded from the alignment. With the resulting reduced alignment INFDEP is repeated. Hence, we could increase the number of detected true positive correlations for the simulated data as well as for the riboswitch alignment.

For the evolution of the nucleotides, we used the HKY model but more general models can also be applied as well as rate heterogeneity.

So far our test is designed to detect dependencies between pairs of sites. In general, INFDEP could be used for higher order correlations which would correspond to the case $d > 2$. For higher order correlations one has to face the difficulty that the investigated alphabet considerably increases with $\mathcal{A}^d$.

Summary

The focus of this thesis was the statistical inference of structural elements within RNA sequences. Our analysis is based on comparative methods that incorporate phylogenies relating the RNA sequences.

In the first part of chapter 2 we elucidated some problems that arise when the phylogenies are not considered in the analysis. Therefore, we investigate alignments derived from star trees. We observed that the ability in detecting correlations depends on the number of sequences and the branch length of the tree. Furthermore, we investigated the influence of ancestral correlation caused by the internal branchings of bifurcating trees in detecting correlated sites. We showed that the number of false positives is drastically increased as compared to star phylogenies.

In the second part of chapter 2 we introduced a novel strategy called StarDep to detect pairwise correlations. StarDep is based on the analysis of subtrees of the full phylogeny. We could show that this method gives encouraging results for synthetic and real data. The limitation of StarDep is that it can be applied only when the genetic distance between sequences is large.

In chapter 3 we introduced INFDEP. This method allows detection of correlated site incorporating the full phylogeny. In simulation and on real data this method was able to detect the expected secondary structure. An

essential part of this work was the improvement of the accuracy by means of reduction false positive correlations, that were discussed in chapters 2 and 3.

In the direct comparison, StarDep performed better than INFDEP for the tRNA alignment. StarDep detects 17 true positive base pairs (15 of the secondary structure, two from the tertiary structure) and INFDEP detects 14 (13 secondary structure; one tertiary structure). However, INFDEP can be applied to any sequence alignment, whereas StarDep can only be applied to alignments where the pairwise genetic distance between sequences is large. This was shown for the riboswitch alignment, where StarDep could not be applied.

Appendix A

Parameter Settings and Data

A.1 Data

Purine Riboswitch Sequences

The sequences of the purine riboswitch can be obtained from the NCBI homepage (*http://www.ncbi.nlm.nih.gov/*). The accession numbers are given in the following table. The purine riboswitch is a subsequence of these sequences. The first and last number of the accession number is the start and end position of this subsequence.

NC_000964_625975-625913	NC_006510_282596-282660
NC_002570_806889-806949	NC_006510_274273-274337
NC_002662_1159525-1159585	NC_006510_272489-272553
NC_002973_617828-617764	NC_007530_260657-260721
NC_003030_1002192-1002253	NC_007530_262617-262681
NC_003030_2824935-2824875	NC_007530_295347-295405
NC_003030_2905032-2904968	NC_007530_1497710-1497768
NC_003098_1634841-1634899	NC_007530_342338-342274

NC_003366_422836-422900	NC_007530_3605407-3605343
NC_003366_2871183-2871121	l77246_1_1237-1334
NC_003366_2618403-2618343	ap001509_1_53309-53408
NC_003454_1645802-1645741	ap001512_1_93774-93675
NC_003909_382608-382548	u51115_1_15589-15688
NC_003923_410562-410627	d88802_1_12464-12366
NC_004193_786783-786846	ap001509_1_209873-209971
NC_004368_1163472-1163414	ap004595_1_169586-169684
NC_004461_2433029-2432964	ap004596_1_203843-20394
NC_004557_2551374-2551314	ap004595_1_186670-186768
NC_004567_2410494-2410555	ap004595_1_160373-160472
NC_004567_2968812-2968751	al596170_1_223345-223247
NC_004605_1369737-1369799	al596165_1_154156-154057
NC_004668_2288408-2288348	al591975_1_251119-251020
NC_004722_298790-298848	al591981_1_205922-205824
NC_004722_259617-259681	ap003359_2_80811-80910
NC_004722_343829-343765	ae016752_1_24569-24470
NC_005362_1949403-1949463	ae007775_1_3557-3458
NC_005363_3414621-3414681	ae007768_1_1788-1690
NC_006086_857680-857738	ae007602_1_8615-8714
NC_006274_265891-265955	ap003186_2_121422-121519
NC_006274_3685834-3685770	ap003186_2_211688-211589
ba000043_1_282580-282681	cp000002_2_4024215-4024313
ae017333_1_4024498-4024398	ae017333_1_2295789-2295694
ae017333_1_696847-696940	ae017333_1_692988-693082
d83026_1_18553-18454	ap003193_2_214121-214023

u51115_1_11655-11754	ap003194_2_163701-163603
ap001509_1_79475-79574	ae015944_1_141656-141558
j02732_1_196-295	ae013027_1_8237-8336
ap001509_1_51442-51541	ae016954_1_153893-153795
ae017024_1_260641-260743	ae006347_1_1212-1310
ae017265_1_94745-94643	af327738_1_2512-2607
ae017269_1_211313-211415	ae014241_1_16113-16017
ae016998_1_259601-259703	ae007476_1_6452-6548
ae016999_1_36564-36462	ae010036_1_1276-137
ae017265_1_8580-8682	ae010606_1_4680-4581
ae017010_1_138243-138141	bx842655_1_288908-289004
ae017265_1_48309-48411	ap005088_1_167671-167771
ae017265_1_6624-6726	ae016809_1_202495-202592
ae016998_1_298774-298876	ba000043_1_274257-274360
NC_006322_4024340-4024404	ae017002_1_301083-301185
NC_006322_696854-696918	z99107_2_14363-14263
NC_006322_692997-693061	z99107_2_86081-86183
NC_006322_2295770-2295708	z99115_2_111605-111505
NC_006371_1538878-1538816	z99123_2_194901-195003
NC_006448_1185097-1185155	z99107_2_82145-82247
NC_006449_1182964-1183022	ab008757_1_115-16
NC_003995_794076-794177	

tRNA sequences

The tRNA sequences were obtained from the tRNA compilation homepage (Sprinzl *et al.*, 1998): *http://www.staff.uni-bayreuth.de/ btc914/search/index.html* The accession numbers are given in the following table.

RA1140	RG1381	RL1540	RR1141	RV1660	RG1140
RA1180	RG1540	RL1660	RR1540	RV1661	RG1180
RA1540	RG1580	RL1661	RR1660	RV1662	RG1310
RA1660	RG1660	RL1662	RR1661	RV2120	RG1380
RA1661	RG1661	RL1700	RR1662	RW1140	RK1660
RA1662	RG1662	RL2020	RR1663	RW1141	RL1140
RC1140	RG1700	RL2100	RR1664	RW1250	RL1141
RC1660	RG1701	RL2101	RS1140	RW1251	RL1142
RD1140	RH1140	RL2120	RS1141	RW1540	RQ1140
RD1580	RH1660	RM1140	RS1180	RW1660	RQ1660
RD1660	RH1700	RM1540	RS1540	RX1140	RQ1661
RE1140	RI1140	RM1580	RS1541	RX1180	RR1140
RE1660	RI1141	RM1660	RS1542	RX1300	RT1661
RE1661	RI1180	RN1140	RS1660	RX1540	RV1140
RE1662	RI1540	RN1660	RS1661	RX1580	RV1180
RE2140	RI1580	RN1720	RS1662	RX1581	RV1540
RF1140	RI1660	RN1721	RS1663	RX1660	RY1660
RF1540	RI1661	RP1140	RS1664	RX1661	RY1661
RF1580	RI1662	RP1180	RT1140	RX2060	RY2120
RF1660	RK1140	RP1540	RT1141	RX2100	RZ1665
RF2020	RK1141	RP1700	RT1180	RY1140	
RF2060	RK1540	RP1701	RT1540	RY1540	
RF2120	RK1541	RP1702	RT1660	RY1541	

	A	C	G	U
A	0.003	0.0049	0.0042	0.1539
C	0.0049	0.0035	0.2508	0.0032
G	0.0042	0.2508	0.0018	0.0762
U	0.1539	0.0032	0.0762	0.0052

Table A.1: The dinucleotide distribution used for the SH-model (Equation 1.6), e.g. $\pi_{AU} = 0.1539$

A.2 Simulated Data

The alignments of the synthetic data from chapter 2 and chapter 3 were generated using SISSI (Gesell and von Haeseler, 2006). The base paired regions (stems) were generated using the SH-model (see Equation 1.6). The dinucleotide frequencies $\boldsymbol{\pi}_d = \{\pi_{AA}, \pi_{AC} \ldots \pi_{UU}\}$ are displayed in Table A.1. Positions that were not base paired evolved according to the HKY single nucleotide substitution model (Table 1.1 with nucleotide frequencies:
$\boldsymbol{\pi}_s = \{\pi_A, \pi_C, \pi_G, \pi_U\} = \{0.166, 0.262, 0.333, 0.239\}$.
The transition and transversion parameters equal one.

Bibliography

Akmaev, V., S. Kelley, and G. Stormo, 1999 A phylogenetic approach to RNA structure prediction. Proc. Int. Conf. Intell. Syst. Mol. Biol. **7**: 10–17.

Akmaev, V., S. Kelley, and G. Stormo, 2000 Phylogenetically enhanced statistical tools for RNA structure prediction. Bioinformatics **16**: 501–512.

Batey, R., S. Gilbert, and R. Montange, 2004 Structure of a natural guanine-responsive riboswitch complex with the metabolite hypoxanthine. Nature **432**: 411.

Bremaud, P., 1999 *Markov chains, Gibbs Fields, Monte Carlo Simulation and queues*. Springer-Verlag New York.

Bronstein, I. N. and K. A. Semendjajew, 1996 *Teubner-Taschenbuch der Mathematik (Teil 1)*. Teubner Verlagsgesellschaft Leipzig.

Chen, Y., D. B. Carlini, J. F. Baines, J. Parsch, J. M. Braverman, S. Tanda, and W. Stephan, 1999 RNA secondary structure and compensatory evolution. Genes Genet Syst **74**: 271–86.

Chiu, D. and T. Kolodziejczak, 1991 Inferring consensus structure from nucleic acid sequences. CABIOS **7**: 347–352.

Cox, D. R., 1962 Further results on tests of separate families of hypotheses. J. Roy. Statist. Soc. B **24**: 406–424.

Crick, F., 1958 On protein synthesis. Sym. Soc. Exp. Biol. **12**: 138–163.

Dowell, R. D. and S. R. Eddy, 2004 Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics **5**: 71.

Dytham, C., 2003 *Choosing and Using Statistics*. Blackwell Publishing, Oxford, UK.

Evans, M. and J. Rosenthal, 2003 *Probability and Statistics*. W.H. Freeman and Company.

Faith, D. P., 1992 Conservation evaluation and phylogenetic diversity. Biol. Conservat. **61**: 1–10.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**: 368–376.

Felsenstein, J., 2004 *Infering Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Fisher, R., 1922 On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society **85**: 87–94.

Fitch, W. M., 1971 Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. **20**: 406–416.

Ge, Y., S. Dudoit, and T. Speed, 2003 Resampling-based multiple testing for microarray data analysis. TEST **12**: 1–44.

Gesell, T. and A. von Haeseler, 2006 In silico sequence evolution with site-specific interactions along phylogenetic trees. Bioinformatics **22**: 716–722.

Goldman, N., 1993 Statistical tests of models of DNA substitutions. J. Mol. Evol. **36**: 182–198.

Goldman, N., J. L. Thorne, and D. T. Jones, 1996 Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J Mol Biol **263**: 196–208.

Graef, S., J. H. Teune, D. Strothmann, S. Kurtz, and G. Steger, 2005 A computational approach to search for non-coding RNAs in large genomic data. In *Nucleic Acids and Molecular Biology, Vol. 17*, edited by C. Hammann and W. Nellen, Springer-Verlag.

Gulko, B. and D. Haussler, 1996 Using multiple alignments and phylogenetic trees to detect RNA secondary structure. Pac Symp. Biocomput. pp. 350–367.

Gutell, R., A. Power, G. Hertz, and G. Putz, E.J.and Stormo, 1992 Identifying constraints on the higher-order structure of RNA: continued development of comparative sequence analysis methods. Nucl. Acids Res. **20**: 5785–5795.

Hasegawa, M., H. Kishino, and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**: 160–174.

Higgs, P., 2000 RNA secondary structure; physical and computational aspects. Q. Rev. Biophys. **30**: 199–253.

Hofacker, I., M. Fekete, and P. Stadler, 2002 Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. **319**: 1059–1066.

Jensen, J. L. and A. M. K. Pedersen, 2000 Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. **32**: 499–517.

Ji, Y., X. Xu, and G. Stormo, 2004 A graph theoretical approach for prediction common RNA secondary structure motifs including pseudoknots in unaligned sequences. Bioinformatics **20**: 1591–1602.

Juan, V. and C. Wilson, 1999 RNA secondary structure prediction based on free energy and phylogenetic analysis. J Mol Biol **289**: 935–47.

Jukes, T. and C. Cantor, 1969 Evolution of protein molecules. In Mammalian protein metabolism (Munroe,H.H.,ed.) **3**: 21–132.

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.

Klingler, T. and D. Brutlag, 1993 Detection of correlations in tRNA sequences with structural implications. Proc. Int. Conf. Intell. Syst. Mol. Biol. **1**: 225–233.

Knudsen, B. and J. Hein, 1999 RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics **15**: 446–454.

Lapedes, A., B. Giraud, L. Liu, and G. Stormo, 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.

Lauritzen, S., 1996 *Graphical models*. Oxford: Clarendon Press.

Lück, R., S. Gräf, and G. Steger, 1999 ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. Nucl. Acids Res. **27**: 4208–4217.

Marchetti, G. M. and M. Drton, 2006 *ggm: Graphical Gaussian Models, Functions for fitting Gaussian Markov models.*.

Mathews, D. H., J. Sabina, M. Zuker, and D. H. Turner, 1999 Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288**: 911–940.

Mattick, J. S. and I. V. Makunin, 2006 Non-coding RNA. Hum Mol Genet **15 Spec No 1**: R17–29.

McCaskill, J. S., 1990 The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29**: 1105–19.

Meli, M., B. Albert-Fournier, and M. C. Maurel, 2001 Recent findings in the modern RNA world. Int Microbiol **4**: 5–11.

Mossel, E., 2003 On the impossibility of reconstructing ancestral data and phylogenies. J Comput Biol **10**: 669–76.

Muse, S. V., 1995 Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139**: 1429–1439.

Navidi, W. C., G. A. Churchill, and A. von Haeseler, 1991 Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol Biol Evol **8**: 128–43.

Notredame, C., 2002 Recent progress in multiple sequence alignments: a survey. Phamacogenomics **3**: 131–144.

Paradis, E., K. Strimmer, J. Claude, G. Jobb, R. Opgen-Rhein, J. Dutheil, Y. Noel, and B. Bolker, 2004 *ape: Analysis of Phylogenetics and Evolution*. R package version 1.4.

Pollock, D. D., W. R. Taylor, and N. Goldman, 1999 Coevolving protein residues: Maximum likelihood identification and relationship to structure. J. Mol. Biol. **287**: 187–198.

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.

Rambaut, A. and N. C. Grassly, 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 Z. Yang, 1996 Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. **43**: 304–311.

Rodriguez, F., J. L. Oliver, A. Main, and J. R. Medina, 1990 The general stochastic model of nucleotide substitution. J. Theor. Biol. **142**: 485–501.

Rzhetsky, A., 1995 Estimating substitution rates in ribosomal RNA genes. Genetics **141**: 771–783.

Sachs, L., 1992 *Angewandte Statistik*. Springer Verlag.

Saitou, N. and M. Nei, 1987 The neighbor–joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**: 406–425.

Savill, N., H. D.C., and H. P.G., 2001 RNA sequence evolution with secondary structure constraints: comparison of substitutions rate models using maximum-likelihood methods. Genetics **157**: 399–411.

Schöniger, M. and A. von Haeseler, 1994 A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. **3**: 240–247.

Schöniger, M. and A. von Haeseler, 1999 Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. J. Mol. Evol. **49**: 691–698.

Semple, C. and M. Steel, 2003 *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and Its Applications*. Oxford University Press, Oxford, UK.

Siepel, A. and D. Haussler, 2004 Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. **21**: 468–488.

Sprinzl, M., C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg, 1998 Compilation of tRNA sequences and sequences of tRNA genes. Nucl. Acids Res. **26 No.1**: 148–153.

Steger, G., 2003 *Bioinformatik Methoden zur Vorhersage von RNA-und Proteinstrukturen*. Birkhäuser-Verlag.

Steinberg, S. and R. Cedergren, 1995 A correlation between N2-dimethylguanosine presence and alternate tRNA conformers. RNA **1**: 886–91.

Strimmer, K. and A. von Haeseler, 2003 Nucleotide substitution models. In *The Phylogenetic Handbook*, edited by M. Salminen, pp. 348–377, Cambridge University Press, Cambridge, UK.

Tabaska, J. E., R. B. Cary, H. N. Gabow, and G. D. Stormo, 1998 An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics **14**: 691–699.

Tamura, K. and M. Nei, 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.

Tavaré, S., 1986 Some probabilistic and statistical problems on the analysis of DNA sequences. Lec. Math. Life Sci. **17**: 57–86.

Tillier, E. R. M., 1994 Maximum likelihood with multiparameter models of substitutions. J. Mol. Evol. **39**: 409–417.

Tillier, E. R. M. and R. A. Collins, 1998 High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics **148**: 1993–2002.

Vinh, L. and A. von Haeseler, 2004 IQPNNI: Moving fast through tree space and stopping in time. Mol. Biol. Evol. **21**: 1565–1571.

von Haeseler, A. and M. Schoniger, 1998 Evolution of DNA or amino acid sequences with dependent sites. J Comput Biol **5**: 149–63.

Wallace, M., G. Blachshields, and D. Higgins, 2005 Mutltiple sequence alignment. Cur. Opin. Struct. Biol. **15**: 261–266.

Waterman, M. S., 1995 *Introduction to Computational Biology-RNA Secondary Structure*. Chapman and Hall, London.

Zuker, M., 2000 Calculating nucleic acid secondary structure. Curr. Opin. Struct. Biol. **10**: 303–310.

Die hier vorgelegte Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 26.02.2007

(Thomas Schlegel)