Codon-based Models of Evolution and Applications in Mammalian Phylogeny

A dissertation submitted to the ETH ZURICH

for the degree of Doctor of Sciences

presented by
ADRIAN SCHNEIDER
Dipl. Informatik-Ing. ETH
born 14 April 1979
citizen of Wetzikon and Volketswil, Switzerland

accepted on the recommendation of
Prof. Dr. Gaston H. Gonnet, examiner
Dr. Gina M. Cannarozzi, co-examiner
Prof. Dr. Steven A. Benner, co-examiner
Prof. Dr. Axel Janke, co-examiner

2009
Abstract

Many applications and analyses in computational biology require a model to describe the evolutionary process at the level of DNA or protein sequences. Typically a Markov model is employed which defines the probabilities that a nucleotide or amino acid is replaced by another over time. Besides the reconstruction of phylogenetic trees, models of molecular evolution are also needed for reconstructing the evolutionary history of a particular gene, for searching homologous sequences and their classification as paralogs or orthologs, for determining if a gene or part of it was subject to selection and when this occurred, or for dating evolutionary events (such as speciations or duplications).

This thesis is concerned with modeling sequence evolution and its use in phylogenetic reconstruction. In the first part, an empirical model of evolution for codons (nucleotide triplets) is presented. This model combines the DNA level information of nucleotide substitutions with the protein level information of amino acid substitutions. The range of applicability is established by comparing the codon model to an amino acid model for doing alignments of coding sequences and for phylogenetic reconstructions. Furthermore, the codon model is used as the basis for a method called "SynPAM" to estimate evolutionary distances based only on synonymous substitutions (substitutions that conserve the encoded amino acid). These substitutions are much less dependent on functional selection pressure and thus provide an evolutionary signal which is more clock-like than that of the nonsynonymous substitutions. Comparisons of SynPAM estimates to other distance measures based on synonymous substitutions show that SynPAM estimates have less variance, contain more phylogenetic signal and increase linearly with time over a longer range.

In the second part of my thesis, evolutionary models are applied to problems in mammalian phylogeny. A reanalysis of a large set of genomic data from human, dog and mouse using several different methods results in strong support for a human-dog clade for most, but not all of the methods. A large-scale analysis on thousands of mammalian quartets shows that it is not uncommon that phylogenetic methods can return strong
statistical support for an incorrect topology. This phenomenon is investigated by means of simulations as well as theoretical analysis, to determine the conditions under which a method can be mislead and thus provide strong support for an incorrect tree. Finally, a method is presented which quantifies the fact that support for a branch should decrease, when increasingly distant outgroups are used. Simulations show that this method allows the identification of the correct branchings even when most phylogenetic methods fail.
Zusammenfassung

Viele Bioinformatikanwendungen benötigen ein Modell, das den evolutionären Prozess auf der Ebene von Protein- oder DNA-Sequenzen beschreibt. Typischerweise wird dazu ein Markov-Modell verwendet, welches die Wahrscheinlichkeiten definiert, dass bestimmte Nukleotide oder Aminosäuren im Laufe der Zeit durch andere ersetzt werden. Ausser zum Rekonstruieren von phylogenetischen Bäumen (evolutionären Stammbäumen), werden Modelle für die molekulare Evolution auch verwendet, um die evolutionäre Geschichte eines Gens zu bestimmen, um homologe Sequenzen zu finden und für deren Klassifizierung als Orthologe oder Paraloge, um zu bestimmen, ob ein Gen oder ein Teil davon Selektion ausgesetzt war und wann das geschah, oder zum Datieren von evolutionären Ereignissen (wie Artenbildungen oder Genduplikationen).


Im zweiten Teil meiner Arbeit werden evolutionäre Modelle auf Probleme im Säugetierstammbaum angewandt. Eine Neuanalyse von einer grossen Menge Genom-
Acknowledgments

It all started with my Master thesis which I did under the supervision of Gina Cannarozzi in the CBRG group of Gaston Gonnet. Throughout my time as a PhD student, we continued working together on those topics and started new projects. Therefore, my foremost thanks go to Gina for guiding me through all those years, for motivating me, for teaching me a lot about science and scientific writing and last but not least for the fun we had working together. But I equally want to thank Gaston Gonnet for allowing me to work on interesting bioinformatics problems, for the helpful discussions and also for the scientific freedom he gave me.

In my opinion, the most important factor for enjoying every day’s work is the friendly people you work with. Therefore, I would like to thank all current and former members of the CBRG group for the projects we worked on together, but also for having interesting and funny discussions, coffee breaks, Friday lunches in the city, after-work beers and for the various excursions and events we could enjoy together. I am very happy that I could spend most of my time sharing the office with Adrian Altenhoff. But I also enjoyed the company of all the other people which are (in order of appearance): Markus Friberg, Christiane Baumann, Peter von Rohr, Alexander Roth, Christian Gut, Daniel Margadant, Manuel Gil, Christophe Dessimoz, Emanuele Costa, Daniel Dalquen, Adam Szalkowski and Maria Anisimova.

And finally, I consider myself very lucky to have parents that supported me to study at the ETH. Without their help and encouragement, all this would not have been possible. Therefore a big “thank you” also to them.
Contents

Abstract iii
Acknowledgments vii
Introduction and Outline 1

1 General Background 3
   1.1 The Genetic Code 3
   1.2 The Markov Model of Evolution 4
      1.2.1 Introduction and Notation 4
      1.2.2 Markov Chains 5
      1.2.3 Multiple Substitutions 5
      1.2.4 Continuous-time Processes 6
      1.2.5 Time-reversibility 6
      1.2.6 Definition of PAM 7
   1.3 Parametric and Empirical Models 7
   1.4 OMA Orthologs Project 8

I Codon-based Models of Evolution 11

2 Review of Codon Models 13
   2.1 Overview 13
   2.2 Parametric Models 14
4.3.2 Phylogenetic Accuracy ........................................... 54
4.3.3 Linearity with Time .............................................. 56
4.4 Discussion and Conclusions ........................................ 60

II Models of Evolution Applied to Mammalian Phylogenies 63

5 The Relative Positions of Human, Dog and Mouse in the Mammalian Tree 67
5.1 Introduction ......................................................... 67
5.2 Methods and Results .............................................. 68
  5.2.1 Distance Methods .............................................. 68
  5.2.2 Parsimony ....................................................... 71
  5.2.3 Likelihood Trees .............................................. 75
5.3 Discussion ......................................................... 77

6 Strong Support for Conflicting Topologies 79
6.1 Introduction ......................................................... 79
6.2 Assessing the Amount of Conflicting Quartets ...................... 80
  6.2.1 About Conflicts .............................................. 80
  6.2.2 Data and Methods ............................................ 81
  6.2.3 Results ......................................................... 82
6.3 The Effect of Model Violations ................................... 84
  6.3.1 Theoretical Analysis on Binary Characters .................. 85
  6.3.2 Simulations: Non-linearity of Distance Estimation ............ 86
  6.3.3 Simulations: Incorrect Quartet Reconstruction ............... 87
  6.3.4 Simulations: Conflicting Quartets .......................... 88
6.4 Conclusions ....................................................... 90

7 Phylogenetic Signals from Outgroup Analysis 93
7.1 Introduction ....................................................... 93
  7.1.1 Mammalian Phylogeny ....................................... 93
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1.2 Recent Advances in Phylogenetic Methodology</td>
<td>94</td>
</tr>
<tr>
<td>7.1.3 The \textit{pOG} Score</td>
<td>96</td>
</tr>
<tr>
<td>7.2 Methods</td>
<td>97</td>
</tr>
<tr>
<td>7.2.1 Gene Tree Support</td>
<td>97</td>
</tr>
<tr>
<td>7.2.2 The \textit{pOG} Score</td>
<td>98</td>
</tr>
<tr>
<td>7.2.3 Example of the \textit{pOG} Score</td>
<td>99</td>
</tr>
<tr>
<td>7.2.4 Simulations</td>
<td>100</td>
</tr>
<tr>
<td>7.2.5 Quartet Reconstruction</td>
<td>102</td>
</tr>
<tr>
<td>7.2.6 Simulated Model Violations</td>
<td>103</td>
</tr>
<tr>
<td>7.2.7 Genomic Sequences</td>
<td>103</td>
</tr>
<tr>
<td>7.2.8 Tree Optimization</td>
<td>105</td>
</tr>
<tr>
<td>7.3 Results and Discussion</td>
<td>106</td>
</tr>
<tr>
<td>7.3.1 Simulations</td>
<td>106</td>
</tr>
<tr>
<td>7.3.2 Testing Phylogenies Using the \textit{pOG} Score</td>
<td>109</td>
</tr>
<tr>
<td>7.3.3 Tree Based on the \textit{pOG} Score</td>
<td>116</td>
</tr>
<tr>
<td>7.4 Conclusions</td>
<td>117</td>
</tr>
</tbody>
</table>

**Bibliography** 122

**Curriculum Vitae** 137
Introduction and Outline

Models to describe evolution at the level of molecular sequences, i.e. DNA or protein sequences, are the basis for a wide range of applications in computational biology, including phylogenetic tree reconstruction, reconstructing the evolutionary history of a particular gene, identification of orthologs, identification of periods of selection in the evolutionary history of a gene, or dating evolutionary events (such as speciations) by assigning a rate (time/substitution) to the evolutionary process.

For my thesis, I developed an empirical model of evolution which treats codons (nucleotide triplets) as a unit. Such a model combines the DNA-level information of nucleotide substitutions with the protein-level information of substitutions between amino acids (since codons encode amino acids). This model finds application in alignments of coding sequences and for phylogenetic purposes. Furthermore, it is the basis for a model describing only synonymous substitutions (substitutions that conserve the encoded amino acid), which is much less dependent on functional selection pressure and thus contains a more clock-like evolutionary signal.

In the second part of my thesis, applications of the codon and other models to problems in mammalian phylogeny are presented. The phylogenetic problem of the relationship of human, dog and mouse is presented. An extensive analysis of the effects of model violations (a situation in which the evolutionary model does not capture the complexity of the true evolutionary process) on the reconstruction of phylogenetic trees is investigated. Finally, a method is developed, which uses a secondary phylogenetic signal and several different outgroup species to find potentially biased reconstructions.

The first two chapters provide background information for the thesis. The first treats the very basic biological and mathematical foundations for sequence analysis and evolutionary modeling, the second chapter is a review of existing models of codon evolution. Both chapters are review chapters, containing no original research and have not been published before.

The third chapter is about the empirical codon model, a project which was initiated
by Gina Cannarozzi for my diploma thesis in 2004. After improvements to both data and methods, we published it in 2005 in *BMC Bioinformatics* with Gina Cannarozzi and Gaston Gonnet as co-authors (Schneider et al., 2005). The chapter contains more detailed methodology and more verifications than the original article.

The model for synonymous codon substitutions is presented in the fourth chapter. The basis for this model was also in my diploma thesis. Together with Gina Cannarozzi and Gaston Gonnet, the model was refined and presented in 2006 at the ICCS in Reading, UK and published in the conference proceedings (Schneider et al., 2006). As a follow-up, it was invited for an extended publication in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (Schneider et al., 2007b). For the thesis chapter, several new analyses have been performed, which were not published in either of the mentioned articles.

The fifth chapter, a phylogenomic study of human, dog and mouse, was initiated by Gaston Gonnet as an offspring from the OMA orthologs project and was done together with Gina Cannarozzi. We published the work in 2007 in *PLoS Computational Biology* (Cannarozzi et al., 2007). The work presented in my thesis is a complete reanalysis with updated genomic data, refined methodology and somewhat different conclusions.

The sixth chapter contains a large-scale empirical analysis using real mammalian sequence data to establish the failure rate of quartet reconstruction and simulations as well as theoretical analysis to explore the conditions under which phylogenetic reconstruction can become biased towards an incorrect result. The work presented in this chapter has not yet been published.

The seventh and last chapter presents a method based on outgroup analysis which can be useful for detecting biased reconstructions. It was joint work with Gina Cannarozzi and was published in *Molecular Biology and Evolution* (Schneider and Cannarozzi, 2009).
Chapter 1

General Background

This chapter introduces in a condensed form essential knowledge to understand the thesis, and references to more detailed information, typically review articles, are given. First, the genetic code and some of the associated terminology is reviewed. Then, the Markovian model of evolution with its properties and their implications on the evolutionary models are described. Since various notations can be found in the literature, a consistent notation will be introduced. Furthermore, the difference between parametric and empirical models is explained and examples of both kinds of models are given. Finally, a brief overview of OMA, the orthologous matrix project, is presented, which provided the data for most of the work presented in this thesis.

1.1 The Genetic Code

The genome of an organism is encoded in the DNA (deoxyribonucleic acid, in some viruses in the RNA, ribonucleic acid). DNA molecules are linear chains of of small units consisting of a phosphate, deoxyribose sugar and one of the four nitrogenous nucleotide bases. The purine bases adenine (A) and guanine (G) are larger and consist of two aromatic rings. The pyrimidine bases cytosine (C) and thymine (T) are smaller and consist of only one aromatic ring. Two DNA molecules are joined by base pairing into a double helix. These base pairings are normally between an adenine base on one strand and a thymine on the other strand and between a cytosine base on one strand and a guanine base on the other. This means that in a given double helix the number of A and T residues is the same as the number of C and G residues. A residue is a specific unit in a chain. In the context of DNA, it refers to a nucleotide, whereas in the context of proteins, it refers to an amino acid.
During transcription, parts of the genome are transcribed to single stranded RNA. These products called transcripts can either perform a function, or can be translated to proteins, possibly after being modified. During translation, the protein is assembled by the ribosome as a sequence of amino acids. Triplets of nucleotides determine which one of the 20 standard amino acids is used. These triplets are called codons and the mapping of codons to amino acids is called the genetic code and almost the same for all known forms of life on earth (see e.g. the review by Knight et al. (1999)). There are \( 4^3 = 64 \) possible codons of which 61 are assigned to an amino acid. The remaining three codons are called stop codons and determine the end of a coding sequence.

The genetic code is redundant, because 61 codons are mapped to only 20 amino acids. Therefore, most amino acids can be encoded by more than one codon. Codons that code for the same amino acid are called synonymous codons. If the code for different amino acids, they are called nonsynonymous.

### 1.2 The Markov Model of Evolution

#### 1.2.1 Introduction and Notation

Markovian models have played an important role in molecular evolution since they were first proposed for this task about forty years ago by Jukes and Cantor (1969). The models have in common that each residue of a biological sequence is considered a random variable which can change independently and with the same probabilities as every other residue in the sequence.

Before the concepts of the Markovian model are introduced, some notation needs to be defined. A matrix containing substitution probabilities is denoted \( M \), where \( M_{i,j} \) is the probability that a character in state \( j \) changes to state \( i \). The order of the indices is not universal and in the theory of Markov chains, the indices are often interpreted as \( M_{i,j} \) meaning "from \( i \) to \( j \)". However, since the the codon models presented in my thesis follow the tradition of the first empirical evolutionary model by Dayhoff et al. (1978), I will use their notation ("from \( j \) to \( i \)""). The Darwin (Gonnet et al., 2000) implementations of the algorithms described in this thesis also follow that interpretation of the indices.
1.2. Markov Model of Evolution

1.2.2 Markov Chains

Markov chains are a stochastic model of how random variables can change over time. The main property of a Markov model is that given the present state, the future states do not depend on the past, only the present state.

In molecular evolution, “states” are typically the four nucleotide bases, the 20 amino acids, or the 61 sense-codons (transitions between sense- and stop-codons usually have very dramatic consequences for the function of protein and thus must be treated separately). The transition probability matrix $M(t)$ contains the probabilities that a character $X$ changes from state $j$ to state $i$ in time $t$:

$$M_{i,j}(t) = Pr\{X(t) = i \mid X(0) = j\}$$  \hspace{1cm} (1.1)

More generally, if $\pi(0)$ is a probability vector with $\pi_i(0)$ denoting the probability that a character is in state $i$ at time $0$, then the distribution after time $t$ is $\pi(t) = M(t)\pi(0)$. \hspace{1cm} (1.2)

If the initial and the target distribution are the same, $\pi(t) = \pi(0)$, then the chain will stay in this distribution forever. This distribution is called the steady-state distribution or equilibrium frequencies $\pi$.

1.2.3 Multiple Substitutions

Given two matrices $M(t_1)$ and $M(t_2)$, describing the transition probabilities over times $t_1$ and $t_2$, respectively, the Chapman-Kolmogorov theorem explains the behavior of a Markov process over time $t_1 + t_2$: A character in state $j$ mutates to any state $k$ after $t_1$ with probability $M_{k,j}(t_1)$. Afterwards, it mutates to $i$ from any of these $k$ possible states with a probability of $M_{i,k}(t_2)$. Therefore, the probability $M_{i,j}(t_1 + t_2)$ of mutating from $j$ to $i$ over time $t_1 + t_2$ is

$$M_{i,j}(t_1 + t_2) = \sum_k M_{k,j}(t_1)M_{i,k}(t_2)$$ \hspace{1cm} (1.3)

and consequently

$$M(t_1 + t_2) = M(t_1)M(t_2).$$ \hspace{1cm} (1.4)

Thus, given a transition matrix $M(1)$ for a unit of time, the transition matrix for any other integer time $t$ can be derived as

$$M(t) = M(1)^t.$$ \hspace{1cm} (1.5)
1.2.4 Continuous-time Processes

The biological processes treated by models of molecular evolution are continuous-time processes, therefore transition probabilities for fixed time steps are not sufficient to describe the process. Thus, instead of a transition matrix $M(1)$ of a unit time step, the Markov process is generated by a rate matrix $Q$, describing the instantaneous rates of change:

$$\frac{dM(t)}{dt} = QM(t) \quad (1.6)$$

This differential equation together with the boundary condition $M(0) = I$, the identity matrix, has the solution

$$M(t) = e^{Qt} \quad (1.7)$$

which allows for the computation of $M(t)$ for any real $t \geq 0$.

If $Q$ is independent of $t$, a property that is often assumed but not necessarily biologically justified, then $Q$ is called time-homogeneous. A rate matrix has some properties which ensure that matrices derived via Equation 1.7 are valid substitution matrices: all off-diagonal entries of $Q$ are non-negative and the sum of each column is 0. Thus, the diagonal element $Q_{i,i} = -\sum_{j \neq i} Q_{j,i}$ is always negative and corresponds to the substitution rate of state $i$, the rate with which the Markov chain leaves that state.

1.2.5 Time-reversibility

Most models of molecular evolution are time-reversible, which is defined as

$$\pi_i Q_{j,i} = \pi_j Q_{i,j} \quad \text{for all } i \neq j. \quad (1.8)$$

The assumption of time-reversibility is again more mathematically than biologically motivated. However, there are several theoretical and practical benefits from this property. For example, it implies that for likelihood computations the placement of the root is irrelevant, which is very convenient since the the exact position of the root is often not known. Further conveniences are that the eigenvalues of $Q$ all become positive and that $Q$ can be expressed as the product of two symmetric matrices:

$$Q = \begin{bmatrix}
\cdot & s_{1,2} & \cdots & s_{1,n-1} & s_{1,n} \\
\cdot & \cdot & \cdots & \cdot & \cdot \\
\cdot & \cdots & \cdots & \cdot & \cdot \\
\cdot & \cdots & \cdots & \cdots & \cdot \\
s_{1,n-1} & s_{2,n-1} & \cdots & s_{n-1,n} & 0 \\
s_{1,n} & s_{2,n} & \cdots & s_{n-1,n} & 0 \\
\cdot & \cdots & \cdots & \cdots & \cdot \\
\end{bmatrix} \begin{bmatrix}
\pi_1 & 0 & \cdots & 0 & 0 \\
0 & \pi_2 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & \pi_{n-1} & 0 \\
0 & 0 & \cdots & 0 & \pi_n \\
\end{bmatrix} \quad (1.9)$$
where the diagonal elements of the first matrix (called exchangeability matrix) are chosen such that the columns of $Q$ add up to 0. The second matrix is a diagonal matrix with the equilibrium frequencies $\pi_i$ and $n$ is the number of states, e.g. four for DNA. The most generic form of such a model is called general time-reversible (GTR) (Tavare, 1986) and has $n(n + 1)/2 - 2$ parameters ($n(n - 1)/2$ for the exchangeability matrix and $n - 1$ for the frequencies, minus one because $Q$ can be normalized).

1.2.6 Definition of PAM

It follows from Equation 1.7 that a constant factor applied to all elements of $Q$ does not change the behavior of the Markov process but only the unit of $t$. Dayhoff et al. (1978) defined the "PAM" unit (point accepted mutations) where 1 PAM is defined as the evolutionary distance over which on the average one percent of all positions undergo substitution:

$$\sum_i \pi_i (1 - M_{i,i}(1)) = 0.01 \quad (1.10)$$

The factor to be applied to $Q$ such that the unit of $t$ in Equation 1.7 corresponds to 1 PAM can be found numerically.

1.3 Parametric and Empirical Models

One distinguishes between parametric and empirical models. In parametric (also called mechanistic) models, all transition probabilities are derived from a few parameters which are either based on theoretical considerations or estimated directly from the data under investigation. Empirical models often have more parameters (typically all the transition probabilities, corresponding to the GTR model) which are estimated once from a very large data set and then are fixed.

Parametric models are more flexible to adapt to peculiarities in specific data, e.g. the transition/transversion ratio was found to differ among genes and lineages (Yang, 1999). Thus, if the transition/transversion ratio is a parameter to be estimated from the data, the model may fit a particular data set better. More generally, the more parameters that can be estimated from the data, the better the fit. However, there are drawbacks of estimating too many parameters from the data. Since the amount of data used for analyses is often very limited, the estimated parameters suffer from large variances. Particularly if some parameters compensate for each other, it becomes very difficult
to estimate them from small data sets. There is also the risk of overfitting, where the model no longer describes a general process but some particularities of a specific data set.

The nucleotide-level evolution of DNA is most commonly described by parametric models, as the number of possible parameters is small. The simplest such model is the parameter-free model by Jukes and Cantor (1969) (it is often described with a rate parameter $\alpha$ for all substitutions, but since the rate matrix is normalized, no degree of freedom is left). In the following three decades, the models have been refined to allow for differing transition/transversion rate ratios (Kimura, 1980), to incorporate different equilibrium frequencies (Felsenstein, 1981; Hasegawa et al., 1985), and to model different rates for purine and pyrimidine transitions (Tamura and Nei, 1993). In some studies, the GTR model with eight free parameters is used for modeling DNA evolution.

In contrast, to model amino acid evolution mostly empirical models were proposed. One important reason is that the GTR model for amino acids has 208 free parameters, requiring very large data sets to estimate them for each problem separately and there seems to be no intuitive reduction to only a few parameters. The pioneers for empirical amino acid models were Dayhoff et al. (1978). Later, as sequence data and computational power increased, other researchers also created empirical models (Gonnet et al., 1992; Jones et al., 1992; Whelan and Goldman, 2001).

Prior to 2005, codon-based models were exclusively parametric. This was partly because the evolutionary process on codon level is quite diverse (different synonymous to nonsynonymous rate ratios are the most prominent example), but also a very large amount of data is needed to estimate the 2078 parameters of the GTR model reliably. A detailed review of parametric and semi-parametric codon models is given in Chapter 2. The first empirical model is then presented in the subsequent chapter.

1.4 OMA Orthologs Project

The OMA project (Dessimoz et al., 2005; Roth et al., 2008) is a large-scale project for finding orthologs between completely sequenced genomes. The process is based on an all-against-all comparison of the proteins in genome databases using dynamic programming and pairwise ML distance estimations. The main distinction of the OMA project is its systematic and exhaustive verification step where every potential ortholog is tested against paralogy using sequences in all possible third genomes.

Noteworthy is also the size of OMA. Since its beginning in the summer 2004, new
1.4. OMA Orthologs Project

Genomes have continuously been added and updated, reaching several hundred genomes by 2009, making it the largest project of this kind. The species cover all domains of life and although the majority are bacteria, most sequenced and publicly available metazoans are also included in OMA. A large-scale comparison of OMA to other orthology projects showed that the OMA predictions are as good as or better than those of any other project (Altenhoff and Dessimoz, 2009).

The orthologs predictions of OMA come in two variants: one is the OMA groups, which are cliques of orthologs between genomes, where every sequence is orthologous to every other sequence in the group. Therefore, OMA groups have at most one member in each species, which is a very desirable property for phylogenetic applications. However, because of gene duplications, orthologs between two species are not always a one-to-one relation, a sequence in one genome can have more than one ortholog in another genome. Thus, the OMA groups cover only a subset of all possible orthologous pairs. The complete set of orthologs as well as the OMA groups can be explored and downloaded from the OMA Browser web server (Schneider et al., 2007a) at http://www.omabrowser.org.
Part I

Codon-based Models of Evolution
Chapter 2

Review of Codon Models

In this chapter, all relevant codon models of evolution are reviewed, except the completely empirical model by Schneider et al. (2005) which will be discussed in the next chapter. For each model, the methodology is outlined and its particularities as well as its reception and use is discussed. Finally, the parameters and properties of all models are compared to each other.

2.1 Overview

Before 2005, when our empirical codon matrix was published (Schneider et al., 2005), only parametric models for codon evolution had been used. One of the main reasons is that the estimation of an empirical codon model requires a very large amount of coding DNA. Only the large-scale sequencing of complete genomes in the last few years has produced the amount of high-quality data needed for this task.

The first models describing evolution on codon level were introduced independently by Goldman and Yang (1994) and Muse and Gaut (1994) in the same issue of Molecular Biology and Evolution in 1994. The main application of codon models beside phylogeny is their use in detecting genes, sites or lineages under evolutionary selection pressure. Selection is usually expressed as the ratio between "synonymous substitutions per synonymous sites" (called $dS$ or $Ks$) and "nonsynonymous substitutions per nonsynonymous sites" (called $dN$ or $Ka$). If $dN/dS > 1$, positive selection is assumed, if $dN/dS < 1$ it is considered negative selection and $dN/dS = 1$ is interpreted as neutral evolution.

All of the codon models described in this chapter are used in an ML framework, in
which the parameters of the model and a tree relating (codon-wise aligned) sequences are optimized to maximize the likelihood of the data. The model parameters and the tree can be optimized simultaneously or alternatingly. Depending on the specific problem to be addressed, either the tree itself or the optimized parameter values are of interest. Often, likelihood ratio tests (LRTs) are performed in order to compare hypotheses about parameters, for example to test for positive selection.

Codon-based models of evolution can be divided into three categories: The parametric models that were presented in the 1990s, the completely empirical model as it was presented in 2005 (see the next Chapter) and combinations of these two approaches, which were first presented in 2007 (Kosiol et al., 2007; Doron-Faigenboim and Pupko, 2007).

### 2.2 Parametric Models

#### 2.2.1 Model by Muse and Gaut (1994)

**Introduction**

Muse and Gaut introduced their model as a method to estimate synonymous and non-synonymous substitution rates. They criticized previous approaches to this task for using parsimony methods that underestimate the number of evolutionary events on a branch. Their proposed ML-based model would not suffer from that problem and it since it is based on a Markov process, the explicit counting and weighting of possible multiple substitution paths typical of traditional approaches can be avoided.

**Method**

The model uses six parameters, $\alpha$ and $\beta$ for the synonymous and non-synonymous substitution rates as well as the four nucleotide frequencies $\pi_x$. The instantaneous rate between codons $i$ and $j$ is only positive if they differ by exactly one nucleotide:

$$Q_{i,j} = \begin{cases} 
\alpha \pi_n & \text{if synonymous} \\
\beta \pi_n & \text{if nonsynonymous} \\
0 & \text{if more than one change}
\end{cases}$$

(2.1)

Here, $\pi_n$ is the the frequency of the "target nucleotide", e.g. the rate for the synonymous substitution from AGG to AGA is $\alpha \pi_A$. 

2.2. Parametric Models

The model describes a $61 \times 61$ matrix for the sense codons, as stop codons are not treated. The equilibrium frequencies of the codons can be derived as $\pi_i \pi_j \pi_k / (1 - \Pi_{\text{stop}})$ for the codon consisting of nucleotides $i, j$ and $k$ where $\Pi_{\text{stop}}$ is the sum of the products $\pi_i \pi_j \pi_k$ of the stop codons.

Discussion

Although the authors claim that their model is similar to the one by Hasegawa et al. (1985), they do not consider different rates for transitions and transversion. Also, since $Q$ can be normalized, the two parameters $\alpha$ and $\beta$ could be replaced by a single parameter for the synonymous/nonsynonymous rate ratio, as it was done in later models by Yang and co-workers (see below). This would allow for a more intuitive and direct interpretation of the estimated parameters.

The two parameter approach, however, was used by the authors to separately test for synonymous and nonsynonymous rate differences in different lineages. They applied LRTs to two lineages of a tree by either allowing arbitrary $\alpha$’s and $\beta$’s in both lineages or to constrain one or both of them to be the same. A test very similar to this was presented again a few years later by Yang (1998), but the Markov model introduced by Muse and Gaut was lacking some of the desirable properties of the model by Goldman and Yang (1994), which could explain why it did not find widespread use (see next section).

2.2.2 Models by Yang and co-workers (1994–2002)

Introduction

Whereas Muse and Gaut (1994) presented their model as a new way to identify selection in genes, Goldman and Yang (1994) introduced their model mainly as a Markov model for phylogenetic reconstructions based on codon substitutions which should better fit the sequence data than mononucleotide models. As an additional advantage they found that the model could also be used to obtain ML estimates of transition/transversion as well as synonymous/nonsynonymous rate ratios.

Method

The model is again a $61 \times 61$ Markov model. The model contains a parameter $\kappa$ which is applied to the rates describing transitions. Additionally, the physicochemical distances
between amino acids from Grantham (1974) were used to account for the similarities between different amino acids. These distances are based on several properties of the amino acids and are denoted \( d_{aa_i,aa_j} \) with \( aa_i \) being the amino acid encoded by codon \( i \). Finally, they used a parameter \( V \) representing "the variability of the gene or its tendency to undergo nonsynonymous substitution". It is negatively correlated with the synonymous/nonsynonymous rate ratio, but not in a linear or otherwise simple way.

The substitution rate between from codon \( j \) to codon \( i \) was only assumed to be positive for codon substitutions involving a single nucleotide change:

\[
Q_{i,j} = \begin{cases} 
\pi_i \cdot \exp\left(-d_{aa_i,aa_j}/V\right) & \text{if transversion} \\
\kappa \pi_i \cdot \exp\left(-d_{aa_i,aa_j}/V\right) & \text{if transition} \\
0 & \text{if more than one change}
\end{cases}
\] (2.2)

Here, \( \pi_i \) is the equilibrium frequency of codon \( i \). The original formula also included a scaling factor \( \mu \) to normalize the rate matrix, which is omitted here.

Once \( \kappa \) and \( V \) are estimated, the synonymous substitution rate per codon can be computed as the weighted sum of the rates of all synonymous substitutions \( i,j \) with \( i \neq j \).

\[
\rho_s = \sum_{\substack{i \neq j \\text{aa}_i=\text{aa}_j}} \pi_j Q_{i,j}.
\] (2.3)

The nonsynonymous rate \( \rho_n \) can either be computed analogous to Equation 2.3 or – since \( Q \) is normalized – simply as \( 1 - \rho_s \). The ratio \( \rho_s/\rho_n \) expresses the relative rates of synonymous and nonsynonymous substitutions and the authors proposed to use this ratio instead of similar ratios obtained by counting methods. The transition/transversion rate ratio can be computed in a similar way.

**Discussion**

The model presented by Goldman and Yang was later used to test for selection in specific lineages (Yang, 1998), specific sites (Nielsen and Yang, 1998), or in sites and lineages simultaneously (Yang and Nielsen, 2002). For this purpose, the model from 1994 was modified in order to allow for a more intuitive representation of the synonymous to nonsynonymous rate ratio \( dN/dS \) which was called \( \omega \). This, however, came at the cost of no longer including the amino acid distances. The rate matrix used in all
2.2. Parametric Models

these publications was defined as follows:

\[
Q_{i,j} = \begin{cases} 
\pi_i & \text{if synonymous transversion} \\
\kappa \pi_i & \text{if synonymous transition} \\
\omega \pi_i & \text{if nonsynonymous transversion} \\
\omega \kappa \pi_i & \text{if nonsynonymous transition} \\
0 & \text{if more than one change}
\end{cases}
\]  

(2.4)

Because \(Q\) is again normalized and it no longer contains the exponential factors depending on the amino acid distance and \(V\), the two parameters \(\kappa\) and \(\omega\) directly correspond to the transition/transversion and nonsynonymous/synonymous rate ratios, respectively.

All these methods have the commonality that they were created to compare two or more hypotheses about selection. To test for lineage specific selection (Yang, 1998), some (in practice, two, three or all) branches are allowed to have an arbitrary ratio \(\omega\). These parameters are then optimized on a tree and the obtained likelihood is compared to the one of the null hypothesis of all branches having the same \(\omega \leq 1\).

In the site-specific model (Nielsen and Yang, 1998), some proportions of sites are allowed to have different (possibly constrained) \(\omega\)’s. The parameters of the test include then the proportions of site categories as well as possibly several \(\omega\)’s that are optimized. Typically, a proportion \(p_1\) of sites would be fixed to be neutral \((\omega_1 = 1)\) and another proportion \(p_2\) would be constrained to model negative selection \((\omega_2 < 1)\), while the rest would have positive selection \((\omega > 1)\). Again, the likelihood of this model is compared to a null hypothesis where no \(\omega\) is allowed to be larger than 1.

The combination of the two approaches above was presented four years later (Yang and Nielsen, 2002). In order to limit the parameters needed to simultaneously search for selection among sites and lineages, the tree is a priori divided into "foreground" and "background" branches. The background branches are constrained to have only neutral and negative selection, whereas the foreground branches are tested for having sites with positive selection.

The models of Yang an co-workers are widely used and are currently the most popular codon models. In addition to desirable properties like modeling the two main aspects of DNA evolution with the parameters \(\kappa\) and \(\omega\), an important reason for their success is their integration into the \textit{paml} software package (Yang, 1997).
2.2.3 Model by Whelan and Goldman (2004)

Introduction

The codon models discussed so far did not consider instantaneous rates between codons which differ by more than one nucleotide. This was explained by the assumption that only one nucleotide can change at a time. Additionally, such a model is easier to formulate and the interpretation of the parameters is more intuitive. Whelan and Goldman (2004) were the first to explore how justified this simplifying assumption is by proposing a model which also takes double- and triple-nucleotide substitution rates into account.

Method

The exact method to derive the rate matrix $Q$ is rather complicated and would go beyond the scope of this short review. Therefore, the procedure is only outlined here and the parameters are introduced.

The main idea of the model by Whelan and Goldman is that they consider substitutions of two or three consecutive nucleotides as one possible evolutionary event. The complications from incorporating these events into a codon model of evolutions arise from the fact that these multiple substitutions events can affect more than one codon at a time. The overall rate matrix $Q$ is defined as the sum of all rate matrices describing (s)ingle-, (d)ouble- and (t)riple-nucleotide substitutions:

$$Q = \sigma Q^{(s)} + \delta Q^{(d)} + \tau Q^{(t)}$$

(2.5)

where $Q^{(s)}$, $Q^{(d)}$ and $Q^{(t)}$ contain the rates for the respective categories and the parameters $\sigma$, $\delta$ and $\tau$ define the relative frequencies of the three kinds of substitutions ($\sigma + \delta + \tau = 1$).

The rate matrix $Q^{(s)}$ for the single nucleotide substitutions is the simplest to describe and is very close to previously discussed models. It contains the target nucleotide frequencies $\pi$, the transition rate factor $\kappa$ and a factor for nonsynonymous substitutions $\omega_1$.

The matrices $Q^{(d)}$ and $Q^{(t)}$ for double and triple nucleotide substitutions are more complicated to define. The main task is to correctly account for substitution events that affect two neighboring codons. For example, a double-nucleotide substitution can change the last nucleotide of one codon and the first of the next codon. These events are assumed to occur uniformly across the whole sequence, independent of intron and exon
structure. The matrix $Q^{(d)}$ is defined by the appropriately weighted probabilities of the double-nucleotide events that affect only the first position of the codon, those that affect the first two or the last two and those that affect only the last position. $Q^{(t)}$ is defined along the same idea, but for triple-nucleotide events. Multiple substitution event probabilities are based on the nucleotide frequencies $\pi$, and two parameters $\omega_1$ and $\omega_2$ indicating if one or two nonsynonymous substitutions events occurred (in the sense of "number of amino acids changed").

**Discussion**

By fitting their model to real sequences, the authors report significant rates of double and triple substitutions (compared to the null hypotheses of no such events, $\delta = \tau = 0$). Also, those events are found to occur "far more frequently than previously thought". It remains unclear, however, if these results really correspond to such events in evolution or if they are the effect of some larger-scale events (gene conversion, inversions or small-scale recombinations).

Probably because of its rather complicated definition and the non-intuitive parameter set, this model is not used very often for phylogenetic studies or selection inference. It seems that its main purpose was to establish the fact that instantaneous rates of multiple nucleotide substitutions are not negligible and necessary for a realistic codon model.

**2.3 Combined Parametric and Empirical Models**

As stated in the Introduction, there are various forces driving the evolution on codon level and the varying contributions of these different factors are difficult to capture by a fixed (i.e. unparameterized) model of evolution. Therefore, in 2007, two approaches were published that combine an empirically based codon-model with parameters to be estimated from the data under investigation (Doron-Faigenboim and Pupko, 2007; Kosiol et al., 2007).
2.3.1 Model by Doron-Faigenboim and Pupko (2007)

Introduction

Doron-Faigenboim and Pupko (2007) presented a "combined empirical and mechanistic" model of codon evolution in which empirical transition rates between amino acids were included in a parametric codon model. Their model is similar to the one by Goldman and Yang (1994), since both models weight codon substitutions with a factor for the amino acid change. But there are also differences: the Grantham distances used in the Yang model are not proper substitution rates between amino acid but "distances" based on their physicochemical properties. In this combined model, empirical amino acid replacement rates are used. Furthermore, whereas the Yang model allowed only instantaneous rates between codons that differ at most one position, this model does not restrict the number of nucleotide changes.

Methods

Doron-Faigenboim and Pupko (2007) use the empirical amino acid replacement matrix by Jones et al. (1992) (or other, more purpose-specific matrices) to construct a codon replacement model. The relationship between the amino acid substitution rates $A_{i,j}$ with equilibrium frequencies $\psi_j$ and the codon substitution rates $Q_{l,s}^*$ with codon frequencies $\pi_s$ has been pointed out by Yang et al. (1998): $A_{i,j}$ is a weighted sum over the rates $Q_{l,s}^*$ of all possible codon substitution between amino acids $i$ and $j$:

$$\psi_j A_{i,j} = \sum_{l: aa_l = i} \sum_{s: aa_s = j} \pi_s Q_{l,s}^*$$

(2.6)

The codon rate matrix $Q^*$ contains six parameters for different combinations of transitions and transversions as well as factors $x_{(aa_1, aa_2)}$ describing the weights for the amino acid substitutions. $Q_{l,s}^*$ for any $l \neq s$ is defined as follows:

$$Q_{l,s}^* = \kappa(i, j) \cdot \pi_l \cdot x_{(aa_1, aa_s)}$$

(2.7)

where $\kappa(i, j)$ stands for the parameter used to weight the respective combination of transitions and transversions. The six possible parameters are $t_r$ and $t_v$ denoting one transition or one transversion, respectively, $t_{rr}$, $t_{rv}$, and $t_{vv}$ for the possible combinations of two changes and $t_{sup}$ for all substitutions with three nucleotide changes. The $20 \times 20$ $x$-factors can be determined by solving Equation 2.6 after the other parameters of $Q^*$ have been estimated.
2.3. Combined Parametric and Empirical Models

The $Q^*$ matrix is the basis for the construction of a matrix $Q'$ which includes the parameter $\omega$, allowing for the estimation of the nonsynonymous/synonymous rate ratio:

$$Q'_{l,s} = \begin{cases} 
\omega \cdot Q^*_{l,s} & \text{for nonsynonymous substitutions} \\
Q^*_{l,s} & \text{for synonymous substitutions}
\end{cases} \quad (2.8)$$

Because of the selection implied from the amino acid substitution rates, an $\omega$ of 1 in this model does not correspond to neutral evolution. Therefore, a matrix $Q^0$ is introduced which models the neutral codon evolution by leaving away all $x$ factors:

$$Q^0_{l,s} = \kappa(i,j) \cdot \pi_l \quad (2.9)$$

Finally, the complete codon model is assembled by combining the selection matrix $Q'$ and the neutral matrix $Q^0$:

$$Q = f \cdot Q' + (1 - f) \cdot Q^0 \quad (2.10)$$

with $f$ being the probability that the sites are under selection. It is also a parameter estimated from the data.

Discussion

The authors use an Akaike information criterion (Akaike, 1974) to compare their codon model to two mechanistic models by Yang et al. (2000) and the completely empirical model presented in Chapter 3. They observe that most of the times, their model can be better fit to the biological data sets used in those tests, concluding that allowing more than one nucleotide change as well as incorporating rates for amino acid replacements significantly improves the degree of realism of the codon models.

2.3.2 Model by Kosiol et al. (2007)

Introduction

Another version of a combination of empirical and parametric model was presented by Kosiol et al. (2007). Whereas Doron-Faigenboim and Pupko (2007) started from an amino acid transition matrix and then extend it to a $61 \times 61$ codon model, these authors start with an empirical codon model and introduce parameters to allow for different $dN/dS$ ratios as well as different transition/transversion rate ratios among genes.
Chapter 2. Review of Codon Models

The empirical codon matrix that is the basis for their model was not derived using the counting based method presented in Chapter 3. Instead, they followed the method by Whelan and Goldman (2001) where a rate matrix is estimated by ML from a set of alignments and trees.

Methods

Starting from 7,332 multiple sequence alignments and corresponding trees from the Pandit database (Whelan et al., 2003, 2006) a rate matrix $Q^*$ for a GTR model was estimated. This matrix can be decomposed into a symmetric exchangeability matrix $S^*$ and codon frequencies $\pi^*$:

$$Q^*_{i,j} = \pi^*_i S^*_{i,j} \text{ for all } i \neq j$$

(2.11)

The matrix $S^*$ is the basis for the parametric rate matrix $Q$, which contains the parameters $\pi$ (the codon frequencies estimated for the data set under consideration), $\omega$ (the nonsynonymous/synonymous rate ratio) and $\kappa(i,j)$ (for weighting transition and transversion rates, see discussion below):

$$Q_{i,j} = \begin{cases} 0 & \text{if } i \text{ or } j \text{ is a stop codon} \\ \pi_i \cdot S^*_{i,j} \cdot \kappa(i,j) & \text{for synonymous substitutions} \\ \pi_i \cdot S^*_{i,j} \cdot \kappa(i,j) \cdot \omega & \text{for nonsynonymous substitutions} \end{cases}$$

(2.12)

Five different models are suggested for $\kappa(i,j)$ to incorporate transition and transversion rates. The first one simply uses the average ratio from the empirical model ($\kappa(i,j) = 1$ for all $i$ and $j$). Two models have only one parameter, either $t_r$ or $t_v$, where either the transversions or the transitions are weighted with a factor. This factor is assumed to be multiplicative, e.g. substitutions with two transitions have a factor of $t_r^2$. A further variant contains both of these factors, so a substitutions with with one transition and one transversion would be weighted with by $t_r \cdot t_v$. The most complex version would have nine parameters where each possible combination of transversions and transitions is weighted by a specific factor.

Discussion

Using a subset of 200 alignments from Pandit, the authors test the different variants of their model and compare them to purely parametric models. They conclude that the version with nine $\kappa(i,j)$ parameters does not significantly better fit the data and is therefore over parameterized. The tests suggest that the model with a transition and a
transversion parameter gives the best results. When comparing their semi-parametric model with parametric models by Yang et al. (2000), they observe very significant improvements in terms of likelihoods.

The authors also discuss the significant improvements obtained by allowing also instantaneous rates for substitutions involving more than one nucleotide change. Unlike Whelan and Goldman (2004) who assumed that substitutions of two or three consecutive nucleotides can occur through a single event, Kosiol et al. (2007) propose the idea that substitutions that happen independently but compensate for each other have a larger chance of getting fixed in the population when they occur within a short time interval.

Because their model is based on empirical codon substitution patterns, there is already an average transition/transversion as well as nonsynonymous/synonymous rate bias included in the model. The $\kappa$ and $\omega$ parameters only reflect the deviations from the average bias. In order to interpret those parameters, they need to be normalized by dividing the estimates by values expected under neutral evolution.

### 2.4 Summary

The codon models presented in this chapter have in common that they are presented to be used in an ML framework for finding phylogenies or estimating selection. All models are defined by a rate matrix $Q$ describing the instantaneous substitution rates between the 61 sense codons and contain parameters that are fit to the data sets under investigation. However, the models differ in what aspects of codon evolution is parameterized and how the parameters are incorporated. These differences are summarized in Table 2.1.

The consensus result from the later models is clearly that substitutions that change the amino acids cannot be treated equally. There is no reason why the amino acid similarities should not be reflected in a codon model. A second essential aspect is codon substitutions in which more than one nucleotide changes, even for instantaneous rates. All studies agree that allowing such substitutions significantly improves the performance of the models, but no conclusive explanation for this phenomenon has been given so far.
Chapter 2. Review of Codon Models


<table>
<thead>
<tr>
<th>Feature</th>
<th>MG94</th>
<th>GY94</th>
<th>Y++</th>
<th>WG04</th>
<th>DP07</th>
<th>KHG07</th>
</tr>
</thead>
<tbody>
<tr>
<td>transition/transversion rate ratio</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>dN/dS rate ratios</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>dN/dS as a simple parameter</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>unequal amino acid transitions</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>multiple nucleotide substitutions</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>uses codon frequencies</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Chapter 3

An Empirical Markov Model of Codon Evolution

The previous chapter was a review of parametric and semi-parametric models of codon evolution. In this chapter, the completely empirical model by Schneider et al. (2005) is introduced, which was the first empirical codon model, based on alignments of orthologs from five vertebrate species. After the methodology, the various aspects of the codon matrix are compared to empirical amino acid matrices, in order to establish the optimal ranges of applicability for both types of matrices.

3.1 Introduction

Codon models proposed prior to 2005 were parametric and modeled only certain aspects of the evolution of DNA. None of the models included exchangeabilities between amino acids while allowing more than one nucleotide change at a time. Fully parameterized models could also fail to capture important features of codon-wise replacements that are not modeled by any parameter. Thus, to study the complete pattern of evolution on codon level, an empirical model of codon substitutions was created from a large set of alignments of vertebrate coding sequences.

For an empirical model, a complete substitution matrix is estimated once from a large data set and then fixed. From this substitution matrix, matrices describing other evolutionary distances as well as a rate matrix can be derived using the theory of Markov models (see Section 1.2). The main use of an empirical codon model of evolution is not in an ML framework, where a tree and a model is fit to the sequence data under
consideration and where estimation of the parameters is an important step to improve the fit of the model to the data. However, for creating sequence alignments, either pairwise or of multiple sequences, log-odds matrices are used. These matrices are derived from substitution probabilities matrices and are typically parameter-less, thus empirical models are especially suited for this task.

Sequence alignment using empirical amino acid substitution matrices has a long tradition and many useful applications, since alignments are the basis for most phylogenetic or molecular evolution studies. When sequences are aligned on a codon instead of an amino acid basis, the information from DNA level substitutions can also be used in order to get alignments of better quality. The empirical codon model presented in this chapter is based on the data from pairwise sequence alignments and thus its main application is in that area.

## 3.2 Methods

### 3.2.1 Selecting an Unbiased Set of Alignments

Empirical models of evolution are based on large sets of alignments. These must fulfill a variety of criteria: certainly they must be of good quality, otherwise too many nonhomologous characters are counted. The problem with "good quality", however, is that the true alignment is not known. But two measures were taken to ensure as high quality as possible. First, all sequences were eliminated that contained any unknown nucleotides (Xs). The occurrence of Xs in a sequence usually indicates that there was some uncertainty in the sequencing process. Additionally, the use of dynamic programming (Smith and Waterman, 1981; Gotoh, 1982) to create local alignments is the best way to consider only those parts of the sequences that can be reliably aligned and where the confidence is high that the residues actually share common ancestry.

Furthermore, the distance range must be selected carefully. If the distance between aligned sequences is too small, then only the most frequent substitutions can be observed but there is no information about the rarer changes. If, on the other hand, the distances are too high, then the problem of multiple and back substitutions occurs which also blurs the true pattern of molecular evolution. Also, the range of the used distance should be as small as possible to avoid averaging over too many different distances which would obscure the estimation of the Markovian process. But at the same time, the distance range must be large enough to contain a sufficient number of alignments.
After some preliminary analyses, a distance range of 25 to 60 PAM was chosen for the alignments. The PAM distance alone, however, is not enough to exclude saturated alignments, since sometimes even at a low PAM distance, many synonymous substitutions can be observed, most likely because a protein has been very conserved over a long period of time. For this reason, the alignments were further reduced to those with an $f^2$ value between 0.50 and 0.95. $f^2$ is the fraction of conserved two-fold codons, a very simple measure of the amount of synonymous substitutions.

A decision had to be made about the taxonomic range of the sequences used to estimate the matrix. Prokaryotic and other single-cellular organisms are known to have a strongly biased selection between synonymous codons and that bias can be very different in different organisms. In vertebrates, however, the codon usage bias does not differ much among the different species (Ikemura, 1985). Additionally, the vertebrates are separated from each other by evolutionary distances compatible with the ranges given above. For these reasons, only alignments from vertebrate sequences were used to estimate the codon matrix.

Another important aspect is to find an unbiased data set. This means that no gene family, no period of evolution and no lineage should be overrepresented. Using orthologous groups from the OMA project should ensure diversity with respect to gene families, since for every family at most one ortholog between any two species will be used. Thus, the orthologs from the vertebrate set of OMA were selected as the basis to estimate the empirical codon matrix. To avoid counting substitutions on internal branches of the species tree too often, not all pairs of species were used, but the species tree was followed in a circular tour, which ensures that each branch is counted at most twice.

At the time when the codon matrix was created, the OMA project contained five vertebrates: human (*Homo sapiens*), mouse (*Mus musculus*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*) and zebrafish (*Danio rerio*). The genome databases all came from Ensembl (Hubbard et al., 2007), a database of annotated eukaryotic genomes. After following the circular tour and applying the filtering according to the distance ranges and the synonymous change criteria outline above, a total of 17,502 alignments with 8.3 million aligned codons were left to estimate the matrix. The number of alignments from each pair of genomes is tabulated in Table 3.1.

### 3.2.2 Estimating a Probability Matrix from Observed Substitutions

Estimating an empirical matrix for a particular type of data for the first time is related to the "chicken or egg problem". That is, the matrix must be estimated from alignments,
but the alignments can only be created using an existing matrix. Fortunately, there is a relatively easy solution to this specific instance of the problem. The initial alignments can be created with a rough estimate of a scoring matrix (e.g. an arbitrary positive score for exact matches and a negative score for all replacements). Then, a first better matrix can be estimated and, in an iterative process, the sequences can be realigned using the new matrix, which in turn allows for the estimation of an improved matrix.

In the case of the empirical codon matrix, this procedure was simplified since the protein sequences could be aligned using empirical amino acid matrices (Gonnet et al., 1992). The aligned proteins were then mapped to the corresponding coding DNA, resulting in codon-wise aligned DNA sequences. The stop codons which are not translated into an amino acid, had to be treated separately: whenever the last codons of two sequences were aligned to each other, then the respective stop codons were appended to the aligned sequences. Aligning stop codons with regular codons was not considered, since this would have a too dramatic effect on the function of a protein and can thus be considered to be very rare. In subsequent iterations, the previously estimated codon matrices could be used to directly align the coding sequences. After six iterations, sufficient convergence was reached.

Given the aligned coding sequences, the observed mutations were counted and the counts stored in a matrix called $C$ with $C_{i,j}$ being the number of times a substitution of codon $j$ to codon $i$ was observed. However, since pairwise alignments were used, the direction of the changes was not known. Thus, whenever $i$ and $j$ were different, $C_{i,j}$ as well as $C_{j,i}$ was increased by $1/2$. $C$ is a symmetric matrix and the sum of the rows or the columns corresponds to the number of times each codon was observed. For the reasons outlined above, there were no counts between stop codons and regular codons. Therefore $C$ is a block-diagonal matrix consisting of a $61 \times 61$ matrix of sense codon substitutions and a $3 \times 3$ matrix of stop codon substitutions. Often, only the $61 \times 61$

<table>
<thead>
<tr>
<th>Genome 1</th>
<th>Genome 2</th>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>mouse</td>
<td>3,107</td>
</tr>
<tr>
<td>mouse</td>
<td>chicken</td>
<td>3,691</td>
</tr>
<tr>
<td>chicken</td>
<td>frog</td>
<td>3,671</td>
</tr>
<tr>
<td>frog</td>
<td>fish</td>
<td>3,441</td>
</tr>
<tr>
<td>fish</td>
<td>human</td>
<td>3,592</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>17,502</td>
</tr>
</tbody>
</table>

Table 3.1: Genome pairs from the circular tour around the vertebrates tree and the number of orthologous alignments used from each pair
3.2. Methods

matrix of the sense codon is used. However, throughout this thesis the empirical codon matrix will be used as a $64 \times 64$ matrix, even though the stop codons often have no effect on the results. Table 3.2 displays some properties of the observed matrix $C$. Note that for finding the most frequent and rarest substitutions, the off-diagonal elements of $C$ had to be doubled.

Given the counts matrix $C$, a matrix $M$ containing the substitution probabilities can be derived. As stated earlier, $M_{i,j}$ denotes the probability of codon $j$ being changed to codon $i$. Consequently, the columns of $M$ must sum up to 1: $\sum_j M_{x,j} = 1$. The matrix $M$ can be computed from $C$ as follows:

$$M_{i,j} = \frac{C_{i,j}}{\sum_x C_{x,j}}.$$

(3.1)

The equilibrium frequencies $\pi$ of the codons are derived from the column sums of $C$:

$$\pi_i = \frac{\sum_x C_{x,i}}{\sum_a \sum_b C_{a,b}}.$$

(3.2)

By taking the logarithm of $M$ (see Equation 1.7), a rate matrix $Q$ can be obtained. Unlike $C$, the matrices $M$ and $Q$ are no longer symmetric. However, since $C$ is symmetric, the matrices describe a time-reversible process (Equation 1.8). The evolution that caused the observed patterns most likely did not follow exactly the Markovian model. Therefore, it is not guaranteed that a rate matrix $Q$ can be derived that would explain the observed counts. Indeed, the logarithm of $M$ results in a matrix with negative off-diagonal values, corresponding to negative transition rates for some rare substitutions. This is problematic, since it would lead to negative substitution probabilities for small distances. In order to prevent this, all negative off-diagonal entries of $Q$ were replaced with a rate of 0.

<table>
<thead>
<tr>
<th>Total substitutions</th>
<th>8,343,554</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most frequent (on diagonal)</td>
<td>GAG – GAG 153,040</td>
</tr>
<tr>
<td>Most frequent (off-diagonal)</td>
<td>GAA – GAG 169,440</td>
</tr>
<tr>
<td>Rarest (on diagonal)</td>
<td>TAG – TAG 566</td>
</tr>
<tr>
<td>Rarest (diagonal, not stop)</td>
<td>TCG – TCG 3,297</td>
</tr>
<tr>
<td>Rarest (off-diagonal, not stop)</td>
<td>GAC – TGG 45</td>
</tr>
<tr>
<td>Most frequent codon</td>
<td>GAG 426,383</td>
</tr>
<tr>
<td>Rarest codon</td>
<td>TAA 1,283</td>
</tr>
<tr>
<td>Rarest codon (not stop)</td>
<td>ATA 25,688</td>
</tr>
</tbody>
</table>

Table 3.2: Statistics about the count matrix $C$ of observed codon substitutions.
and the diagonal elements of \( Q \) were adjusted to satisfy the condition that the columns of \( Q \) must sum to 0.

### 3.2.3 Definition of CodonPAM and Relation to PAM

Analogous to the definition of PAM (Section 1.2.6), the unit of the codon matrix is called CodonPAM and 1 CodonPAM is the evolutionary distance over which one percent of the codons are expected to undergo substitution. Equation 1.10 (definition of 1 PAM) is also fulfilled for \( M \) being the 1 CodonPAM matrix and \( \pi \) the codon frequencies. Following the procedure described in Section 1.2.6, the codon rate matrix \( Q \) was scaled such that \( e^Q \) resulted in the 1 CodonPAM probability matrix. The scaling factor was sought numerically and was found to be 141.7, which means that the estimated matrix \( M \) describes the substitution probabilities at an average of 141.7 CodonPAM.

Given a codon substitution matrix \( M(t) \) for \( t \) CodonPAM, it is possible to compute the \( 20 \times 20 \) matrix \( M' \) containing the same probabilities for the amino acid substitutions but ignoring the synonymous substitutions:

\[
M'_{i,j} = \sum_a \sum_b \pi_b M_{a,b}(t) \frac{\pi_a}{\pi_b}
\]  

(3.3)

with the sums of \( a \) and \( b \) over all codons encoding amino acids \( i \) and \( j \), respectively. \( M(t) \) and \( \pi \) must not contain frequencies and substitution probabilities for the stop codons, since those probabilities would not be used in the computation of \( M' \) and thus the column sums of \( M' \) would be less than 1. Once \( M' \) is computed, its PAM value can be found as described above. Consequently, there is a bijective relationship between PAM and CodonPAM, since for every CodonPAM value the corresponding PAM can be computed and for every PAM value, the CodonPAM matrix with the same amount of amino acid change can be found.

The relation between CodonPAM and PAM has been computed for a CodonPAM range of 0 to 1000. The result is plotted in Figure 3.1. The relation is almost linear, but the conversion factor CodonPAM/PAM becomes slightly larger for longer CodonPAM distances. At 10 CodonPAM, the factor is 3.55 (resulting in 2.8 PAM) whereas at 1000 CodonPAM, the factor is 3.61 (resulting in 277 PAM).
Figure 3.1: Relation between CodonPAM and PAM. The curve is flattens almost unnoticeably with increasing CodonPAM.


3.2.4 Scoring Matrices and Gap Costs

In order to use a Markovian model of evolution for dynamic programming algorithms, scoring matrices as well as gap costs need to be defined. The scoring matrices \( S(t) \) depend on the evolutionary distance (here CodonPAM) and have been defined by Dayhoff et al. (1978) as the following log-odds:

\[
S_{i,j}(t) = 10 \log_{10} \frac{\pi_j M_{i,j}(t)}{\pi_i \pi_j} = 10 \log_{10} \frac{M_{i,j}(t)}{\pi_i} \tag{3.4}
\]

The interpretation of the fraction in this term is to compare the probability of observing a substitution of character \( j \) to character \( i \) (over distance \( t \)) to the probability of a random pairing of the two characters. Taking the logarithm converts the likelihoods (the numerator) into an additive score which is necessary for dynamic programming alignment algorithms.

The most common type of gap costs used for dynamic programming are affine costs, where a gap of length \( k \) is scored with costs of the form \( a + (k - 1)b \) where \( a \) is the cost of opening a new gap and \( b \) is the cost of extending an existing gap by one more character. Benner et al. (1993) derived empirical gap costs from a large set of protein alignments in which the parameter \( a \) is a function of the evolutionary distance \( t \) (in PAM). After refinements (new database versions) the gap costs were determined to be

\[
\text{Costs}(k, t) = -37.64 + 7.434 \log_{10}(t) - 1.3961(k - 1). \tag{3.5}
\]

Since one codon corresponds to one amino acid, it is justified to use the same gap costs for codon alignments. Thus, for a codon alignments at a CodonPAM distance \( t \), the equivalent PAM distance is determined and from that the gap costs according to Equation 3.5 can be computed.

3.3 Results and Discussion

Since one of the main application of the empirical codon matrices is their use as scoring matrices for dynamic programming, it makes sense to assess the performance of codon matrices for aligning sequences. The most relevant competitor for this task are the amino acid substitution matrices which are commonly used for alignments. In this chapter, two tests are presented: first, the accuracy (the percentage of correctly aligned) positions is investigated. This has to be done on simulated sequences, since for real data the true alignment is not known. Additionally, the scores which are an
indicator of significance of an alignment are compared between codon and amino acid matrices. This analysis was performed using real sequences from different subsets of species.

The use of the CodonPAM measure as a distance estimator and for phylogenetic reconstruction will be analyzed in the next chapter together with the SynPAM measure presented there.

### 3.3.1 Alignment Accuracy

To compare the accuracy of sequence alignments between codon and amino acid matrices, simulated alignments were created. For the alignments, a PAM range of 5 to 250 was chosen since this is the relevant range for most purposes. This corresponds to a CodonPAM range of 18 to 903. At each step of 5 PAM, 1000 random codon sequences of length 500 codons were created and then randomly mutated over the CodonPAM corresponding to the respective PAM distance. Insertions and deletions were simulated according to an exponential distribution (Walpen, 2008).

Each alignment from the simulation was taken to be a “true” alignment $A_T$ because it is known what the truly homologous characters are. By removing the gap information from $A_T$, two homologous codon sequences of possibly different lengths were created. These were then aligned using the CodonPAM scoring matrix for the simulated distance, leading to a codon alignment $A_C$. Furthermore, the two codon sequences were translated to proteins and aligned using the PAM scoring matrix derived from the codon matrix as described in Equation 3.3, leading to an amino acid alignment $A_A$. This way it is ensured that the amino acid scoring matrix corresponds exactly to the codon matrix, but without the synonymous substitutions probabilities.

The alignments $A_T$, $A_C$ and $A_A$ can be represented as sets of aligned character pairs. Each member of an alignment set is either of the form ”character $x$ in sequence 1 is aligned to character $y$ in sequence 2” or ”character $z$ in sequence $i$ is aligned to a gap”.

The fractions $f_C$ and $f_A$ of correctly aligned positions in $A_C$ and $A_A$ can then be defined using set operators:

$$f_C = \frac{|A_C \cap A_T|}{|A_T|} \quad \text{and} \quad f_A = \frac{|A_A \cap A_T|}{|A_T|}.$$  \hspace{2cm} (3.6)

Figure 3.2 shows the percentages of correctly aligned positions using codon or amino acid matrices. The fraction of correct positions decreases almost linearly with increasing distance and falls below 50% at around 600 CodonPAM (approx. 170 PAM). The percentage of correct alignments (alignments with 100% correct positions) is also
shown in the figure. These curves fall very steeply and beyond 300 CodonPAM (approx. 85 PAM) almost no alignments are correct. The difference between the two types of matrices appears to be very small, in the average codon matrices align 0.8% more positions correctly. But it should be considered, that no dramatic difference can be expected, since almost the same matrices are used. The codon matrix contains information about the synonymous substitutions, but this information only makes a difference for a few positions. However, the 0.8% mentioned above correspond to 4 positions in a 500 codon long alignment, and since most positions in an alignment are not disputed, these 4 positions can make an important difference.

Figure 3.2: Percentages of correctly aligned positions and alignments using codon or amino acid scoring matrices as a function of CodonPAM distance between the two sequences.
3.3. Results and Discussion

3.3.2 Alignment Scores

Comparing models of evolution using simulations has its merits, most notably that the true answers to the problems are known and therefore the outcome of any method can be easily evaluated. The drawback, however, is that the simulations themselves are based on a model of evolution and these models might not accurately reflect reality. For this reason the use of simulation studies is limited and thus the comparison of codon based and amino acid based alignments was also performed on real sequences.

When real sequences are used, the true alignment is not known and therefore the alignment accuracy cannot be computed. What can be compared, however, is the score of the alignment. The score expresses how much more likely it is under a given model that the two sequences are homologous compared to being matched by chance. Since both the scores of the amino acid alignments as well as of the codon alignments are based on the same type of model, it is valid to compare the scores and to draw conclusions about the confidence in the alignment using the two models.

This analysis was performed on orthologs from the OMA project. There is no reason for using orthologs for this task except that it is convenient, as the search for homologous sequences has already been performed. Two sets of data were used: the vertebrates, which is the taxonomic range from which the codon matrices have been created, although that was on a small subset of the current data, and the cyanobacteria, as an independent test set for which the codon matrices are not optimized.

All pairs of orthologous sequences were aligned using the optimal codon matrix and the gaps were removed from the alignments. The reason for this is that the gaps also contribute to the score, but this contribution should not be considered here, since the matrices and not the gap scores are compared. The optimal (i.e. the highest scoring) codon matrix was then sought for the gapless DNA alignment, as well as the optimal amino acid scoring matrix for the translated gapless sequences. As in the previous section, the amino acid matrices were derived from the codon matrices and therefore describe the same evolutionary process, but on a different level.

The scores were divided by the sequence lengths in order to normalize them and make them comparable across alignments of different lengths and distances. The resulting scores per site are shown as a function of evolutionary distance in Figures 3.3 and 3.4. The lines connect average scores per site from a range of 10 CodonPAM.

On both plots, it can be seen that codon-wise alignments produce clearly higher scores for distances up to 100 CodonPAM (28 PAM). Up to this distance the additional information from the DNA increases the confidence in the alignments. From there on upwards,
with increasing saturation of the synonymous substitution, the difference of the two methods becomes very small. This means most likely that the amino acid substitutions dominate in this distance range and the DNA contributed very little information. A difference can be observed between the two data sets: On the vertebrate data, for which the codon matrices were optimized, the codon based scores are always higher than the amino acid based ones. However, when sequences from cyanobacteria are used, then the codon model performs slightly less well for distances greater than 120 CodonPAM.

Figure 3.3: Average scores per site of 2,272,815 alignments from vertebrate sequences as a function of evolutionary distances. The difference between the codon and amino acid scores is difficult to see in the plot and is therefore shown separately using a different y-axis.
3.3. Results and Discussion

Figure 3.4: Average scores per site of 330,978 alignments from cyanobacteria sequences as a function of evolutionary distances. The thin blue line visualizes the difference between the codon based (red line) and amino acid based scores (green, dashed line).
3.4 Conclusions

The purely empirical model of codon evolution presented here was the first of its kind. Because of it being parameterless, its primary use is as scoring matrices for alignments of coding DNA. In tests on simulated sequences, it was shown to produce more accurate alignments than amino acid scoring matrices. And based on the scores of alignments of real sequences, it can be concluded that codon based alignments outperform amino acid alignments for short distances and at least on vertebrate DNA, they are competitive also for the longer distances. On prokaryotic DNA however, the amino acid model has an advantage for higher distances. This is not unexpected, since, as stated in the introduction, single-cellular organisms tend to optimize their codon usage to very specific distributions and this can mislead models that are based on averages from large data sets.

The use of the codon model to estimate only synonymous change and a comparison of CodonPAM and other measures on other, mostly phylogenetic problems, will be discussed in the next chapter.
Chapter 4

A Model for Synonymous Codon Substitutions

This chapter describes the construction of submatrices from codon substitution matrices that describe only synonymous substitutions. The SynPAM distance measure is introduced which is a distance measure based on the synonymous substitution matrices. The SynPAM distance estimates are compared to distance estimates from two other methods to estimate synonymous substitutions. This comparison is done on real biological sequences, as simulations often favor the model which is closest to the one used to simulate. The drawback of comparisons on real data is that the true result is normally not known, thus indirect measures have to be found. Several such tests are presented here.

4.1 Introduction

It has been observed already in the early times of molecular evolution that the relative rates of nonsynonymous and synonymous substitutions vary considerably among genes (see e.g. Miyata and Yasunaga (1980)). This phenomenon is treated differently by the various codon models (see previous chapter for details). The purely parametric models typically have a parameter (commonly called $\omega$) modeling the rate ratio of nonsynonymous and synonymous substitutions. Fully empirical models do not consider this difference among genes but use the same average ratio for all analyses. One option to compensate for varying rate ratios among genes is by re-introducing parameters into an empirical model, as it was done by Kosiol et al. (2007). Another option is to estimate the amount of synonymous substitutions independently of the nonsynonymous
Chapter 4. A Model for Synonymous Codon Substitutions

substitutions.

A distance based only on the synonymous substitutions has the advantage that it is not influenced by amino acid (and thus potentially functional) changes. Therefore, neither periods of rapid adaption to new life conditions nor changes in the amino acid composition influence such a measure. However, the time range where synonymous substitution based measures can be applied is smaller, since saturation affects the synonymous substitutions sooner than the amino acid changing substitutions.

For short distances, where the substitutions on DNA level do not suffer from saturation effects, estimates based on synonymous substitutions outperform amino acid based estimates for phylogenetic tree reconstruction (see the analysis in Section 4.3.2). A particularly suitable application for distances based on synonymous change is molecular dating, for which it is desirable to have substitutions occurring as clock-like as possible. If a metric captures the signal from changes that do not evolve because of selection but of random drift, then the obtained distance estimates should be proportional to the time since the divergence of the sequences. If the rate of change per time can also be established, then molecular dating methods are able to estimate divergence times of genes or species.

Traditionally, the amount of synonymous change between two coding sequences was expressed as the "number of synonymous substitutions per synonymous site", termed $K_s$ or $d_S$. Although codon-wise replacements have to be considered, all these methods express the distance on a per nucleotide level. These methods can be divided into two main categories. One contains the "counting methods", in which the number of synonymous sites and changes are determined by comparing two sequences and counting the respective events. The other covers the "ML methods" which usually require a multiple sequences alignments to which a tree and parameters of a model (see previous chapter) are fit. From those estimates $d_S$ can be derived.

The next two subsections review common methods to estimate $d_S$ using the counting or the ML method. After that, SynPAM, a novel method to estimate distances based on synonymous substitutions will be introduced. Various tests to evaluate and compare the different metrics will be presented in the last section of this chapter.

### 4.1.1 Methods based on Counting

The first publication describing a method for separately measuring synonymous and nonsynonymous is found in a study of the chicken preproinsulin gene from 1980 (Perler et al., 1980). In the same year, a methodological article about the problem was pub-
lished and applied to most of the genes sequenced at that time (Miyata and Yasunaga, 1980). Their methods were later refined and extended (Li et al., 1985). Nei and Gojobori (1986) proposed a method that simplified the previous approaches and showed with simulations that these simplifications came at almost no cost in terms of accuracy. For this reason and since their method was clearly simpler, it became the most widely used method for estimating $dN$ and $dS$ (more than 1700 citations according to Google Scholar). Therefore, this approach is used here to illustrate the principle of the "counting methods".

The Nei and Gojobori (1986) method, as most methods of the counting approach, is defined on a pairwise sequence comparison and consists of three steps:

1. Count the number of synonymous and nonsynonymous sites in the sequences.

2. Compute the number of synonymous and nonsynonymous differences between the sequences.

3. Convert the number of observed changes per site into distances by estimating the amount of multiple substitutions.

Counting Synonymous and Nonsynonymous Sites

A codon is composed of three nucleotides and thus is defined to have "three sites". Because of the complexity of the genetic code, most of the sites cannot be divided into either synonymous or nonsynonymous. Instead, fractions are assigned. Consider for example the codon TTT coding for phenylalanine (Figure 4.1). Changes at the first two positions are always nonsynonymous, whereas the third site is only synonymous if it changes to a C and nonsynonymous if it changes to A or G. Thus, the last site is deemed 1/3 synonymous and 2/3 nonsynonymous. With the first two sites being completely nonsynonymous the codon TTT has in total 1/3 synonymous and 8/3 nonsynonymous sites. Nei and Gojobori (1986) do not mention how stop codons should be treated, but Yang (2006), has suggested that changes to stop codons are not counted for either category. For example the last site of TAT is synonymous if it changes to C but if it changes to A or G, it is a stop codon. Thus, the site counts as fully synonymous.

The counting is performed for each aligned codon of both sequences and then the averages between the sequences are denoted $S$ and $N$ for the total number of synonymous and nonsynonymous sites, respectively.
Chapter 4. A Model for Synonymous Codon Substitutions

Counting Synonymous and Nonsynonymous Changes

In the next step, the two sequences are compared codon by codon and for each aligned codon pair the number of synonymous and nonsynonymous differences are determined. If the two codons differ at only one position, then the change is unambiguously either synonymous or nonsynonymous. More complicated are the cases of two or three changes. In these situations, more than one pathway (orders in which the changes happened) are possible. In the case of two changes, there are two pathways and both are weighted by 1/2, in the case of three changes even six pathways need to be considered and each is weighted by 1/6. For example in the comparison of TTT and GTA, two pathways are possible (see Figure 4.1): either the first change is from TTT to GTT (nonsynonymous) and then the change from GTT to GTA is synonymous. Or, if the changes happen in the different order, then both changes (TTT to TTA and TTA to GTA) are nonsynonymous. Thus, there are 1/2 synonymous and 3/2 nonsynonymous differences between TTT and GTA. Again, any paths involving stop codons are ignored. The sum of the number of synonymous and nonsynonymous differences between two sequences are denoted $S_d$ and $N_d$ respectively.

![Diagram showing the neighborhood of TTT and the two possible substitution paths between TTT and GTA.](image)

Figure 4.1: Illustrations of the counting methods. Right: "neighborhood" of TTT defining the number of synonymous and nonsynonymous sites. Only one of the nine possible substitutions (to TTC, marked in red) is synonymous, thus TTT is assigned 1/3 synonymous and 8/3 nonsynonymous sites. Left: the two possible substitution paths between TTT and GTA. Only one half of the upper path is synonymous (shown in red), thus TTT and GTA are separated by 1/2 synonymous and 3/2 nonsynonymous substitutions.
4.1. Introduction

Correcting for Multiple Changes

Both the synonymous and nonsynonymous changes can be considered independent Markov processes. The method by Nei and Gojobori (1986) assumes that all transitions are equally likely and all nucleotides equally frequent. Therefore, the method by Jukes and Cantor (1969) can be employed to estimate the expected number of substitutions from the observed changes:

\[ d = -\frac{3}{4} \ln(1 - \frac{4}{3}p) \]

(4.1)

where \( p \) is the fraction of observed changes, either \( S_d/S \) or \( N_d/N \). The resulting estimates are called \( dS \) for "synonymous substitutions per synonymous site" and \( dN \) for "nonsynonymous substitutions per nonsynonymous site".

It has been pointed out (Lewontin, 1989) that this step is not correct. The Jukes-Cantor model is designed for noncoding DNA, where every nucleotide can freely change to any of the three others. But particularly the synonymous sites most often have less than three possibilities for synonymous substitutions. However, the authors showed with simulations that this method performs well, and because of its simplicity it is frequently used for estimating \( dS \) and \( dN \).

Variants

Interestingly, the development of these measures did not follow the usual direction from simple models to more complex ones. For example, the different substitution paths that have to be considered when two codons differ in more than one position, are not equally likely. Miyata and Yasunaga (1980) proposed to weight the different paths proportional to a physicochemical distance measure between the encoded amino acids, whereas Li et al. (1985) used a combination of physicochemical distances and empirical substitution rates between different codons to weight the pathways. Li et al. (1985) also proposed to treat transitions and transversions differently and thus use the Kimura two-parameter model (Kimura, 1980) instead of the simpler model by Jukes and Cantor (1969).

Over the years many more methods for estimating \( dS \) and \( dN \) were presented using more complex nucleotide substitution models as well as more sophisticated counting methods (Li, 1993; Pamilo and Bianchi, 1993; Comeron, 1995; Ina, 1995). These methods were shown to produce considerably different estimates of \( dN \) and \( dS \). Yang (2006) summarizes the four main observations from various studies:
1. Ignoring transition to transversion rate ratio differences leads to an overestimation of $dS$ and thus to an underestimation of $dN/dS$.

2. Ignoring biased codon usage leads to the opposite effect (underestimation of $dS$). This has the ironic consequence that the method by Nei and Gojobori (1986) which ignored both codon bias and transition to transversion rate ratios, often outperforms more sophisticated methods.

3. Different methods and model assumptions lead to very different estimates. Unlike distance estimations (of nucleotides or amino acids), where the model choice has only a small influence, the estimation of $dN$ and $dS$ appears to lack robustness.

4. Under the same model assumptions, the differences between counting methods and ML methods often produce similar results. Thus, the assumptions appear to matter more than the method.

Another noteworthy measure of "synonymous substitutions per synonymous site" was introduced as "NED" by Caraco (2002) and later refined as "TREx" by Li et al. (2006). Their method considers only synonymous substitutions of at the third nucleotide of two-fold redundant codons. This has the advantage that no multiple nucleotide substitutions have to be considered and that all substitutions are transitions. In opposition to the traditionally used Markov models to account for multiple substitutions, the distance correction is derived from a chemical decay-to-equilibrium process (but resulting in the same formula). The NED method and also the more complex TREx method are easy to implement and therefore useful for quick analyses.

4.1.2 ML Methods for Estimating $dS$ and $dN$

A different approach for estimating $dS$ and $dN$ was presented with the introduction of Markov models of codon evolution (Goldman and Yang, 1994; Muse, 1996). These models have in common that in addition to the evolutionary distance $t$ (in "substitutions per site"), several parameters of the rate matrix $Q$ are estimated from the data using ML. From the rate matrix, $dS$ and $dN$ can be derived. The results depend on the parameters of $Q$, typically $\omega$, the nonsynonymous/synonymous rate ratio, but also the transition/transversion rate ratio $\kappa$ and the codon frequencies $\pi$.

Given a rate matrix $Q$ with all its parameters estimated, the expected number of substitutions from codon $i$ to codon $j$ over time $t$ is $\pi_i Q_{j,i} t$. (Note that this is different from the expected number of observed differences, which would be $\pi_i M_{j,i}(t)$ with $M(t) = e^{Qt}$.)
The expected number of synonymous and nonsynonymous substitutions per codon (denoted $S_d$ and $N_d$) at distance $t$ can be derived by summing up these terms for all synonymous or all nonsynonymous substitutions:

$$S_d = t \rho_S = \sum_i \sum_{j \neq i: a_i = a_j} \pi_i Q_{j,i} t$$

$$N_d = t \rho_N = \sum_i \sum_{j \neq i: a_i \neq a_j} \pi_i Q_{j,i} t$$

(4.2)

Since $Q$ is normalized to 1 substitution per codon, $\rho_S + \rho_N = 1$ and consequently $S_d + N_d = t$.

The number of synonymous and nonsynonymous sites per codon (denoted $S$ and $N$) are derived in a similar way as $S_d$ and $N_d$, but using a matrix $Q^{(1)}$ describing neutral evolution (e.g. $\omega = 1$ in simple codon models). Furthermore, since the unit of $Q^{(1)}$ is “substitutions per codon” the result is multiplied by three to get the number per site:

$$S = 3 \rho_S^{(1)} = 3 \sum_i \sum_{j \neq i: a_i = a_j} \pi_i Q_{j,i}^{(1)}$$

$$N = 3 \rho_N^{(1)} = 3 \sum_i \sum_{j \neq i: a_i \neq a_j} \pi_i Q_{j,i}^{(1)}$$

(4.3)

Using the results from Equations 4.2 and 4.3, $dN$ and $dS$ are computed as follows:

$$dS = S_d / S = t \rho_S / (3 \rho_S^{(1)})$$

$$dN = N_d / N = t \rho_N / (3 \rho_N^{(1)})$$

(4.4)

Estimating the parameters of $Q$ under ML for computing $dN$ and $dS$ is computationally more expensive than the counting methods and the derivation is less intuitive. The advantage is the clear statistical framework that makes the counting and weighting of pathways no longer necessary. The simplifying assumptions are now only in the evolutionary codon model. The resulting $dN$ and $dS$, however, are not what can be observed from the sequences, but the expected values at distance $t$ with the particular parameters of $Q$.

### 4.2 Methods

A different approach than either the counting or the ML methods is the SynPAM method (Schneider et al., 2006, 2007b). It employs the empirical codon substitution matrix from the previous chapter to infer an evolutionary distance estimated from synonymous codon replacements. The use of an empirical codon model is not a fundamental part of the method, but it presents itself since a parameter-less model allows for faster estimation.
Chapter 4. A Model for Synonymous Codon Substitutions

The method itself resembles in some aspects the NED method since it also considers only aligned codon pairs that code for the same amino acid. But unlike NED where only two-fold redundant codons are used and the decay to the equilibrium is modeled explicitly, SynPAM employs a Markov model of evolution which takes all possible synonymous replacements into account.

4.2.1 Synonymous Substitution Matrix

The synonymous substitution probabilities are derived from a codon substitution matrix $M(t)$ where every entry $M_{a,b}(t)$ is the probability that codon $b$ mutates to codon $a$ over distance $t$ (the unit of $t$ does not matter at this point). The SynPAM presented here is based on the empirical codon matrix from the previous chapter and also the comparison studies in the next section are based on that matrix. In order to measure the evolutionary distance based only on synonymous substitutions, the full codon substitution matrix is transformed to a matrix containing probabilities for only the synonymous substitutions: Starting with a codon substitution matrix $M(t)$ for any evolutionary distance $t$, the probabilities of all nonsynonymous substitutions are set to 0. The remaining probabilities are then rescaled such that all possible codon substitutions between codons coding for a given amino acid sum to 1, resulting in a synonymous codon substitution matrix $M^*$:

\[
M^*_{a,b}(t) = \begin{cases} 
\frac{M_{a,b}(t)}{\sum_{x:aa_x=aa_b}M_{x,b}(t)} & \text{if } a \text{ and } b \text{ synonymous} \\
0 & \text{if } a \text{ and } b \text{ nonsynonymous}
\end{cases} \quad (4.5)
\]

From the initial codon substitution matrix, a wide range of matrices describing the substitution process at different evolutionary distances are derived through matrix exponentiation (Equation 1.7) and then converted to synonymous substitution matrices. It is important that the exponentiation is executed on the full substitution matrix and only then the transformation described above should be performed. The reason is that the reduced substitution matrix no longer describes the full Markov process of codon evolution. E.g. some codons have a larger chance to undergo nonsynonymous substitutions than others, but this information would be lost, if only the synonymous matrices were exponentiated. In addition, by including all elements of the substitution matrix until the time of reduction, all alternative pathways, back mutations and multiple hits are taken into consideration.
4.2. Methods

4.2.2 Definition of SynPAM

Following the definitions of PAM (Section 1.2.6) and CodonPAM (Section 3.2.3), 1 SynPAM is defined as the amount of evolution in which 1 percent of the conserved amino acids were subject to a synonymous substitution at the codon level:

$$\sum_{i=0}^{64} \pi_i (1 - M^*_i, i (1)) = 0.01$$

The matrix $M^*(1)$ is the 1 SynPAM substitution matrix. As described before, the exponent to transform the initial codon matrix to the substitution matrix from which $M^*(1)$ can be derived has to be found numerically.

The SynPAM distance between two aligned coding sequences is the $t$ which maximizes the likelihood of the observed synonymous substitutions. In order to find this distance efficiently, scoring matrices in the tradition of Dayhoff (Equation 3.4) are defined:

$$S_{i,j}(t) = 10 \log_{10} \frac{\pi_j \cdot M^*_{i,j}(t)}{\pi^*_i \cdot \pi_j} = 10 \log_{10} \frac{M^*_{i,j}(t)}{\pi^*_i}$$

with $\pi^*_j$ being the relative codon frequencies for a given amino acid:

$$\pi^*_i = \frac{\pi_i}{\sum_{j:aa_j=aa_i} \pi_j}$$

The scores can be interpreted as the ratio of the probability of the two synonymous codons stemming from a common ancestor over a total evolutionary distance $t$ to the chance of being two randomly chosen synonymous codons. Nonsynonymous substitutions are assigned a score of 0 for any distance. Thus, they have no influence on the SynPAM distance estimation.

Typically, SynPAM scoring matrices are precomputed for a range of SynPAM distances (e.g. from 0 to 1000 SynPAM). Then, a given alignment can be scored with a selection of SynPAM matrices in order to determine which one produces the highest score. The scoring is performed by summing the scores $S_{i,j}$ corresponding to the aligned codon pairs in an alignment.

4.2.3 Example

To demonstrate the application of the above formulas, the transformation is presented on a small subset of the full codon matrix, namely on the codons AAA and AAG (coding
for lysine) and AAC and AAT (coding for asparagine). Table 4.1 shows the three steps from the full substitution matrix at a distance of 20 SynPAM (top), the corresponding matrix containing the conditional substitution probabilities for the synonymous mutations (left, after applying Equation 4.5), to the log-likelihood scores for 20 SynPAM which form a symmetric matrix (right, after applying Equation 4.7).

4.3 Results

Comparing different evolutionary methods is a difficult task. The experiments have to be designed carefully to avoid biases towards any of the methods. Most often, comparison studies are based on simulated evolution. This has the advantage that all methods can be applied to the simulated sequences and the resulting estimates can directly be compared to the simulation parameters. The problem with this approach is that the evolutionary model chosen for the simulation has a strong influence on the results. Most often, the method seems to perform best which uses a similar (or the same) model as the simulation. This phenomenon was demonstrated in the semester project by Keller (2008). Therefore, only verification and comparison on real sequences is presented here.

The OMA project (see Section 1.4) provides the necessary data for this: large sets of orthologous sequences. This is useful for several reasons. First of all orthologs are homologs, therefore the very time-consuming search for homologs, which is the basis

<table>
<thead>
<tr>
<th></th>
<th>AAA</th>
<th>AAG</th>
<th>AAC</th>
<th>AAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>0.802</td>
<td>0.112</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td>AAG</td>
<td>0.127</td>
<td>0.826</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>AAC</td>
<td>0.003</td>
<td>0.002</td>
<td>0.811</td>
<td>0.119</td>
</tr>
<tr>
<td>AAT</td>
<td>0.004</td>
<td>0.002</td>
<td>0.105</td>
<td>0.782</td>
</tr>
</tbody>
</table>

Table 4.1: Example of the SynPAM calculation. Top: Substitution matrix at 20 SynPAM, left: probability matrix for synonymous substitutions, right: the synonymous scoring matrix.
for any analysis, can be avoided. Furthermore, orthologs carry the phylogenetic signal of the species tree and all orthologs between two species have in principal almost the same divergence date (given that the lineage sorting of the ancestral speciation event was completed within a short period).

The latter fact will be used for two tests: one of them analyzes the variance of all distance estimates between two species. The second test looks at all averaged pairwise distances and examines their correlation with divergence time. An other test uses the phylogeny feature of orthologs by building gene trees from orthologous groups and comparing them to an accepted species tree.

The SynPAM method presented here is compared to other distance measures based on synonymous substitutions ($dS$). One is the ML method implemented in the codeml program of the PAML software package (Yang, 1997). The other is the counting method by Nei and Gojobori (1986) presented above and implemented in Darwin. These two methods will be referred to as "MLdS" and "NG86", respectively.

### 4.3.1 Variance

The first test based on real sequences uses the sets of orthologs between pairs of species. Because all those sequence pairs are orthologs, they all diverged from common ancestral sequences at the speciation event of the last common ancestral species. If all sequences were infinitely long and if the substitution process and rate at any given time would have been the same for all genes, then all distances between the orthologs should be the same, since they are, in principal, separated by the same period of time. An ideal evolutionary measure would in this case estimate the same distance for all orthologs between two species.

However, since the sequences are only of limited length and since the occurrence of substitutions is a stochastic process, different orthologs will undergo different numbers of substitutions. Varying selection pressure among genes or sites increases these differences. Synonymous substitutions are generally expected to evolve under very little functional constraints, but it cannot be ruled out in all cases (Bielawski et al., 2000; Dunn et al., 2001). For these reasons, even an optimal distance measure would not estimate the same distance for all orthologs between a species pair.

Therefore, the variance of the distance estimates are compared among the different methods and for a range of evolutionary distances. The estimated variance is made up of two parts: the variance inherent to the different orthologs for the reasons stated above, as well as an additional variance that stems from the deviation of the evolution-
ary model from the reality.

**Histograms**

Histograms of the estimated distances between all orthologs of two species display the distribution of the estimates. Although the histogram are not suited for a comparison of the variance (since the estimates of the different methods are of different magnitudes), it does allow for a visual comparison of the distributions.

Figure 4.2 displays the histograms of the three distance estimation methods on five different pairs of metazoan species with increasing divergence times. For the most recent of them, the human–mouse divergence, all three methods have a small variance. With increasing evolutionary distance, the variances of the distributions increase, but until the human–chicken divergence, the histograms for all three methods are centered around a mean estimate with relatively few outliers. Starting at the human–frog split, the MLdS histogram becomes flatter and at human–ciona, almost no signal can be detected for any method.

Curiously, the ML estimations at longer distances show an unexpected pattern: Starting at about the human–fish distance, the ML method returns many $dS$ estimates of around 65. This can be seen in Figure 4.3 where the last two histograms are shown again, but with a different domain (0 to 80 $dS$ instead of 0 to 8). Estimates of this magnitude cannot be correct and must be an artifact from the ML way of computing $dS$, because it is impossible to estimate 65 substitutions per site, and certainly not with the relatively low variance suggested by the histograms. But additionally, if an approximate linearity of $dS$ with time is assumed (see discussion below) then given an average $dS$ of approximately 2 for the human–bird split which occurred about 300 million years ago, this would date a sequence pair with a $dS$ value of 65 at 9.75 billion years, about twice the age of the earth.

**Coefficient of Variance**

The variance observed in the histograms can also be quantified. However, since the metrics compared are on different scales, not the variance itself, but the coefficient of variance (CV, computed as the square-root of the variance divided by the mean) is used.

Table 4.2 lists the CVs for several pairs of species (human vs. another metazoan species) as well as the number of estimates that had to be excluded because at least
one of the methods was not able to return useful estimates. Sequence pairs with less than 100 aligned synonymous codons were also excluded since such short sequences yield unreliable estimates. Further criteria for excluding an estimate was a SynPAM value of 1000 (the maximum distance possible in the current implementation, corresponding to 10 substitutions per synonymous codon). For the \( dS \) metrics a cut-off of 15 was chosen, corresponding approximately to the same amount of evolution. If a sequence pair yielded an invalid estimate by any one of the methods, it was excluded from the analysis for all methods. This way it is ensured that the comparison is performed on the same set of sequences for all methods.

It can be seen that SynPAM estimates consistently have the lowest variance up to the human–frog distance. From the human–fish divergence, NG86 has the least amount of variation. However, many sequences had to be excluded because the MLdS and NG86 were not computable, leaving only those estimates that were within the range of the NG86 and MLdS computation. This biases those results and make the variance artificially smaller.

<table>
<thead>
<tr>
<th>Species</th>
<th>N</th>
<th>CV</th>
<th>excluded because of</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SynPAM</td>
<td>MLdS</td>
<td>NG86</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>19930</td>
<td>1.40</td>
<td>2.27</td>
</tr>
<tr>
<td>Macaque</td>
<td>16584</td>
<td>0.59</td>
<td>1.19</td>
</tr>
<tr>
<td>Mouse</td>
<td>15057</td>
<td>0.29</td>
<td>0.58</td>
</tr>
<tr>
<td>Opossum</td>
<td>12147</td>
<td>0.34</td>
<td>0.83</td>
</tr>
<tr>
<td>Platypus</td>
<td>6126</td>
<td>0.34</td>
<td>0.86</td>
</tr>
<tr>
<td>Chicken</td>
<td>7757</td>
<td>0.34</td>
<td>0.83</td>
</tr>
<tr>
<td>Frog</td>
<td>3342</td>
<td>0.37</td>
<td>0.79</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>1698</td>
<td>0.45</td>
<td>0.67</td>
</tr>
<tr>
<td>Ciona</td>
<td>54</td>
<td>0.78</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 4.2: Coefficients of variance for synonymous distances for orthologs between human and other metazoan species. The last four columns indicate the number of sequences that caused invalid estimates and were therefore excluded from the analysis. Numbers in bold highlight the smallest CV in each row.
Figure 4.2: Histograms of distance estimates for five pairs of metazoan species. The domain is 0 to 500 for SynPAM and 0 to 8 for the two $dS$ variants.
Figure 4.3: Histograms of distance estimates for two pairs of metazoan species, showing the unexpected accumulation of MLdS estimates centered about 65 for longer evolutionary distances. Unlike in Figure 4.2, here the domain for the $dS$ estimates is 0 to 80.
4.3.2 Phylogenetic Accuracy

A small variance, however, is not enough to conclude the usefulness of a method. It is not difficult to construct methods with very small variance but very inaccurate results. Therefore, another test is necessary to determine the accuracy of the estimated distance.

As mentioned above, the distance estimates cannot be compared to the true distances, because those are not known. The pairwise distances between genes, however, can be used to build gene trees and those can be compared to the correct tree (if it is known). If we use only orthologous genes then there is at most one sequence per organism and therefore the gene tree should have the same topology as the species tree. Thus we can estimate pairwise distances for many orthologous groups and build the least-squares distance tree for each group and for each method and compare the resulting gene trees to the correct (species) tree. The fraction of correctly reconstructed gene trees is an indication of how well a method is able to accurately measure distances: results that are biased would not be useful for reconstructing the correct phylogeny.

The species tree can be inferred on the basis of full genomes and also additional information (e.g. from the fossil record) and is therefore much more reliable than a single gene tree. For the metazoan set of full genomes used in this analysis, there is only one disputed branching in the phylogeny – whether Rodentia (mouse and rat of the genomes used here) or Laurasiatheria (dog and cow) are closer to Primates. The trees based on thousands of orthologs from complete genomes provide very strong evidence for Rodentia being the outgroup (Cannarozzi et al. (2007), see Chapter 5). But since this is contradicted by other analyses (e.g. Kriegs et al. (2006)), gene trees for both of the two hypotheses were considered to be correct for this study. One of the two species trees is shown in Figure 4.4, the other tree can be obtained by exchanging the (dog,cow) subtree with the (mouse,rat) subtree.

The tree was divided into subtrees covering increasing evolutionary distances. The roots of these subtrees are labeled T1 to T10. For each subtree and all OMA groups with a member in all species of the subtree, the gene trees were created for each of the tested distance methods using the MinSquareTree function of Darwin. Unlike in the previous test, all distances have been used without upper limits on the estimates. The resulting trees were compared to the species (sub-)tree and the number of correctly reconstructed branches (also called "splits") of the trees were counted. The trees were always considered as unrooted, therefore each branch divides a tree into two "half trees". These half trees were viewed only as sets of species, the structures inside the two individual trees was not considered when the correctness of a branch was
4.3. Results

The results are given in Table 4.3 and are visualized in Figure 4.5. For each subtree T1 to T10 the number of OMA groups with members in all species of the subtree as well as the averages of the longest PAM distance found in the distance matrices used to construct the respective subtrees is presented. This measure was used for the x-axis in the plot of the result and gives an indication of the evolutionary distances in the subtrees.

For all subtrees the CodonPAM distances result in the highest percentage of correctly reconstructed splits. A similar observation was made by Ren et al. (2005) in which codon models were found to outperform amino acid and DNA models for ML tree con-

![Figure 4.4: One of the reference species trees used to evaluate the gene trees based on the distance estimates from several methods.](image-url)
inside the mammals (up to T4), the synonymous measures perform clearly better than the amino acid based PAM distances, but beyond the human–frog split (T6), PAM outperforms synonymous measures. This is a strong indication that for shorter distances, the codon substitutions carry a stronger phylogenetic signal than the amino acid replacements, but whereas the amino acid signals remains almost constant with increasing distance, the synonymous substitutions start to become unreliable. The CodonPAM method benefits from combining the amino acid level information with the DNA signal.

Among the synonymous substitution based measures, the difference is very small for the smallest trees, but from T3 on (human–opossum and beyond), SynPAM always produces the highest number of correct splits. The MLdS performs clearly worse, but is still competitive. However, the shortcomings of the NG86 version of $dS$ have a too strong effect outside the range of the mammals.

### 4.3.3 Linearity with Time

Since the first molecular evolution studies, it has been suggested that mutations on the molecular level occur in an almost clock-like manner (Zuckerkandl and Pauling, 1965).

<table>
<thead>
<tr>
<th>Subtree</th>
<th>N</th>
<th>max. PAM</th>
<th>PAM</th>
<th>CodonPAM</th>
<th>SynPAM</th>
<th>MLdS</th>
<th>NG86</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>9959</td>
<td>13.7</td>
<td>87.3</td>
<td>98.5</td>
<td>96.2</td>
<td>95.3</td>
<td>96.3</td>
</tr>
<tr>
<td>T2</td>
<td>8268</td>
<td>17.5</td>
<td>81.8</td>
<td>93.5</td>
<td>90.7</td>
<td>91.3</td>
<td>91.0</td>
</tr>
<tr>
<td>T3</td>
<td>6946</td>
<td>24.3</td>
<td>76.7</td>
<td>89.0</td>
<td>83.0</td>
<td>79.8</td>
<td>75.3</td>
</tr>
<tr>
<td>T4</td>
<td>3525</td>
<td>26.3</td>
<td>78.4</td>
<td>89.4</td>
<td>82.1</td>
<td>78.1</td>
<td>66.1</td>
</tr>
<tr>
<td>T5</td>
<td>2787</td>
<td>28.4</td>
<td>74.4</td>
<td>86.7</td>
<td>78.7</td>
<td>73.3</td>
<td>49.1</td>
</tr>
<tr>
<td>T6</td>
<td>1947</td>
<td>32.9</td>
<td>72.2</td>
<td>85.2</td>
<td>74.2</td>
<td>61.4</td>
<td>10.6</td>
</tr>
<tr>
<td>T7</td>
<td>1027</td>
<td>39.6</td>
<td>73.2</td>
<td>84.4</td>
<td>67.1</td>
<td>44.1</td>
<td>0.8</td>
</tr>
<tr>
<td>T8</td>
<td>233</td>
<td>60.9</td>
<td>70.4</td>
<td>84.4</td>
<td>58.6</td>
<td>41.6</td>
<td>0.0</td>
</tr>
<tr>
<td>T9</td>
<td>126</td>
<td>62.5</td>
<td>71.5</td>
<td>85.3</td>
<td>56.7</td>
<td>40.2</td>
<td>0.0</td>
</tr>
<tr>
<td>T10</td>
<td>84</td>
<td>65.6</td>
<td>69.3</td>
<td>84.6</td>
<td>51.8</td>
<td>39.5</td>
<td>0.0</td>
</tr>
<tr>
<td>All</td>
<td>34902</td>
<td></td>
<td>78.1</td>
<td>90.1</td>
<td>82.7</td>
<td>77.6</td>
<td>62.9</td>
</tr>
</tbody>
</table>

Table 4.3: Percentages of correctly reconstructed splits by five distance methods on subtrees of the tree of Figure 4.4. For subtrees T3–T10, two alternative topologies were accepted (see text). N: number of OMA groups covering the subtree, max. PAM: longest PAM distance encountered in the subtree.
If most of the mutations are neutral, then also the number of substitutions that get fixed in the population should accumulate clock-like and thus increase linearly with time (Kimura and Ohta, 1972). Amino acid substitutions often influence or change the function of a protein and the substitution rate can be very different among genes, sites or lineages. Synonymous substitutions, however, evolve practically without any functional constraints (Miyata et al., 1980). It can therefore be expected, that distances based on synonymous substitutions are almost proportional to time.

For these reasons, it makes sense to examine the range in which distance estimates increase linearly with time by plotting distance estimates versus divergence time estimates. Instead of using the estimates from every sequence pair separately, all estimates between the same two species are averaged for SynPAM and the two versions of $dS$. This reduces the variance and allows for reasonable estimates over distances where the variances of single gene distances are too large. The estimates were not

![Tree Reconstruction](image)

Figure 4.5: Percentages of correctly reconstructed splits of Table 4.3 as a function of maximum PAM distance in the subtree.
filtered, all valid results were used to compute the averages. Excluded were only sequences, where codeml failed to return a result, where the fraction of conserved synonymous substitutions was less than 0.25 and thus the logarithm for the computation of NG86 could not be taken, or where the SynPAM optimization stopped at 1000, the maximum. Invalid results were only excluded for the respective method.

For correlating distance estimates with time, the divergence times of the species need to be known. Divergence times are typically inferred from the fossil record. The time estimates for this test stem from the review by Benton and Donoghue (2007). In that article, the authors explain the difficulties of fossil calibration and give comprehensive minimum and maximum constraints for many major divergences of metazoan species. These divergence times from the fossil record are independent from molecular analysis and therefore allow for a fair comparison of molecular distance estimates. For the linearity analysis presented here, the averages of the lower and upper bounds of the time estimates were used. Divergence times are often given in "million years ago" (MYA).

The plot of distance estimates versus fossil times is shown in Figure 4.6. Each dot corresponds to one average from all orthologs between a pair of species. Interpolation lines were fit to the points with the constraint that they go through the origin (since in zero time, zero evolution should happen). Furthermore, the line was allowed to have one change in slope. The point where it changes was determined for each method separately by minimizing the least-squares fit between the points and the lines. To approximate the points by a piecewise linear function seems justified since it decreases the least-squares residuals by a factor 9.3 for SynPAM, 9.2 for MLdS and 2.0 for NG86.

The first part of the line for SynPAM goes up to 418 MYA (tetrapod–fish split), then the line becomes steeper. The MLdS fits the linear part up to 321 MYA (mammal–bird split), after that point the artifactual estimates of around 65 (discussed above) start to influence the result. The NG86 maintains the linearity up to 340 MYA (human–frog divergence), then the curve starts to flatten, probably due to the effect that only the less divergent sequences produce valid results at longer distances.

The slopes of the first parts of the interpolation lines suggest rates of 0.41 SynPAM/MY, 0.015 MLdS/MY and 0.0058 NG86/MY. It is remarkable that there is a factor of almost 3 between the two rates of the dS methods, emphasizing again the importance of method and assumptions for the estimation of dS.

As an additional analysis, the coefficients of variance are computed for all points corresponding to a particular time value on the x-axis. This analysis is similar to the one in Section 4.3.1, but there the CV of all estimates between two species were investigated. Here, the variance of the averages among species pairs with the same divergence time
is analyzed. These are only little influenced by the variance of single estimates, but it should reflect the consistency of the methods when estimating distances that have the same divergence time.

The results in Table 4.4 show that for all time ranges of less than 340 million years, the SynPAM estimates have the lowest CV. At the longer distances the pattern is unclear with different methods having lowest CVs. However, since at those distances all methods are outside the range where they produce estimates that proportional to time, the CV is only of theoretical interest.

Figure 4.6: Linearity of distance estimates with time. Each point corresponds to one average from all orthologs between a pair of species. The lines are least-squares fits of lines through the origin allowing one change of slope.
4.4 Discussion and Conclusions

SynPAM has been introduced as a new method for estimating evolutionary distance based on synonymous codon substitutions. It is a novel idea to use a Markov model and a maximum likelihood framework to estimate the amount of synonymous change based on differing synonymous codons. Additionally, the use of empirical codon substitution matrices makes the estimation relatively fast and allows for the incorporation of many features of codon evolution that are not treated in parametric models.

Various tests demonstrated the competitiveness of SynPAM: it was shown to result in reliable estimates with a lower variance than other methods. Distances estimated by SynPAM allow for better tree reconstruction over a longer time range than any other synonymous substitution based methods. Furthermore, for distances inside mammals, it was shown that the information from synonymous substitutions carries more phylogenetic signal than the amino acid based PAM distances. Finally, it was demonstrated that SynPAM estimates increase linearly with time within vertebrates, which is for longer distances than the other methods it was compared to.

As the empirical codon substitution matrix was built from vertebrate DNA, its usable range is for the vertebrates. Tests with plants, yeasts and bacteria have not been successful, among other factors because of very high substitution rates which quickly lead to saturation. Substitution matrices constructed from groups of closely related species could improve the performance for these subgroups. The creation of matrices for only pairs of species would allow for the incorporation of species-specific substitution pat-

<table>
<thead>
<tr>
<th>MYA split</th>
<th>N</th>
<th>SynPAM</th>
<th>MLdS</th>
<th>NG86</th>
</tr>
</thead>
<tbody>
<tr>
<td>81 Primates–Glires</td>
<td>14</td>
<td>0.156</td>
<td>0.166</td>
<td>0.165</td>
</tr>
<tr>
<td>104 Euarchopeae–other Eutheria</td>
<td>64</td>
<td>0.239</td>
<td>0.299</td>
<td>0.256</td>
</tr>
<tr>
<td>132 Eutheria–Marsupialia</td>
<td>14</td>
<td>0.0709</td>
<td>0.137</td>
<td>0.089</td>
</tr>
<tr>
<td>177 Metatheria–Monotremata</td>
<td>15</td>
<td>0.0582</td>
<td>0.180</td>
<td>0.0772</td>
</tr>
<tr>
<td>321 Mammalia–Sauropsida</td>
<td>16</td>
<td>0.0425</td>
<td>0.202</td>
<td>0.0538</td>
</tr>
<tr>
<td>340 Amniota–Amphibia</td>
<td>17</td>
<td>0.0476</td>
<td>0.132</td>
<td><strong>0.032</strong></td>
</tr>
<tr>
<td>419 Tetrapoda–Teleostei</td>
<td>90</td>
<td><strong>0.0573</strong></td>
<td>0.0583</td>
<td>0.11</td>
</tr>
<tr>
<td>542 Vertebrata–Arthropoda</td>
<td>115</td>
<td>0.0828</td>
<td><strong>0.0803</strong></td>
<td>0.0993</td>
</tr>
</tbody>
</table>

Table 4.4: CVs of averaged estimates for species pairs with same divergence times. N: number of species pair for a given divergence time. The numbers in bold indicate the lowest CV in each row.
terns without the need to adjust parameters to the actual sequences.

Because synonymous substitutions are evolving under fewer functional constraints than amino acid substitutions, they are expected to occur more regularly. Therefore they are of interest for use in molecular dating, in which clock-like distance measures and known divergence times are used to extrapolate actual time estimates for unknown divergences. Since the SynPAM method of measuring synonymous distances has advantages over existing distance metrics, the use of SynPAM for molecular dating in the realm of the vertebrates is promising.
Part II

Models of Evolution Applied to Mammalian Phylogenies
Overview

Models of molecular evolution are a valuable tool for determining phylogenetic relationships. Models based on changes in gene and protein sequences have clear advantages over traditional methods based on morphology by using an explicit probabilistic model of the evolutionary changes but also by being able to use very large amounts of data. Particularly interesting for this type of analysis is the phylogeny of the mammals because they are our closest relatives, but also because the diversification of the mammals is important for the understanding of the recent history of the earth. Mammalian genomes contain billions of base pairs which can be used in evolutionary analyses and in recent years many mammalian genomes have been sequenced, leading to an unprecedented amount of data for phylogenetic studies of mammals.

The second part of this thesis will first present the results of applying several phylogenetic methods to the problem of the phylogenetic problem of Primates (represented by human and two apes), Carnivora (the dog) and Rodentia (mouse and rat). The methods are applied to complete genomes and with this large amount of data come to very strong conclusions. However, by extending the analysis to more mammalian genomes, it will be shown that significant results can be obtained for several topologies, some of them mutually exclusive. Simulations will then be employed to demonstrate that differences between the evolutionary model and the model used for creating the data can lead to strongly biased results which could explain the contradicting support observed before. Finally, an approach based on the analysis of using different outgroups will be presented, which can help to identify biased reconstructions.
Chapter 5

The Relative Positions of Human, Dog and Mouse in the Mammalian Tree

This chapter presents the results of a study of the evolutionary relationships of human, mouse and dog using the data from completely sequenced genomes. Several types of methods (distance trees, parsimony and ML trees) and various variants of the methods are applied to thousands of orthologous genes, leading in the most cases to a strong support for a sister relationship of human and dog to the exclusion of mouse.

5.1 Introduction

Understanding the phylogenetic relationships between organisms are fundamental to almost any evolutionary study. Of particular interest are the diversification of the mammals, since they are our closest relatives, but also since understanding mammalian evolution helps solving questions relating to the biogeographic history of the world. However, some relationships still need further examination before being considered as conclusively resolved. Paleontological data show a sudden radiation of mammals towards the end of the Cretaceous, 145.5–65.5 million years ago (Archibald, 2003). Multiple speciations during a short period of time can lead to many morphologically distinct groups, with very few characters in extant species linking them to other groups. This makes it difficult for morphological analyses to correctly identify the relationships among these groups. However, the correct succession of events during the radiation might be resolved using molecular data, since there may be enough data in the genomes to produce results with high confidence levels.

This study focuses on the relative positions of Primates, Carnivora and Rodentia. This
is an important question in mammalian phylogeny as it has implications in medical research as well as for biogeography. It is a difficult problem, as the branch separating alternative topologies is short and relatively deep in the tree and with lineages evolving at different rates. On the other hand, representative species for this problem (human, dog and mouse) have been sequenced with the genomic sequence being of high quality. Thus, a large amount of reliable data is available for molecular studies of this problem.

In order to determine the rooted tree of three species, an outgroup is required, since most phylogenetic methods produce unrooted trees. The position of the branch leading to the outgroup determines the root of the three ingroup species. The closest outgroup to the placental mammals for which the genomic sequence is also available, is the opossum. The unrooted triplet and the three possible rooted quartets are shown in Figure 5.1. The hypotheses are either Carnivora forming a clade with Primates to the exclusion of Rodentia (Tree A) or a clade of Primates and Rodentia (called Supraprimates or Euarchontoglires) to the exclusion of Carnivora (Tree B). Tree C is not accepted as a realistic possibility.

Several molecular studies exist in support of either of the two hypotheses, a selection of relevant articles is given in Table 5.1. In 2007, *A Phylogenomic Study of Human, Dog and Mouse* (Cannarozzi et al., 2007) was published, where several phylogenetic methods using data from full genomes provided strong support for Tree A. In this chapter, a re-analysis of that study is presented. Since the original article was published, several genomes have been updated, leading to more orthologous sequences of better quality. Additionally, some of the methods were refined and extended. The effect of these changes will be discussed in Section 5.3 and the following chapters.

### 5.2 Methods and Results

#### 5.2.1 Distance Methods

Distance trees were computed for eight mammalian species for which complete genomes were available: human, chimpanzee, rhesus macaque, mouse, rat, cow, dog and opossum. In the June 20, 2008 version of the mammalian set of OMA, there were 7640 groups with a member in all eight species. Four different methods were used to estimate distances: PAM (amino acid substitution), CodonPAM (codon substitutions), SynPAM (only synonymous codon substitutions) and MLdS (synonymous substitutions). For each pair of species, the distances between all orthologs were estimated. Groups
had to be discarded when all sequences were the same (and thus not informative) or when estimates for DNA based methods were missing (typically because of missing DNA). The matrices of the averages and the variances of all estimates between two species were used as input to the MinSquareTree function of Darwin.

Table 5.2 lists the resulting normalized least-squares residues (MST) from fitting the pairwise distances to the three tested topologies. The topologies were those of Figure 5.1 with the constraints, that rat must be with mouse, cow with dog and chimp and macaque with human. For all three methods, Tree A, human closest to dog, resulted in the lowest MST values. In order to illustrate the relative lengths of the branches, the optimal tree created from the PAM distances is shown in Figure 5.2. Trees from the other distance methods have the same topology and very similar relative branch lengths, but are not shown here.

Figure 5.1: Unrooted trees of three species (human, dog and mouse) display no information of the speciation order (center triplet). Only the use of an outgroup (opossum) places the root on one of the three branches (labeled A, B and C), giving three possible rooted trees.
Although the MST for tree A was always a factor 3.8 to 4.8 smaller than for any alternative topologies, it is difficult to assess the significance of these differences (Goldman et al., 2000). Therefore, bootstrapping was performed to test the robustness of the support of the topology by sampling from the groups. The sampling was done with replacement, in order to compute the bootstrap replicates from the same number of groups as the original tree. For all four methods, 1000 bootstrap replicates were created and all resulted in tree A. This best possible support of 100%, is most likely due to the large number of groups used. With many groups, the variances of the estimates become very small and the tree reconstruction is very robust.

<table>
<thead>
<tr>
<th>Article</th>
<th>Result</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Janke et al. (1994)</td>
<td>A</td>
<td>ML analysis of 7 mt genomes</td>
</tr>
<tr>
<td>Shoshani and McKenna (1998)</td>
<td>B</td>
<td>based on 260 morphological characters</td>
</tr>
<tr>
<td>Reyes et al. (2000)</td>
<td>A</td>
<td>ML and ME trees of 35 mt genomes</td>
</tr>
<tr>
<td>Murphy et al. (2001a,b)</td>
<td>B</td>
<td>Bayesian analysis of 44 complete mtDNA</td>
</tr>
<tr>
<td>Lin et al. (2002)</td>
<td>B</td>
<td>ML analysis of 32 mt genomes</td>
</tr>
<tr>
<td>Arnason et al. (2002a)</td>
<td>A</td>
<td>ML analysis of 60 mt genomes</td>
</tr>
<tr>
<td>Misawa and Janke (2003)</td>
<td>A</td>
<td>NJ trees of nuclear and mt genes of 8 taxa</td>
</tr>
<tr>
<td>Amrine-Madsen et al. (2003)</td>
<td>B</td>
<td>ML analysis of nuclear and mt genes of up to 63 species</td>
</tr>
<tr>
<td>Thomas et al. (2003)</td>
<td>B</td>
<td>genomic regions from 12 taxa, supported by insertions of transposable elements</td>
</tr>
<tr>
<td>Reyes et al. (2004)</td>
<td>B</td>
<td>70 complete mtDNA</td>
</tr>
<tr>
<td>Jorgensen et al. (2005)</td>
<td>A</td>
<td>ML analysis of 865 nuclear genes of 3 mammals and fugu</td>
</tr>
<tr>
<td>Bashir et al. (2005)</td>
<td>B</td>
<td>SINE and LINE insertions in 28 taxa</td>
</tr>
<tr>
<td>Kriegs et al. (2006)</td>
<td>B</td>
<td>insertions of retroposed elements in 17 taxa</td>
</tr>
<tr>
<td>Kullberg et al. (2006)</td>
<td>A</td>
<td>ML analysis of housekeeping genes of 26 taxa</td>
</tr>
<tr>
<td>Huttley et al. (2007)</td>
<td>A</td>
<td>ML analysis of 6 complete genomes</td>
</tr>
</tbody>
</table>

Table 5.1: Overview of relevant articles about the phylogenetic positions of human, dog and mouse. Tree A (human-dog clade) and B (human-mouse clade) refer to the trees in Figure 5.1. mt: mitochondrial, ML: maximum likelihood, ME: minimum evolution, NJ: neighbor joining, MP: maximum parsimony.
5.2. Methods and Results

5.2.2 Parsimony

MSAs for the parsimony analysis were created for four species (human, dog, mouse and opossum) using the Mafft alignment program (Katoh et al., 2002). There were 10,839 OMA groups with sequences in all four species with a total of 6.7 million aligned positions. In order to reduce artifacts from truncated genes (fragments), only the part of the alignments between the first position with four aligned amino acids and the last position with four aligned amino acids were considered, gaps at the beginning or end of the alignment were ignored.

To decide which of the three quartets is the most parsimonious one, only those align-

<table>
<thead>
<tr>
<th>Method</th>
<th>N</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM</td>
<td>7604</td>
<td>1.27</td>
<td>4.89</td>
<td>5.05</td>
</tr>
<tr>
<td>CodonPAM</td>
<td>7152</td>
<td>2.27</td>
<td>9.86</td>
<td>10.2</td>
</tr>
<tr>
<td>SynPAM</td>
<td>6866</td>
<td>2.73</td>
<td>13.2</td>
<td>13.6</td>
</tr>
<tr>
<td>dS</td>
<td>6599</td>
<td>0.559</td>
<td>2.12</td>
<td>2.15</td>
</tr>
</tbody>
</table>

Table 5.2: Normalized least-squares residues for the three tested topologies for distance trees using four different methods of distance estimation. N: number of orthologous groups used. A, B and C: MST quality for the three topologies of Figure 5.1 with additional taxa (see text).

Figure 5.2: PAM distance tree of eight mammals. Branch labels indicate PAM distances.
ment positions at which two species share one character and the other two share another character (2-2 cases) are informative. (Positions where all species share the same characters or have all four different characters as well as 3-1 splits are uninformative. The 2-1-1 splits have parsimony costs of 2 for all three topologies and are also uninformative.)

**Influence of gaps on parsimony**  Gap regions are difficult to align and thus the alignments with many gaps might be more error-prone. Therefore, the parsimony analysis was conducted once with counting gaps as characters and once with ignoring all positions containing gaps. In order to further investigate the influence of gaps on the alignments and thus on the parsimony results, alignments were excluded from the analysis based on certain maximum percentages of gaps.

These results are summarized in the top part of Table 5.3. The rows of the table correspond to the maximal allowed gap percentages. The last row corresponds to 100%, meaning that all 10,839 alignments are counted. Independent of how the gaps are counted, topology A (human-dog) always has the highest support. The support for A is stronger, if the gapped positions are not counted (right part of the table). Furthermore, the support for A also becomes stronger, the stricter the filtering based on gap

<table>
<thead>
<tr>
<th>Gaps</th>
<th>Groups</th>
<th>With gaps</th>
<th></th>
<th></th>
<th>Without gaps</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>A (%)</td>
<td>B (%)</td>
<td>p-val</td>
<td>N</td>
<td>A (%)</td>
</tr>
<tr>
<td>0</td>
<td>747</td>
<td>2,179</td>
<td>44.7</td>
<td>32.5</td>
<td>4e-11</td>
<td>2,179</td>
<td>44.7</td>
</tr>
<tr>
<td>5</td>
<td>7,482</td>
<td>89,451</td>
<td>41.0</td>
<td>34.7</td>
<td>3e-102</td>
<td>81,386</td>
<td>42.5</td>
</tr>
<tr>
<td>10</td>
<td>9,736</td>
<td>125,069</td>
<td>40.3</td>
<td>35.5</td>
<td>2e-85</td>
<td>108,199</td>
<td>42.4</td>
</tr>
<tr>
<td>20</td>
<td>10,729</td>
<td>144,735</td>
<td>39.8</td>
<td>36.1</td>
<td>9e-57</td>
<td>119,192</td>
<td>42.4</td>
</tr>
<tr>
<td>100</td>
<td>10,839</td>
<td>146,984</td>
<td>39.6</td>
<td>36.2</td>
<td>6e-50</td>
<td>120,386</td>
<td>42.4</td>
</tr>
<tr>
<td>slow</td>
<td>3,613</td>
<td>21,040</td>
<td>34.3</td>
<td>43.6</td>
<td>5e-22</td>
<td>13,881</td>
<td>41.7</td>
</tr>
<tr>
<td>medium</td>
<td>3,613</td>
<td>49,038</td>
<td>37.8</td>
<td>37.7</td>
<td>0.4</td>
<td>39,333</td>
<td>41.8</td>
</tr>
<tr>
<td>fast</td>
<td>3,613</td>
<td>76,906</td>
<td>42.2</td>
<td>33.2</td>
<td>5e-179</td>
<td>67,172</td>
<td>42.9</td>
</tr>
</tbody>
</table>

Table 5.3: Result of parsimony analysis on four species. Gaps: maximal percentage of gaps allowed in the MSA, Groups: number of orthologous groups fulfilling this criteria, N: number of informative characters, A and B: percentages of characters supporting the trees of Figure 5.1, p-val: probability that the number of characters supporting A and B come from a binomial distribution with equal probabilities. Bottom part: the results after splitting the alignments into slowly, medium and fast evolving proteins. Numbers in bold indicate significantly best supported topologies.
percentages is applied.

It may seem that the relative number of characters supporting the different topologies are very close, support for human-dog ranges from 39.6% to 44.7% whereas support for human-mouse ranges from 32.5% to 36.2%. However, since a very large number of characters is used, these differences are very significant (see the p-val columns).

Influence of evolutionary rates on parsimony  If the evolutionary distances are long, then parsimony will underestimate the number of changes along a branch. This could bias the results and lead to incorrect phylogenetic reconstructions. To investigate this effect, the alignments were split in three parts, containing the slowly, medium and fast evolving proteins. Speed of evolution was defined as the sum of branch lengths (in PAM) of the quartet tree. This analysis was performed on all alignments, thus 3,613 alignments were available for each rate category. The result is shown in the bottom part of Table 5.3.

If gaps are ignored (right half of the table), then the influence of the evolutionary rate is relatively small. In slowly evolving proteins 41.7% of the informative characters support human-dog, whereas for fast evolving proteins, the support is 42.9%. However, if the gaps are counted as characters, the result changes: 43.6% of the characters in slowly evolving proteins support human-mouse, while in fast groups human-dog is supported by 42.2% of the characters. For medium rates, the support is not significantly different between topologies A and B.

In summary, positions with gaps strongly favor the human-mouse topology, particularly in slowly evolving proteins. Several possible explanations have to be considered: It could be that the evolution of indels is much less affected by any kind of bias that would mislead the results from amino acid substitution analysis. It is also possible that indels are less likely to be homoplasies (i.e. to come from two independent events). This would mean that the gaps are more likely to express the true phylogenetic signal while the amino acid substitutions can be biased towards a wrong topology. On the other hand, the indels supporting human-mouse could also be the result of a bias A likely form of bias are the differing genome qualities. The human and mouse genomes are of a better quality than the opossum and dog genomes. It is therefore possible that sequencing or assembling mistakes in the two latter genomes lead to artificial gaps supporting a human-mouse topology.

There is some evidence supporting this hypothesis: Gaps supporting human-dog have an average length of 3.29 amino acids when they are in human and dog and average 3.25 amino acids when they are in mouse and opossum. Gaps supporting human-
mouse, however, average 2.93 amino acids when in human and mouse, but 6.69 amino acids when in dog and opossum. The indels common to dog and opossum are thus more than twice as long as those common to human and mouse. This could be an indication that some these gaps are not the result of normal evolutionary events, but that they are artifacts, introduced by sequencing or annotation errors.

Further evidence came from manual inspection of alignments supporting human-mouse. It can be observed that gaps common to dog and opossum occur often in regions that are repetitive. An example looks like this:

- Human: LGGRSAASSSASASSSVSFSPGGGGGAAACCRGMSWTPAEHTNALIAVGNLRLEAARYQQLEGAG
- Mouse: LGGRGAASSSASASSSVSFSPGGGGGAACCRGMSWTPAEHTNALIAVGNLRLEAARYQQLEGAG
- Dog: LGGRSAASSSASASSSVSFSPG______________CRGMSWTPAEHTNALIAVGNLRLEAARYQQLEGAG
- Opossum: LGARSAASSS_____________________________TNALIAVGNLRLEAARYQQLEGAG

Repetitive regions, also called low-complexity regions, are more difficult to sequence and assemble. It is plausible that sequencing errors common to the low-quality genomes dog and opossum occur predominantly in such regions. Therefore, an attempt was made to identify repetitive regions and to remove them from the parsimony analysis. The following heuristic criteria were defined for a repetitive region: 5 consecutive MSA position with only 1 type of amino acids, 10 consecutive positions with at most 3 different amino acids or 20 consecutive positions with at most 4 different amino acids.

The result of applying this heuristic is shown in Table 5.4. The medium rate proteins now clearly support human-dog, and the human-mouse support from the slowly evolving proteins became weaker. This demonstrates that at least part of the gap support for human-mouse stems from correlated sequencing errors in dog and opossum. There could also be more such cases that were not detected by the rather simplistic heuristic, but sequencing errors can also occur at positions that are non-repetitive and thus less

| Rate   | Groups | With gaps | | Without gaps |
|--------|--------|-----------| |------------|
|        |        | N  A (%)  B (%) p-val     | N  A (%)  B (%) p-val |
| slow   | 3,613  | 19,363    | 36.8  **39.7** 3e-06 | 13,827 41.7 36.2 7e-14 |
| medium | 3,613  | 47,241    | **38.9** 36.0 3e-14 | 39,265 41.8 33.9 2e-72 |
| fast   | 3,613  | 76,264    | **42.5** 32.8 9e-210 | 67,143 42.9 32.6 8e-207 |

Table 5.4: Result of parsimony analysis as a function of evolutionary rate after removing repetitive regions. N: number of informative characters, A and B: percentages of characters supporting the trees of Figure 5.1, p-val: probability that the number of characters supporting A and B come from a binomial distribution with equal probabilities.
simple to identify. However, there is still a significant support for human-mouse from the slowly evolving groups. Also, it is very difficult to distinguish between true gaps caused by evolution and artificial gaps caused by low sequence quality.

In summary, except the subset of slowly evolving proteins where the gaps support human-mouse, the parsimony supports the human-dog topology with very high confidence. The analysis, however, demonstrates also the influence of possible biases in the data. Some issues may be detected and taken care of, but there might exist several biases that go unnoticed.

5.2.3 Likelihood Trees

The most complex, but also computationally most expensive phylogenetic method is maximum-likelihood (ML) analysis. Unlike the parsimony analysis, in which the tree that minimizes the number of necessary changes to explain the observed sequences is sought, the ML analysis is based on a Markov model of evolution. The likelihood of aligned sequences can be computed given an evolutionary model and a tree relating the sequences. Thus, the tree that maximizes the likelihood can be found.

The 10,839 MSAs from the parsimony analysis were also used for the ML analysis. All positions containing gaps were removed from the alignments since the likelihood computation for gaps is not well defined. For each group, the logarithms of the likelihoods (log-likelihoods) of the tree possible topologies were computed using the PhyML program (Guindon and Gascuel, 2003) under the JTT model (Jones et al., 1992) of amino acid evolution. The likelihoods were computed under the assumption of no rate variation among sites as well as with assuming a gamma distribution with four discrete rate classes (Yang, 1994).

The 10,839 log-likelihoods for the each topology were summed up, allowing for a comparison of the three topologies with the trees for each group having optimal branch lengths. The sum of the log-likelihoods as well as the number of individual groups supporting the three hypotheses are shown in Table 5.5.

The results show that under the simple model of constant rates among sites, the support for human-dog is very strong. It is difficult to assess the exact significance of the likelihood ratio for different topologies, but for a log-likelihood difference of 3,501, there is no doubt of the significance. Interestingly, if a gamma distribution is applied to the rates among sites, the result changes dramatically. Under the “JTT+Γ” model, the likelihood method supports human-mouse. The support is weaker, but with a log-likelihood difference of 1,951 still strong. Interestingly is also the distribution of the shape-parameter
α of the gamma-distribution. This parameter was estimated for each alignment separately and the histogram is shown in Figure 5.3. The distribution has a median value of 0.563, which implies a relatively high degree of rate variation among sites.

<table>
<thead>
<tr>
<th>Model</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>JTT</td>
<td>(-28,310,285)</td>
<td>(-28,313,786)</td>
<td>(-28,337,176)</td>
</tr>
<tr>
<td>delta logLs</td>
<td>0</td>
<td>(-3,501)</td>
<td>(-26,890)</td>
</tr>
<tr>
<td>number of groups</td>
<td>(4727.33)</td>
<td>(4080.83)</td>
<td>(2030.83)</td>
</tr>
<tr>
<td>JTT+Γ</td>
<td>(-28,162,786)</td>
<td>(-28,160,836)</td>
<td>(-28,173,694)</td>
</tr>
<tr>
<td>delta logLs</td>
<td>(-1,951)</td>
<td>(0)</td>
<td>(-12,858)</td>
</tr>
<tr>
<td>number of groups</td>
<td>(4239.67)</td>
<td>(4320.67)</td>
<td>(2278.67)</td>
</tr>
</tbody>
</table>

Table 5.5: The results of the ML analysis of the four main species (human, dog, mouse and opossum): sum of the log-likelihoods, the difference to the maximal log-likelihood and the number of groups supporting each topology. (If two or three topologies had the same likelihood, each was counted as 1/2 or 1/3 support, respectively.) A (human-dog), B (human-mouse) and C (mouse-dog) refer to the trees of Figure 5.1.

Figure 5.3: Distribution of α values estimated by phyML from the 10,839 alignments from human, dog, mouse and opossum. 921 values are larger than 3 and are not displayed in this histogram.
These results demonstrate the importance of the evolutionary model. Under one model, it is possible to obtain significant results for one topology and under another model, another (mutually exclusive) topology is significantly more likely.

5.3 Discussion

The original article pointed out the strong statistical support for the human-dog topology which was unambiguously supported by all methods used. This lead to the conclusion that such a difficult phylogenetic problem, with short internal branches, "can only be solved using thousands of genes, which are only available from complete genomes."

The re-analysis of the problem presented here can largely support those findings, however, some differences have to be noted:

- The distance trees fully confirm the original results. Due to the improved genome quality, the number of genes with complete DNA more than doubled. This allowed for even stronger results for the DNA based methods (CodonPAM, SynPAM and $dS$).

- The parsimony analysis lead to very similar results as the original study. But because of changes in the filtering criteria, the result for the slowly evolving genes changed to human-mouse support when gaps are counted as characters. However, there is evidence that this is in part an artifact stemming from bad sequence quality in the dog and opossum genomes.

- The biggest difference can be found in the likelihood analysis. The original study was conducted with a Darwin implementation of the ML tree optimization and did not allow for rate variation among sites at that time. The new analysis was performed using PhyML, and also using a gamma distribution to model the rate variation. Whereas the analysis assuming constant rates among sites supported the original finding, the use of rate variation changed the result to strong support for the human-mouse topology.

These new results indicate that even the very strong statistical support for one topology, that was found in the original study, could be misleading. The choice of method, data and evolutionary model can strongly influence the result. And since the amount of data is very large, almost every result is significant, and can also be in conflict with other significant results. In the following chapters, some of these issues will be investigated and possible solutions will be proposed.
Chapter 6

Strong Support for Conflicting Topologies

Following the study of the phylogeny of human, dog and mouse, a comprehensive reconstruction of all possible quartets of mammals was undertaken. Because the analysis was performed on thousands of orthologous proteins, support for a particular topology often became very significant, even when the fractions of groups supporting different trees were not very different. Interestingly, some of these well supported topologies contradicted each other. The amount of incorrect quartets needed to cause all observed conflicts was estimated and found to be surprisingly large. Simulations and theoretical analysis was used to show that in the presence of model violations, phylogenetic methods can systematically fail and produce incorrectly reconstructed quartets.

6.1 Introduction

As shown in the previous chapter, different phylogenetic methods applied to the same data set can produce conflicting results. All of those results are statistically significant, but since some of them are conflicting with others, it is clear that some result that are significant are also incorrect.

In this chapter, the phenomenon of well supported but conflicting topologies will be investigated. An empirical study on a set of mammalian genomes demonstrates that not only different methods applied to the same data (as in the previous chapter) can produce conflicts, but also that the same method applied to different subsets of the species can yield contradicting results. From a large-scale analysis of 12,650 different quartets of mammals, the amount of conflicts among the different quartets will be assessed.
This will lead to an estimate for a minimum number of mammalian quartets that are incorrect.

In a second part, the possible reason for incorrect but highly supported reconstructions will be explored by means of simulations and theoretical analysis on a simple model. Particularly the effect of model violations on distance estimation and their influence on tree reconstruction will be investigated.

6.2 Assessing the Amount of Conflicting Quartets

6.2.1 About Conflicts

In this section the amount of incorrectly reconstructed quartets among mammals will be estimated. However, many mammalian relationships are still unclear or at least subject to debate and there exists no tree for which a wide-spread consensus can be found. Therefore, it is generally not possible for a given quartet to decide whether it is "right" or "wrong". But when quartets with overlapping species sets are compared to each other, conflicts between them can be detected. When there is a conflict, at least one of the quartets involved has to be incorrect. Therefore, a minimal number of incorrect quartets can be determined, such that all conflicts are explained.

Definition 6.2.1  Given a set of tree topologies \( T_i \), describing the binary, unrooted topology for a species set \( S_i \), and the union \( S \) of all \( S_i \), a situation is called a "conflict", if there exists no topology \( T \) for \( S \), such that every \( T_i \) can be obtained from \( T \) by removing all leaves of \( T \) that are not in \( S_i \).

Any \( T_i \) for only three species, called a triplet, can never cause a conflict, since an unrooted triplet can always be obtained from \( T \) by removing all other leaves. Two quartets \( T_1 \) and \( T_2 \) can only be conflicting, if \( S_1 = S_2 \). Since all quartets in this study have distinct sets of leaves, three quartets for a total of five species are the smallest scenario for which a conflict can occur. Figure 6.1 shows such a situation: The three quartets (i), (ii) and (iii) cannot be put together to a tree of five species, since depending on which two quartets are used, different quintets are implied. This is the most simple possibility for detecting a conflict and thus was selected for the analysis in this section.
6.2.2 Data and Methods

The Mammalia data set of OMA from June 2008 with 25 species and 58,982 orthologous groups was used to systematically and empirically assess the amount of conflicting quartets. With the 25 mammals, 12,650 different quartets can be formed. Since there are so many quartets of species that need to be reconstructed, a fast phylogenomic method must be used. The choice was weighted least-squares trees (as in Section 5.2.1 of the human-dog-mouse study) using the means and variances of all pairwise distances from all groups that have a member in all four species of the quartet. If all six distances in a group were 0, the group was discarded since it contributed no information on the phylogeny of the species. Bootstrapping was then performed by sampling from the groups to assess the significance of the resulting quartets.

For the species quartets, there were between 12,878 groups (for human, chimp, macaque and mouse) and 732 groups (for guinea pig, armadillo, cat and platypus) available. Of

![Image of conflicting quartets]

Figure 6.1: Example of conflicting quartets: the three quartets on the left cannot be put together to a tree of five species. If quartets (i) and (ii) are used, then leaf E can be at the positions shown in tree (iv), whereas if (ii) and (iii) are used, then E can only be at the position shown in (v). The options for E in (iv) and (v) are non-overlapping, thus there must be a conflict among the quartets (i), (ii) and (iii).
the 12,650 quartets, 11,595 had a bootstrap support of 100%, and only for 26 quartets was the highest support less than 50%. This means that for most of the quartets, the resulting topology was very significant (in terms of bootstrap support), mostly because many orthologous groups could be used.

To determine whether a set of trees, here three quartets, contain a conflict, is called the "tree compatibility problem" (Gusfield, 1991). This can be solved by creating a table of binary parsimony characters describing all branches of the trees. Each branch of every tree in the set defines one character by assigning all leaves on one side of the branch state 0 and all leaves on the other side state 1. Leaves not occurring in the tree can be either 0 or 1 and thus are assigned a question mark (?). If there exists a joint assignment of either 0 or 1 for all question marks such that all characters are pairwise compatible, then the set of trees is conflict free. Two characters (i,j) are incompatible, if all four of the possible state pairs — (0,0), (0,1), (1,0) and (1,1) — occur.

Table 6.1 shows an example of a character compatibility table for the trees of Figure 6.1. If columns (i) and (iii) are considered, then character $C$ must be a 1, but if columns (ii) and (iii) are used, then character $C$ can only be a 0. Therefore, there exists no assignment for $C$ that would not cause a conflict.

6.2.3 Results

With 25 species, there are 53,130 possible quintets. For each of them there are five different quartets and every quartet is part of 21 different quintets. In 9,817 of the 53,130 quintets (18.5%) a conflict was detected, involving 9,326 different quartets. However, not all of these quartets are incorrect. A correct quartet can also be part of a conflict as a "witness" against an incorrect other quartet. On the other hand, if two or more quartets of a quintet are incorrect, but in a consistent manner, then no conflict is detected.

It is not possible to determine exactly, which quartets are incorrect without knowing what

<table>
<thead>
<tr>
<th>Species</th>
<th>(i)</th>
<th>(ii)</th>
<th>(iii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>?$_C$</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>?$_D$</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>?$_E$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.1: Example of a compatibility table for the trees of Figure 6.1
6.2. Assessing the Amount of Conflicting Quartets

the correct phylogeny is. However, the minimal number of quartets that are necessary to explain all conflicts can be estimated. This is achieved by finding a set of quartets such that no conflict is observed if these quartets are removed from the analysis. Finding such a set of minimal size appears to be a computationally hard problem, since most likely any subset of the quartets can be the optimal solution. However, it is more likely that a quartet that is involved in many conflicts has to be removed than a quartet that is only part of one conflict.

This leads to a randomized greedy heuristic, where the quartets are sorted by the number of conflicts in which they are involved, highest number of conflicts first. The algorithm has two phases: in the first phase, one quartet after the other, along the previously determined order, is removed from the analysis if it further decreases the number of conflicts. After this phase, there can no longer exist any conflicts. In the second phase, every quartet, again following the order, is re-included into the analysis if it doesn’t cause a new conflict. Since there are 9,326 quartets that are involved in between 1 and 21 conflicts, the above determined order is highly ambiguous. Therefore, blocks of 50 quartets were randomly shuffled to produce a slightly disturbed order. This procedure was repeated 500 times and the solution with the lowest number of removed quartets was recorded.

By applying this algorithm, 1582 quartets were determined to form the minimal set that had to be removed from the analysis. This means that about 12.5% of all quartets are incorrect. One can assume that the number of quartets in this set is very close to the optimal solution, but the exact set is more difficult to determine, as it is often possible to remove one quartet and add another one without changing the number of conflicts. Nevertheless, quantitative statements can be made about the solution set, under the assumption that they will be similar for different optimal sets.

Interestingly, 1153 of those probably incorrect quartets had 100% bootstrap support, clearly showing that highly supported quartets can be incorrect. The number of groups for incorrect quartets ranges from 732 to 10,774 (average 3,292), so even large amounts of data do not guarantee correct reconstruction. The species that were involved the most in this set of quartets were platypus (in 586 quartets) and opossum (in 539). This is more than 50% more than the next frequent species in incorrect quartets, the mouse (in 338). Platypus and opossum are outgroups to the placental mammals and thus quartets involving any of those two will contain the longest possible distances among mammals. The longer distances are, the more sensitive they become to problems associated with model violations, as will be shown in the next section.
6.3 The Effect of Model Violations

Since most of the quartets concluded to be reconstructed incorrectly are based on large amounts of data and since the results are very significant, random error can be ruled out as the cause for the incorrect and conflicting results. Therefore it is very likely, that the estimated distance suffer from a systematic bias which is strong enough to mislead the reconstruction.

It is known that if the evolutionary model used for the distance estimation fails to accurately model the complexity of the true evolution, phylogenetic reconstruction methods can be inconsistent (converge towards an incorrect solution as more data are added). This effect is particularly strong for parsimony, but can also be shown for distance and likelihood methods (Farris, 1999; Swofford et al., 2001) when the evolutionary model assumptions are not adequate enough for the complexity of the true evolutionary process.

A prominent and simple example of model violations is rate heterogeneity across sites. For example in DNA sequences, bases at the third codon position (at which changes are often synonymous) undergo substitutions more often than bases at the second position. In proteins, residues near active sites evolve more slowly than those at less functional positions. The occurrence of rate heterogeneity in natural sequences is well documented (Fitch and Margoliash, 1967). Also its effects on phylogenetic inference have been investigated and counter-measures have been proposed (Yang, 1996).

In this section, simulations as well as a theoretical analysis based on a simple evolutionary model are used to demonstrate that if the evolution doesn’t follow the standard model assumptions, the distances will be underestimated and quartet reconstruction can fail. Since the empirical assessment of the amount of incorrect quartets was performed on weighted least-squares trees, the same method is also used in this study.

In the following sections, it will first be shown that in the presence of rate variation among sites, the estimated distances do not increase linearly with the simulated (true) distances. Therefore, the estimated distances are no longer additive. Since additivity is the basic assumption for distance tree reconstruction, it will be shown that this can lead to inconsistent quartet reconstructions. Finally, it will be demonstrated, that in the presence of unequal rates among lineages, conflicting quartets, as shown in the previous section, can occur.

Rate variation was simulated in two different ways: 1) keeping a fixed proportion of the sites invariant and 2) letting the rates follow a gamma distribution. In the case of invariant sites, a defined proportion \( f \) of the sequence was kept constant while the rest
of the sequence underwent simulated evolution with a rate of \(1/(1-f)\), resulting in an average rate of 1. In the case of gamma distributed rates, the sequence was divided into 10 equally long subsequences. Each subsequence was assigned a rate, such that the mean rate is 1 and the rate distribution is a discrete approximation of the gamma distribution (Yang, 1994). The subsequences were then simulated along the same tree, but with the distances multiplied by the category specific rate. The gamma distribution used in this context is typically defined by a single parameter \(\alpha\), with a low \(\alpha\) implying strong rate heterogeneity and a large \(\alpha\) indicating weak rate heterogeneity.

6.3.1 Theoretical Analysis on Binary Characters

When evolutionary rates among sites are assumed to be constant, but are not, the number of substitutions at the fast-evolving sites tends to be underestimated, leading to underestimations of the evolutionary distances (Golding, 1983). Gillespie (1986) proved that using the Jukes-Cantor formula for distance estimation when the rates are gamma-distributed, causes the calculated distances to not to increase linearly, but logarithmically with the true distances. A similar proof is shown here, but on a simpler model of binary characters and for a generic distribution of rate categories.

The analysis is performed using a two-state Markov model with identical transition probabilities, the so-called Cavender-Farris-Neyman model (Cavender, 1978) on infinitely long sequences of binary characters. After the sequence has evolved at a given rate \(\alpha\) over a time \(t\), each character is either in the original state or in a different one with the proportion \(p\) of differences being:

\[
p(\alpha, t) = \frac{1}{2}(1 - e^{-2\alpha t})
\]  
(6.1)

The variables \(\alpha\) and \(t\) occur only as a product, since it is not possible to distinguish, for example, between a sequence evolving with a given rate over a certain time and one evolving with twice that rate over half the time. Therefore, it is convenient to work with the distance \(d = \alpha t\) instead.

For a given proportion of differences \(p\), an estimate \(\hat{d}\) for the distance can be derived:

\[
\hat{d}(p) = -\frac{1}{2} \log(1 - 2p)
\]  
(6.2)

What happens if the rates vary among sites? The generic case of \(N > 1\) different rate categories with rates \(r_i \geq 0\) and probabilities \(0 < p_i < 1\), \(\sum_i p_i = 1\) and \(\sum_i p_i r_i = 1\) is
treated here. For a given distance $d$, the observable proportion of differences $p_R$ is:

$$p_R(d) = \sum_{i=1}^{N} \frac{p_i}{2} (1 - e^{-2r_i \cdot d})$$  \hspace{1cm} (6.3)

Applying Equation 6.1 to $p_R$ to estimate the distance $\hat{d}$ results in:

$$\hat{d}(p_R(d)) = -\frac{1}{2} \log(1 - \sum_{i=1}^{N} p_i (1 - e^{-2r_i \cdot d}))$$

$$= -\frac{1}{2} \log(\sum_{i=1}^{N} p_i e^{-2r_i \cdot d})$$  \hspace{1cm} (6.4)

Whereas in the case of only one rate, this simplifies to $\hat{d} = d$, for $N > 1$ it is a logarithm of a sum of exponentials and thus cannot be simplified to a linear relationship between the true distance $d$ and the estimated distance $\hat{d}$.

### 6.3.2 Simulations: Non-linearity of Distance Estimation

The theoretical analysis, that showed that in the presence of among-site rate variation, the estimated distance is not a linear function of the true distance, is only feasible on simple models, such as the binary character model presented above. For complex models, particularly models based on an empirical substitution matrix, analytical solutions cannot be derived, but simulations can be performed.

The simulations were performed by generating a random protein of 20,000 amino acids, drawn from the residue frequencies from empirical amino acid substitution matrices (Gonnet et al., 1992). These matrices were also used to randomly evolve the protein over distances ranging from 1 to 100 PAM. Various degrees of the two types of model violations introduced above were used: 20%, 50% or 80% invariant sites and gamma-distributed site rates for $\alpha$ values of 1, 0.5 and 0.2.

The simulation results are shown in Figure 6.2. The plots demonstrate that the estimated distance does not increase linearly with the simulated distance and that for strong model violations, the non-linearity is considerable.

If the estimated distances are not a linear function of the true distance, then the estimated distance are not additive and thus an important prerequisite of distance tree reconstruction is violated. In the following sections it will be shown that this can lead to incorrectly reconstructed least-squares quartets.
6.3.3 Simulations: Incorrect Quartet Reconstruction

Reconstruction of quartets is particularly error-prone in cases of unequal rates among lineages, i.e. when the branches leading to the leaves have different lengths. One such scenario is shown in the left panel of Figure 6.3: the two branches leading to B and D are twice as long as the branches leading to A and C. If, as shown in the previous section, the longer distances are underestimated more than the shorter ones, then B and D appear to be closer to each other than they actually are, which can under some circumstances lead to an incorrect reconstruction of the quartet.

Therefore, the quartet from Figure 6.3 was used to study the effect of model violations on quartet reconstruction. Simulations were performed by creating a random protein of 10,000 amino acid which was randomly mutated along the tree. The distances between the four sequences at the leaves were then estimated and the resulting distances and their variances were used to reconstruct the quartet using the weighted-least squares method of Darwin. Again, two types of rate heterogeneity were used: invariant sites and gamma-distributed site rates.

The space of all possible quartets is very large, therefore the simulations were performed on a subset by reducing the quartet to the two parameters shown in Figure 6.3: $p$, the length of the branches leading to A and C and half of the length of the two other branches, and $r$, the length of the middle branch. The longest distance, $4p + r$ is between B and D, while the smallest distance, $2p + r$, is between A and C (for $r < p$). In

![Figure 6.2: Simulations of distance estimation under model violations for different amounts of invariant sites (left) and different shape-parameters $\alpha$ for gamma-distributed site rates (right).](image)

Figure 6.2: Simulations of distance estimation under model violations for different amounts of invariant sites (left) and different shape-parameters $\alpha$ for gamma-distributed site rates (right).
order to find the conditions under which the reconstruction is problematic, $r_{50}$, the value of $r$ for which the reconstruction fails in 50% of the cases, was determined as a function of $p$. For each $p$, first the approximate range $r_{50}$ was determined by an exponential search and then for each of 50 different $r$’s in this range 20 runs of the simulations were executed to estimate the success probability for the respective $r$. From these 50 sample points, a linear regression was performed to estimate the value $r_{50}$.

The result is displayed in Figure 6.4. The plots show, that as the distances ($p$) become larger and the degree of model violations increases, the middle branch ($r$) needs to be larger as well in order to allow for successful reconstruction. Particularly for large proportions of invariant sites, the reconstruction can fail even for large middle branches. For example, for $p = 25$ and 80% invariant sites, an $r$ of 7 PAM is needed to achieve a success probability of more than 50%. 80% invariant sites might be unrealistic for most real sequences, but also for moderate degrees of model violations, middle branches of 2 or 3 PAM are a strong requirement for reconstruction success.

### 6.3.4 Simulations: Conflicting Quartets

After the simulations in the previous section showed to which extent model violations can lead to incorrect quartet reconstructions, it remains to be demonstrated, how this can also lead to conflicting quartets, since this requires that for a given quintet, some subquartets are correct while others or incorrect. The situation shown in the right panel of Figure 6.3 is used to analyze this situation: a fifth leaf $E$ is placed together with $D$, but with a shorter branch. In this scenario, it is expected that under some conditions, the reconstruction can fail even when the middle branch is large enough. The figure shows a scenario where the reconstruction fails because the distance between the fifth leaf $E$ and the quartet $A-C$ is small, leading to an incorrect quartet reconstruction.

![Figure 6.3](image-url)
6.3. The Effect of Model Violations

the long branch leading to D is incorrectly reconstructed to be together with B while the shorter branch leading to E is still correctly placed with C.

Sequences of 10,000 amino acids were simulated as described in the previous section, leading to five sequences A, B, C, D and E from which WLS quartets were reconstructed. For each quintet, three different quartets were reconstructed: (A, B, C, D), (A, B, C, E) and (A, B, D, E) and conflicts were determined as described in Section 6.2.2. For the same types and degrees of model violations as used in the previous section, combinations of the parameters $p$ and $r$ were explored to find the parameter space causing conflicts among the quartets. For each each value of $p$ from 1 to 30 and for each value of $r$ between 0.1 and 5 in steps of 0.1, 20 simulations were performed and the fraction of times a conflict occurred was recorded.

The result is displayed in Figure 6.5. The plots show that conflicts occur often in this scenario and that for increasing values of $p$ and increasing degrees of model violations, the critical range for $r$ also increases. An interesting observation can be made in the plot of 80% invariant sites (bottom right plot): It can be seen that for large values of $p$ and small middle branches $r$, the chance of a conflict becomes smaller. This is because for that parameter range the first two quartets, (A, B, C, D) and (A, B, C, E), are both reconstructed incorrectly, which can also lead to a conflict-free (but still incorrect) situation.

Figure 6.4: Results of simulations of quartet reconstruction under model violations for different amounts of invariant sites (left) and different shape-parameters $\alpha$ for gamma-distributed site rates (right). The y-axis is $r_{50}$, the length of the middle branch $r$ for which 50% of the reconstructions are correct.
6.4 Conclusions

Two main conclusions can be drawn from the analyses in this chapter: quartets reconstructed as PAM distance trees from orthologous groups can be incorrect due to systematic biases, and violations of the standard Markovian model can cause such systematic biases.

In the first part of the chapter, it was shown on a large-scale empirical study of 25 mammals, that at least 12% percent of the inferred quartets must be incorrect, as they cause conflicts with other quartets. The mammalian tree has many small internal branches, that are difficult to reconstruct, and thus the number of incorrect quartets is not necessarily representative for other data sets. However, most of the reconstructed quartets, also those determined as incorrect, are based on thousands of orthologs and strongly

Figure 6.5: Simulations of conflicting quartet reconstruction under model violations: different amounts of invariant sites (bottom row) and different shape-parameters $\alpha$ for gamma-distributed site rates (top row). The axis correspond to the quartet parameters $p$ and $r$, the color indicates the proportion of times a conflict was detected.
supported by high bootstrap values. This clearly implies that the reconstruction error must be due to a systematic bias, not due to random error. Furthermore, the evolutionary distances found between mammals are relatively small compared to most other data sets. Among metazoa, for example, the longest distances are four times longer than the longest distances among mammals.

In the second part of the chapter, the effect of model violations on the reconstruction of WLS quartets was analyzed. It is well known, that all phylogenetic methods can fail, when the model assumptions are violated. The purpose of the analysis was therefore to specifically investigate the influence of model violations on the model used for the mammalian quartets, using PAM distance trees on distances and violations comparable to those found among mammals. Rate variation among sites was chosen as an example of model violations, but in reality, many other violations can be present, such as non-stationarity of the substitution process or unequal composition biases in different species and genes, to name only a few possibilities. The following main conclusions can be reached from this analysis:

- In the presence of rate heterogeneity, the estimated distances do not increase linearly, but logarithmically with the true distance. This means, the estimated distance is generally underestimated, and that this bias becomes stronger with larger distances. Therefore, an important prerequisite for distance tree reconstruction, additivity of the estimated distances, is violated.

- Because of these properties, distance tree reconstruction can become inconsistent, particularly if also rate variation among lineages occurs. If longer distances are underestimated more than shorter distances, taxa that evolved at a faster rate are more likely to be reconstructed incorrectly, as their distances to the outgroups are underestimated.

The relevant question is of course if these findings apply only to simulations of pathological cases, or if they are also relevant for actual mammalian phylogenies. The mean $\alpha$ value found in the study of ML trees from the orthologs from human, dog, mouse and opossum was 0.563, and the longest branch found in the mammalian tree is about 15 PAM, corresponding to a $p$ of about 7.5 in the simulations of the previous sections. In this parameter range, a middle branch of about 0.4 PAM is not enough to guarantee correct reconstruction, even with very large amounts of data (see Figure 6.4). Of the 22 internal branches of the tree of 25 mammals, only 5 are larger than 0.4 PAM. If more violations than just rate heterogeneity occur, then all branches of the mammalian tree could be subject to reconstruction biases.
This demonstrates and emphasizes how important it is to use evolutionary models that incorporate as many features of sequence evolution as possible. The most commonly used model extension is gamma-distributed sites (Yang, 1994). There is no reason to assume that the rate distribution of sites in real sequences follow a gamma distribution, but it is convenient since it allows the modeling of differently shaped distributions using only one parameter, the shape parameter $\alpha$. Another approach, suggested by Lartillot and Philippe (2004), is to use so-called "mixture models" where the amino acid substitution process is allowed to be different among sites. This was also shown to improve the phylogenetic accuracy (Lartillot et al., 2007).

Since additional parameters (besides the distance) are difficult to estimate from pairwise alignments, methods that work on MSAs and trees are preferable, such as ML and Bayesian estimations. Furthermore, methods to test for possible reconstruction biases are also needed. One such method will be presented in the following chapter.
Chapter 7

Phylogenetic Signals from Outgroup Analysis

It is known that the accuracy of phylogenetic reconstruction decreases when more distant outgroups are used. In this chapter, this phenomenon is quantified with a novel scoring method, the outgroup score \( pOG \). This score expresses if the support for a particular branch of a tree decreases with increasingly distant outgroups. Large-scale simulations confirmed that the outgroup support follows this expectation and that the \( pOG \) score captures this pattern. The score often identifies the correct topology even when the primary reconstruction methods fail, particularly in the presence of model violations. In simulations of problematic phylogenetic scenarios such as rate variation among lineages (which can lead to long-branch attraction artifacts) and quartet-based reconstruction, the \( pOG \) analysis outperformed the primary reconstruction methods. Since the \( pOG \) method does not make any assumption about the evolutionary model (besides the decreasing support from increasingly distant outgroups), it can detect cases of violations not treated by a specific model or too strong to be fully corrected.

7.1 Introduction

7.1.1 Mammalian Phylogeny

Mammalian classification has been a subject of interest for centuries. Traditionally, morphological characters were used to classify the mammals (Simpson, 1945; Novacek, 1992a,b; Shoshani and McKenna, 1998). With the start of the genomic era, molecular studies added to the discussion, leading to sometimes surprisingly new classification,
such as a clade of African mammals, the demise of Insectivora as a clade, or the para-
phyly of Archonta by "moving" Chiroptera (bats) further away from Primates.

Presently, most molecular studies group the mammalian orders into 4 superorders: Af-
rotheria, Xenarthra, Laurasiatheria and Euarchontoglires. Afrotheria is a clade of
animals suggested to be originating from Africa and includes Proboscidea (elephants),
Afrosoricida (e.g. tenrecs), Hyracoidea (e.g. hyraxes) and Sirenia (e.g. manatees). The
only extant Xenarthra are found in Central and South America, and consist of Cingu-
lata (armadillos) and Pilosa (sloths and anteaters). Laurasiatheria are believed to have
evolved on Laurasia after the break-up of Pangaea, and include the orders Carnivora
(e.g. dogs and cats), Artiodactyla (e.g. cattle), Cetacea (e.g. dolphins), Perissodactyla
(e.g. horses), Chiroptera (bats), Erinaceomorpha (e.g. hedgehogs), Soricomorpha
(e.g. shrews) and Pholidota (pangolins). Euarchontoglires (synonymous to Supra-
primates) probably split from Laurasiatheria about 95 to 85 million years ago during the
Cretaceous. It contains Primates, Rodentia, Lagomorpha (e.g. rabbits), Scandentia
(e.g. tree shrews) and Dermoptera (e.g. flying lemurs). Laurasiatheria and Euarchon-
toglires together form a clade called Boreoeutheria (synonymous to Boreotheria).

7.1.2 Recent Advances in Phylogenetic Methodology

Most phylogenetic studies use nuclear or mitochondrial sequence data, and nucleotides,
amino acid or mixed data sets to estimate the phylogeny. Since the beginning of the
genomic era, many trees and corresponding hypotheses on the mammalian evolution
have been proposed and several issues are still not resolved, despite a constantly in-
creasing amount of genomic data. These ambiguities in phylogenetic reconstruction do
not only stem from data selection, but can also result from the choice of evolutionary
model and reconstruction method. Inconsistency (convergence towards an incorrect
solution as more data are added) can occur for parsimony as well as for distance and
likelihood methods (Farris, 1999; Swoford et al., 2001) when the biological data vio-
lates the standard evolutionary model assumptions of homogeneity of rates across sites
(Chang, 1996; Kuhner and Felsenstein, 1994), homogeneity of rates across branches
(Gaut and Lewis, 1995), stationarity (same character frequencies for all branches) or
time reversibility (Squartini and Arndt, 2008). Assessing the harm caused by model as-
sumptions is not a simple task and for the case of mammalian evolution, the adequacy
of the model can make the difference between topologies with short internal branches.

This dependence of the reconstruction on the method and data set has prompted the
investigation of evolutionary events that would solve and settle the structure of the tree
of life. Rare events such as insertion/deletion events (Murphy et al., 2007) and insertions of transposons (Salem et al., 2003; Kriegs et al., 2006; Nishihara et al., 2006) are expected to produce fewer homoplasies and since they occur in low numbers, they can be curated manually, increasing their reliability. Nikolaev et al. (2007) used genomic data from the ENCODE project to identify "conserved noncoding sequences" which appear to evolve more slowly than most coding sequences, thus reducing the chance of homoplasy.

Presence/absence patterns of retroposons have been proposed as a promising phylogenetic tool, because the insertion of a retroposon was seen as an irreversible and homoplasy-free process (Kriegs et al., 2006). Despite their apparent advantages, these events are not yet well understood, and researchers have found evidence for insertion hotspots, which might cause homoplasies (Cantrell et al., 2001; Bashir et al., 2005). There are also indications for precise incision (van de Lagemaat et al., 2005) and lineage-specific degradation (Murphy et al., 2007) which both imply that such insertions can be reversible. Also, since retroposon insertions are rare events, they are mostly found on longer branches. This makes it difficult or even unlikely to find significant numbers of such events to solve difficult problems that involve very short branches, as for instance found between some Laurasiatherian orders (Nishihara et al., 2005).

In cases of retroposons supporting two conflicting topologies, this was attributed to incomplete lineage sorting (Salem et al., 2003; Nishihara et al., 2006). A similar conclusion was reached by Hallström and Janke (2008) when analyzing 2845 kbp of coding sequences. They observed divergences (particularly the root of placental mammals) that remained ambiguous despite using very large amounts of data, prompting them to suggest hybridization or incomplete lineage sorting as potential causes for this uncertainty. When speciations follow each other in very short intervals (estimated to be 2–4 million years in their case), hybridization might still be possible. This can obscure or even erase any information about the correct order of speciation events. Again, such scenarios would affect mostly the short branches.

Because of the complicated nature of species tree reconstruction, it is important to have methods to assess the quality of branches or trees. However, many such methods, like bootstrapping (Felsenstein, 1985), measure the reliability within the model and methodology used for the reconstruction, which can lead to strong statistical support for incorrect topologies (Bergsten, 2005). It has also been shown that depending on the choice of data, high Bayesian posterior probabilities for mutually exclusive topologies can be found (Misawa and Nei, 2003).

The trade-off between using many genes from a few taxa and including more taxa but
having fewer genes available is also relevant (Geuten et al., 2007). As the number of full mammalian genomes increases, one heads toward the ideal of many genes for many taxa, a combination that can help to mitigate the effects of model violations. However many so-called "fully sequenced" genomes are of low coverage and their phylogenetic distribution is highly biased (Milinkovitch and Tzika, 2007).

For phylogenetic reconstruction, the sampling of species is an important issue. It has been shown that the choice of in- and outgroups can have dramatic consequences for the resulting phylogenetic tree (Graybeal, 1998; Poe and Swofford, 1999; Rosenberg and Kumar, 2001). Outgroups are species that diverged from all in-group taxa before they diverged from each other and are the preferred way to determine the root of a phylogenetic tree. It has previously been suggested that from all possible outgroups of comparable sequencing quality, the closest one is the best choice to determine a rooted tree (Ritland and Clegg, 1991; Muse and Weir, 1992; Smith, 1994) because shorter distances suffer less from statistical error and also the expected number of homoplasies between any ingroup and the outgroup is minimized this way.

### 7.1.3 The $pOG$ Score

A direct implication of the closest outgroup being the optimal one is that the quality of phylogenetic reconstruction decreases, when the distance to the outgroup is increased. Using this observation, the $pOG$ score is introduced. The computation of $pOG$ for a particular branch is based on the analysis of the support (the percentage gene trees supporting this branch) as a function of using increasingly distant outgroups. If the support decreases, corresponding to the expected pattern, then a high $pOG$ score is returned. If, however, the support increases, then the resulting low $pOG$ score expresses violation of the expected pattern.

Methods for constructing phylogenetic trees include minimizing the squared errors of fitting estimated pairwise distances to a tree (least-squares trees), minimizing the number of evolutionary events on the tree (maximum parsimony) or maximizing the likelihood (ML methods) or the posterior probability of the tree (Bayesian methods). We call the result of any of these methods primary results, since they are obtained by directly applying the method to the data. The $pOG$ score, however, is a secondary result, since it uses only the primary results (the number of gene trees with different outgroups supporting a topology) to compute a derived score. The computation of $pOG$ is independent of the primary method used.

In this chapter, the computation of the $pOG$ score is presented, demonstrated on a real
example and verified using large-scale simulations of six different primary methods and two different kinds of violations to the evolutionary model. Then the method is applied to alternative hypotheses of several problems of current mammalian phylogeny. Finally, the product of the $pOG$ values on all branches of the tree is used as an optimization criterion to construct a tree of 23 Placentalia.

7.2 Methods

7.2.1 Gene Tree Support

In this study, the results of several primary methods of tree reconstruction are compared to results of the $pOG$ score. The study is conducted on large sets of data with many groups of orthologs. Gene trees are reconstructed for every group separately and the support of a particular branch is expressed as the percentage of gene trees containing that branch. As the focus of the study is on the change of support when using different outgroups, the specific situation of a quartet ($A$, $B$, $C$, $OG_i$) is considered, with $A$, $B$ and $C$ being three ingroup species of interest and $OG_i$ being one of the outgroups.

For the three ingroup species, there are three possible rooted topologies, denoted $a$, $b$ and $c$. At this point it is not important to which topology these labels refer, it will be made clear from the context. Assuming that there are $n_i$ orthologous groups containing $A$, $B$ and $C$ and outgroup $OG_i$, the numbers of gene trees supporting the three topologies are denoted $a_i$, $b_i$ and $c_i$ (with $a_i + b_i + c_i = n_i$).

Given $k$ different outgroups, the gene tree support for a particular topology $a$ is computed as:

$$\frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} n_i}$$

(7.1)

and likewise for topologies $b$ and $c$. Consequently, the sum of the support values for all three topologies add up to 1. If all gene trees agree on the same rooted triplet, the support for that topology would be 1. However, the inference of gene trees can be subject to stochastic errors, reconstruction artifacts, systematic biases or mistakes in the orthology assignment (Castresana, 2007). Furthermore, some so-called “anomalous gene trees” do not follow the species trees because of lineage sorting effects even if optimally reconstructed (Degnan and Rosenberg, 2006; Ewing et al., 2008). For these reasons, the gene tree support for a topology is expected to be less than one.
7.2.2 The $pOG$ Score

Given that the closest outgroup is the optimal choice, the support of a branch is expected to be the highest if the closest outgroup is chosen. More generally, the support is expected to decrease with increasing distance to the outgroup. Or formally, in the case of $a$ being the correct topology, the following inequalities should hold:

$$\frac{a_1}{n_1} > \frac{a_2}{n_2} > \ldots > \frac{a_k}{n_k}$$

(7.2)

for outgroups $1, 2, \ldots, k$ with increasing distances to the ingroups.

However, this ideal pattern is not always strictly observed and it is possible that one of the other two topologies also has decreasing support from more distant outgroups. Therefore it is necessary to establish if the $k$ fractions in Equation 7.2 are in the most decreasing order. Also the significance of this pattern should be quantified. The resulting probability of this event will be called the $pOG$ score.

If the $k$ fractions of Equation 7.2 are not in strictly decreasing order, the distance from strictly decreasing order must be established. This is achieved by determining the minimal number of necessary neighbor interchanges to make the numbers strictly decreasing. After computing this number for all three topologies, the one with the lowest number of necessary changes is the closest to the expected pattern. In the ideal case, one topology would need zero swaps and the other two would each need $k(k - 1)/2$ changes.

In practice, however, such a clear pattern is rare. Sometimes, two topologies have a more or less decreasing outgroup support pattern. Furthermore, the support fractions $a_i/n_i$ have a variance, and thus the number of necessary swaps are also subject to uncertainty. Therefore, simply using the number of necessary swaps is a too coarse measure to quantify the outgroup support pattern.

Instead, the $pOG$ score quantifies the probability that a given topology has the most decreasing pattern. This is achieved via Monte Carlo simulations by modeling the number of gene trees $a_i$, $b_i$, and $c_i$ as a multinomial distribution. The outcomes from the $k$ different outgroups are assumed to be independent, and thus one run consists of sampling from $k$ multinomial distributions and determining which of the three resulting sequences of support fractions $a_i/n_i$ is the most decreasing one (as described above). This procedure is repeated many times and each time the topology with the lowest number of necessary neighbor interchanges is recorded. If two or three topologies result in the same number of changes, then each is counted $1/2$ or $1/3$, respectively. The number of Monte Carlo runs should be as high as possible to minimize the variance of the es-
timate. However, since this is a time-intensive procedure, in practice the number is limited. For this study, if not stated otherwise, a value of 200 runs was found to be sufficient while allowing for reasonable computation times. Finally, the proportion of times each topology resulted in the most decreasing sequence of outgroup support fractions is determined. This proportion is called the \( p_{OG} \) score.

Since the \( p_{OG} \) score is the probability that a topology has the most decreasing support from increasingly distant outgroups, it ranges from 0 to 1 and the scores for the three alternative topologies always sum up to 1. The \( p_{OG} \) value for a particular topology is high (close to 1) if that topology has a significantly more decreasing pattern than either of the other two topologies.

### 7.2.3 Example of the \( p_{OG} \) Score

An example is given in Table 7.1 to illustrate the computation of the \( p_{OG} \) score. The top part of the table displays the numbers of least-squares trees supporting each of the three possible topologies for the triplet human, mouse, rabbit (rows) and every outgroup (columns). 6478 groups were found with opossum as outgroup, but only 1192 with the ciona outgroup. Since these numbers cannot be compared directly, the fractions of gene trees supporting each topology are used (shown in the lower part of the table).

The average support for a given topology is computed as in Equation 7.1 and is given in the last column of the lower half of the table. The highest support (0.559) from the

<table>
<thead>
<tr>
<th>Outgroup ( (OG_i) )</th>
<th>Opossum</th>
<th>Platypus</th>
<th>Chicken</th>
<th>Frog</th>
<th>Fish</th>
<th>Ciona</th>
<th>( \sum_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_i )</td>
<td>1570</td>
<td>908</td>
<td>1278</td>
<td>1031</td>
<td>967</td>
<td>357</td>
<td>6111</td>
</tr>
<tr>
<td>( b_i )</td>
<td>3444</td>
<td>1947</td>
<td>2916</td>
<td>2368</td>
<td>2149</td>
<td>633</td>
<td>13457</td>
</tr>
<tr>
<td>( c_i )</td>
<td>1464</td>
<td>741</td>
<td>884</td>
<td>670</td>
<td>532</td>
<td>202</td>
<td>4493</td>
</tr>
<tr>
<td>( n_i )</td>
<td>6478</td>
<td>3596</td>
<td>5078</td>
<td>4069</td>
<td>3648</td>
<td>1192</td>
<td>24061</td>
</tr>
</tbody>
</table>

| \( a_i/n_i \)       | 0.242    | 0.253    | 0.252   | 0.253| 0.265| 0.299 | 0.254   |
| \( b_i/n_i \)       | 0.532    | 0.541    | 0.574   | 0.582| 0.589| 0.531 | 0.559   |
| \( c_i/n_i \)       | 0.226    | 0.206    | 0.174   | 0.165| 0.146| 0.169 | 0.187   |

Table 7.1: An example using human, mouse and rabbit to illustrate the scores described in the text. Topology \( a \) is the human-mouse clade, \( b \) the human-rabbit clade and \( c \) the mouse-rabbit clade. Above the line are the number of gene trees supporting each topology (rows) using each outgroup (columns). Below the line these counts are expressed as frequencies.
gene trees is for \( b \), the rabbit forming a clade with human to the exclusion of the mouse. The computation of the \( pOG \) score has two components: computing the ranking distance to strictly decreasing sequences and assessing the significance using Monte Carlo simulations. The necessary neighbor interchanges to sort the fractions for topology \( c \) (bottom row of the table) is 2, since the 0.169 in the last column needs to be moved two positions to the left in order to achieve the sorting. Similarly, the ranking distance is 14 for \( a \) and 10 for \( b \).

From this, one can conclude that \( c \) is the one showing the most decreasing support pattern. However, the fractions shown in the table are close and the ranking obtained could be influenced by uncertainties that naturally exist when measuring distances and building trees. Therefore Monte Carlo is performed by assuming the outcome from each outgroup as multinomially distributed. For example for the opossum outgroup column, three numbers are drawn from a multinomial distribution for 6478 trials, with the three categories having probabilities of 0.242, 0.532 and 0.226. For each replicate of the simulations, six such triplets are generated (one for each outgroup), the three ranking distances are computed (one for each topology) and the most decreasing topology is determined. After 200 replicates, the \( pOG \) score is computed as the percentage of how often each topology had the most decreasing support. In the example here, this was always \( c \), leading to \( pOG \) scores of 0.00, 0.00 and 1.00, for hypotheses \( a \), \( b \) and \( c \), respectively. Thus, unlike the gene tree support (using the primary method), the \( pOG \) based analysis favors the rabbit forming a clade with the mouse.

### 7.2.4 Simulations

Large-scale simulations were performed to assess the behavior of the outgroup support patterns under different conditions using six different tree reconstruction methods applied to thousands of trees. Additionally, two different types of model violations were simulated to compare the \( pOG \) score to the primary methods for scenarios difficult for phylogenetic reconstruction.

Random proteins were simulated using the updated PAM matrices (Gonnet et al., 1992) along the tree shown in Figure 7.1. The triplet (A, B, C) of the ingroup species was constructed with the intention that it is prone to suffer from so-called "long-branch attraction" (LBA), one of the most common reasons for failure in reconstructing phylogenetic trees (Philippe and Laurent, 1998; Bergsten, 2005). Under certain conditions, the long branch leading to B tends to be reconstructed erroneously as an outgroup to A and C, since the number of substitutions between B and the outgroup are underestimated.
This phenomenon can be observed for all phylogenetic methods and models of evolution, particularly if the real or simulated evolution deviates too much from the model used for the reconstruction (Swofford et al., 1996). The problem is enhanced by using few taxa, as including more taxa often helps to mitigate the problems associated with LBA (Philippe and Douzery, 1994; Halanych, 1998).

Each run of the simulation consisted of: 1) choosing branch lengths randomly from Poisson distributions with mean values as labeled in the tree of Figure 7.1, except the small branch separating C from A and B which was drawn from $\text{Poisson}(50)/100$, resulting in an expected value of 0.5 as shown in the tree. The length of the initial sequence at the root was drawn from $\text{Poisson}(500)$. 2) simulating evolution along the tree for $N$ groups, with $N$ drawn from $\text{Poisson}(2000)$. For each of the $N$ groups of genes, an initial protein sequence at the root was generated randomly, based on empirical amino acid frequencies (Gonnet et al., 1992) and then randomly mutated along the tree. For most simulations, not all sites of the sequences were evolved at the same rate, as described in Section 7.2.6. 3) reconstructing quartets of the three ingroups and one of the outgroups using several primary phylogenetic methods (Section 7.2.5). The number of times each topology is supported using the different outgroups is recorded.

Figure 7.1: Tree used for simulations. A, B and C are the ingroup species and OG1 to OG6 are outgroups with increasing distances to the ingroups. The branch lengths are in PAM (accepted mutations per 100 residues) and are the average values used for the simulations.
leading to a table similar to Table 7.1. 4) establishing for each run of the simulations the highest supported topology and the one with the highest \( pOG \) value. Each data point in the plots of the simulation results in Section 7.3.1 corresponds to 100 runs of this procedure.

### 7.2.5 Quartet Reconstruction

The quartets from the simulated sequences were reconstructed using three different methods: maximum-likelihood, Bayesian estimation and least-squares distance trees. Because of the very large amount of groups and trees (25.8 million quartets), the analysis of the Metazoan tree was only performed on distance trees, as this analysis requires a fast method.

ML trees were created with PhyML version 2.4.4 (Guindon and Gascuel, 2003) using the JTT matrix to optimize the topology as well as the branch lengths. Four variants were used: equal rates for all sites (here called "JTT"), estimating a proportion of sites that are invariant ("JTTI"), estimating a discrete gamma distribution with four rate categories ("JTTG") and estimating both the discrete gamma distribution as well as the proportion of invariant sites ("JTTGI").

Bayesian reconstruction of phylogenetic trees were conducted with MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003) using the "mixed" amino acid model, which samples each of 10 models according to its probability. A MCMC analysis was run with 1000 cycles, sampled every 10, with a burn-in value of 25. These numbers are unusually small, but tests showed that for a quartet, 1000 cycles produce the same result as 10,000. As MrBayes is very time-consuming, it was not possible to do the many simulations with more cycles of the MCMC.

Distance trees were constructed from pairwise distance estimates between proteins using the pairwise sequence alignment method of the Darwin bioinformatics software (Gonnet et al., 2000). The PAM distances and their variances were estimated as described in Dessimoz et al. (2006). Based on the \( 4 \times 4 \) distance and variance matrices, the quartet minimizing the sum of the variance-weighted squared errors between the pairwise distances and the distances on the quartet is selected, a method known as weighted least-squares (WLS) (Cavalli-Sforza and Edwards, 1967). In the case of two topologies having the same lowest least-squares fit, the quartet was excluded from the analysis.
7.2. Methods

7.2.6 Simulated Model Violations

When the simulations are conducted according to a simple evolutionary model, most methods (except in some cases parsimony) should be able to reconstruct the correct topology. However, if the model is violated, meaning that the evolution of the sequences differs from the model used to reconstruct the phylogeny, then every method can fail. For example, the existence of rate heterogeneity across sites in natural sequences is well documented (Fitch and Margoliash, 1967) as are its effects on phylogenetic inference (Yang, 1996). Therefore rate variation among sites is a suitable model violation to study the effect on phylogenetic reconstruction methods.

Rate variation was simulated in two different ways: 1) keeping a fixed proportion of the sites invariant and 2) letting the rates follow a gamma distribution. In the case of invariant sites, a defined proportion of the sequence was kept constant while the rest of the sequence underwent simulated evolution along the tree as described above. The simulations were performed for percentages of invariant sites ranging from 0 percent (no violation) to 90 percent (very strong violation). In the case of gamma distributed rates, the sequence was divided into 10 equally long subsequences. Each subsequence was assigned a rate, such that the mean rate is 1 and the rate distribution is a discrete approximation of the gamma distribution (Yang, 1994). The subsequences were then simulated along the tree as described above, but with the distances multiplied by the category specific rate. The gamma distribution used in this context is typically defined by a single parameter $\alpha$. For the simulations, $\alpha$ values ranging from 0.1 (strong rate heterogeneity) to 2 (weak rate heterogeneity) were used.

7.2.7 Genomic Sequences

The analysis of the Metazoan tree is based on orthologous protein sequences of 29 metazoan species. Genomic sequences were downloaded from Ensembl (Hubbard et al., 2007) and integrated into the OMA orthologs project (Dessimoz et al., 2005; Schneider et al., 2007a; Roth et al., 2008). The orthology assignments of the OMA groups were shown to be the most reliable for phylogenetic analyses among current orthology prediction algorithms (Altenhoff and Dessimoz, 2009). All 23 currently available placental mammals are used, as well as six outgroups with increasing distance to Placentalia: opossum, platypus, chicken, frog, zebrafish and ciona. All common names as well as scientific names and assembly versions of the genomes used are listed in Table 7.2.
The ortholog predictions come from the Metazoa set of OMA from April 29, 2008, containing 108,339 orthologous groups. Particularly because the low-coverage genomes have many genes missing, only 3 groups contained all 29 species used. Therefore, the analysis was done on quartets of species, each of them containing three Placentalia and one of the six outgroups. For these quartets, there were between 11,409 (hu-

<table>
<thead>
<tr>
<th>Common name</th>
<th>Scientific name</th>
<th>Version</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td><em>Homo sapiens</em></td>
<td>36</td>
<td>15,584</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td><em>Pan troglodytes</em></td>
<td>2.1</td>
<td>14,585</td>
</tr>
<tr>
<td>Orangutan</td>
<td><em>Pongo pygmaeus abelii</em></td>
<td>2</td>
<td>13,478</td>
</tr>
<tr>
<td>Rhesus macaque</td>
<td><em>Macaca mulatta</em></td>
<td>1.0</td>
<td>13,980</td>
</tr>
<tr>
<td>Bushbaby</td>
<td><em>Otolemur garnettii</em></td>
<td>1</td>
<td>7,791</td>
</tr>
<tr>
<td>Mouse lemur</td>
<td><em>Microcebus murinus</em></td>
<td>1</td>
<td>9,311</td>
</tr>
<tr>
<td>Tree shrew</td>
<td><em>Tupaia belangeri</em></td>
<td>1</td>
<td>7,454</td>
</tr>
<tr>
<td>Mouse</td>
<td><em>Mus musculus</em></td>
<td>37</td>
<td>14,933</td>
</tr>
<tr>
<td>Rat</td>
<td><em>Rattus norvegicus</em></td>
<td>3.4</td>
<td>13,904</td>
</tr>
<tr>
<td>Squirrel</td>
<td><em>Spermophilus tridecemlineatus</em></td>
<td>1</td>
<td>6,645</td>
</tr>
<tr>
<td>Guinea Pig</td>
<td><em>Cavia porcellus</em></td>
<td>2</td>
<td>5,776</td>
</tr>
<tr>
<td>Rabbit</td>
<td><em>Oryctolagus cuniculus</em></td>
<td>1</td>
<td>8,257</td>
</tr>
<tr>
<td>Pika</td>
<td><em>Ochotona princeps</em></td>
<td>2.0</td>
<td>9,488</td>
</tr>
<tr>
<td>Dog</td>
<td><em>Canis familiaris</em></td>
<td>2</td>
<td>13,947</td>
</tr>
<tr>
<td>Cow</td>
<td><em>Bos taurus</em></td>
<td>3.0</td>
<td>13,712</td>
</tr>
<tr>
<td>Cat</td>
<td><em>Felis catus</em></td>
<td>1</td>
<td>7,559</td>
</tr>
<tr>
<td>Horse</td>
<td><em>Equus caballus</em></td>
<td>2</td>
<td>13,527</td>
</tr>
<tr>
<td>Microbat</td>
<td><em>Myotis lucifugus</em></td>
<td>1</td>
<td>8,764</td>
</tr>
<tr>
<td>Hedgehog</td>
<td><em>Erinaceus europaeus</em></td>
<td>1</td>
<td>7,489</td>
</tr>
<tr>
<td>Common shrew</td>
<td><em>Sorex araneus</em></td>
<td>1</td>
<td>6,359</td>
</tr>
<tr>
<td>African elephant</td>
<td><em>Loxodonta africana</em></td>
<td>1</td>
<td>7,893</td>
</tr>
<tr>
<td>Tenrec</td>
<td><em>Echinops telfairi</em></td>
<td>1</td>
<td>8,152</td>
</tr>
<tr>
<td>Armadillo</td>
<td><em>Dasypus novemcinctus</em></td>
<td>1</td>
<td>5,408</td>
</tr>
<tr>
<td>Opossum</td>
<td><em>Monodelphis domestica</em></td>
<td>5.0</td>
<td>14,326</td>
</tr>
<tr>
<td>Platypus</td>
<td><em>Ornithorhynchus anatinus</em></td>
<td>5</td>
<td>8,079</td>
</tr>
<tr>
<td>Chicken</td>
<td><em>Gallus gallus</em></td>
<td>2.1</td>
<td>10,145</td>
</tr>
<tr>
<td>Frog</td>
<td><em>Xenopus tropicalis</em></td>
<td>4.1</td>
<td>8,901</td>
</tr>
<tr>
<td>Zebrafish</td>
<td><em>Danio rerio</em></td>
<td>6</td>
<td>8,082</td>
</tr>
<tr>
<td>Ciona</td>
<td><em>Ciona intestinalis</em></td>
<td>2</td>
<td>2,598</td>
</tr>
</tbody>
</table>

Table 7.2: List of species, assembly version and number of sequences used in the analysis of the Metazoan tree. The six species below the line were used as outgroups.
man, chimp, mouse and opossum) and 150 (squirrel, armadillo, guinea pig and ciona) groups available. Overall, 18,303 different groups of orthologs were used, with a total of 233,996 mammalian sequences and 52,131 outgroup sequences. The number of sequences used from each species are listed in Table 7.2.

The analysis was performed on the protein sequences. All quartets with all pairwise distances being 0 were excluded from the analysis, since they contain no phylogenetic information. The OMA algorithm ensures that two proteins can only be in the same group, if at least 60% of the longer protein can be aligned to the other sequence. Thus, all alignments are based on a large part of both sequences, preventing aligning fragments or single domains to full sequences.

The highest support for a rooted triplet found in this dataset is 0.911 for mouse and rat forming a clade to the exclusion of cow and the outgroup support pattern was often stable, with 793 of 1771 possible triplets having \( pOG \) values of 1.000.

### 7.2.8 Tree Optimization

The outgroup analysis assigns a \( pOG \) score to a branch of a tree. When applied to all branches of a tree, it can also be used to evaluate a tree topology. Since the branch scores are probabilities, the score of the tree is the product over all branch scores.

In order to find the tree with the optimal \( pOG \) scores on all branches, in principal all possible topologies have to be evaluated. However, for the 23 Placentalia with complete genomes that were used in this study, more than \( 10^{23} \) different rooted trees are possible, making a complete search infeasible. Therefore a heuristic search was performed.

Trees were constructed in two phases: stepwise addition followed by nearest neighbor interchanges (NNI) (described e.g. by Felsenstein (2004)). The stepwise addition consists of starting with a tree of two random species to which one species after the other is added at the best branch (such that the global \( pOG \) score is maximized). This is a greedy algorithm and often leads to a non-optimal topology. By performing NNI (local rearrangements of the tree), the overall score can be improved, but finding the globally optimal tree is still not guaranteed. Therefore, this procedure was repeated 20 times with species being added in different orders and of those 20 results, the highest scoring tree was selected.
7.3 Results and Discussion

The \( pOG \) score introduced in the Method section provides a way to quantify the pattern of decreasing branch support from increasingly distant outgroups. This pattern is expected to hold for all correct branches of a phylogenetic tree. But can it also be observed when the reconstructed branch is not correct? Simulations in the following section show that for incorrect branches, the decreasing support from the outgroups can normally not be observed. This is expressed in a low \( pOG \) score. Thus, the analysis of the \( pOG \) values is a useful method to detect the incorrectly reconstructed branches in a tree.

Following these results, the outgroup analysis was applied to branches of the mammalian tree in order to analyze the outgroup support pattern of several disputed phylogenetic problems and used as an optimization criterion to construct a tree of 23 Placentalia.

7.3.1 Simulations

Large scale simulations were performed to assess how often the outgroup support pattern captured by the \( pOG \) score is able to identify the correct topology, even when the primary reconstruction methods failed. The results of these simulations are shown in Figures 7.2 and 7.3.

The green lines indicate the percentages of simulation runs in which the primary method successfully reconstructed the correct topology for the majority of quartets. The correct topology is the long branch of B together with A to the exclusion of C, as shown in Figure 7.1. These plots clearly confirm the well known fact that all methods can fail if the evolutionary model is violated (see Introduction). On the left-hand side of the figures are the methods that do not assume any rate variation (JTT, WLS and Mr Bayes). These models perform very poorly, particularly in the case of the simulated gamma distributed rates where for an \( \alpha \) of 0.1 all methods fail in 80% of the cases or more. Clearly the poorest performing method is WLS for which already quite weak violations lead to a high failure rate. The methods on the right (JTTI, JTTG and JTTGI) that assume some form of rate variation among sites clearly perform better and only start to fail for strong model violations.

The red lines in the plots indicate in how many times of the 100 runs of the simulations the \( pOG \) score was the highest for the correct topology. Since the \( pOG \) score quantifies the change in support for increasingly distant outgroups, it can also be high for
a topology that is not the most supported by the primary method. This was observed frequently in the simulations, particularly when the support from the gene quartets favored a wrong topology. In cases where there is very little or no model violation, the primary methods always find the correct topology. However, when model violations are

Figure 7.2: Results from simulations with among-site rate variation modeled as a discrete gamma distribution (\(\alpha\) values between 0.1 and 2). Each plot shows the percentage of correct reconstructions for the primary method (green) and the pOG score (red). The primary methods on the left (JTT, WLS, Mr Bayes) do not model any violations, whereas those on the right assume invariant sites (JTTI), gamma distributed sites (JTTG) or both (JTTGI).
so strong that the primary methods fail, the outgroup analysis often still finds the correct tree. Thus, the $pOG$ score is able to capture the correct phylogenetic signal even when the primary methods fail. Although this effect is the strongest for methods that assume constant rates among sites, it can also be observed for the methods that incorporate rate variation.

Figure 7.3: Results from simulations with invariant sites (0 to 90 percent sites kept invariant). Each plot shows the percentage of correct reconstructions for the primary method (green) and the $pOG$ score (red). The primary methods on the left (JTT, WLS, Mr Bayes) do not model any violations, whereas those on the right assume invariant sites (JTTI), gamma distributed sites (JTTG) or both (JTTGI).
The simulations were performed on two kinds of rate variation (invariant and gamma-distributed sites) that are often considered by standard evolutionary models. The pOG based analysis performed well in these cases without explicitly modeling these violations. Since in reality biological sequences can have many evolutionary mechanisms that are not treated by standard models, the outgroup support behavior could be a useful indicator of such unexpected model failures. Also, the pOG based method performed well in simulations on quartets which are often considered to be problematic in phylogenetic reconstruction.

7.3.2 Testing Phylogenies Using the pOG Score

Following the results of the simulations which indicated that the pOG score is able to identify the correct branching pattern of a quartet even under strong violations of the evolutionary model, the method was applied to several current problems in mammalian phylogeny. These trees are shown in Figures 7.4 and 7.5. Each row corresponds to a particular branch of the mammalian tree, for which all three possible topologies are shown together with the respective pOG values.

Orthologous proteins from the OMA project and six metazoan outgroups were used as described in Methods. A convenient property of the pOG score for this application is that the pOG value of a particular branch of the tree depends only on the species in the three subtrees defining the branch, but not on the branching patterns within the subtrees. This means for instance that the pOG score of the branch defining Primates does not depend on the branching order within Primates, only on the selection of Primate species used for the analysis. Therefore, the subtrees in Figure 7.4 are shown only as triangles with a number indicating the number of species used within that subtree.

a) Rooting Placental mammals   The very early divergences of the eutherian mammals into Afrotheria, Xenartha and Boreoeutheria (consisting of Euarchontoglires and Laurasiatheria) are still debated. Three possible hypotheses for the placement of the root have been proposed and are shown in Figure 7.4a. The first scenario (left) roots Placentalia between Boreoeutheria and Atlantogenata (Afrotheria and Xenartha). This corresponds to the hypothesis that the first divergence occurred when the land masses of Gondwana and Laurasia separated, dividing Atlantogenata on Gondwana from Boreoeutheria on Laurasia. Subsequently, the separation of South America from Africa caused a further split between Xenartha and Afrotheria while in the north the divergence of Boreoeutheria is less clear (Wildman et al., 2007). Recent molecular studies
Chapter 7. Phylogenetic Signals from Outgroup Analysis

(Hallström et al., 2007; Wildman et al., 2007) as well as an analysis of coding indels (Murphy et al., 2007) support this scenario. The fossil record in Africa and South America neither supports nor contradicts the correlation between this phylogeny and the biogeographic theory outlined above (Asher et al., 2003). The second hypothesis (middle) roots Placentalia between Xenarthra and Eutheria (Boreoeutheria and Afrotheria)

Figure 7.4: Analysis of five problems in mammalian phylogeny using the $pOG$ score. Each row displays the three possible branching patterns for a particular branch of the mammalian tree and the corresponding $pOG$ scores: a) relative positions of Afrotheria, Xenarthra and Boreoeutheria, b) branching order of Primates, Rodentia and Laurasiatheria, c) position of Lagomorpha, d) relative branchings of guinea pig, squirrel and mouse/rat, e) position of the tree shrew
Results and Discussion

and is supported by studies of morphology (Novacek, 1992b; Shoshani and McKenna, 1998). This topology was also implied by the finding of two L1MB5 retroelements present in Epitheria but not in Xenarthra (Kriegs et al., 2006). However, this finding was not statistically significant and was later suggested to be an artifact of lineage specific degradation (Murphy et al., 2007). The third tree roots Placentalia between Afrotheria and Exafroplacentalia (Boreoeutheria and Xenarthra) and was supported by large nuclear data sets (Nishihara et al., 2007).

The $p_{OG}$ based analysis supports the first hypothesis (Atlantogenata) with a score of 0.915. It should be noted that Xenarthra is only represented by one low-coverage genome, the armadillo, resulting in a low number of orthologous groups being available.

b) The positions of the Superorders Primates, Rodentia and Laurasiatheria  

The branching order of Primates, Rodentia and Laurasiatheria (e.g. Carnivora and Artiodactyla) is also subject to debate. Historically, based on morphology, Primates and Rodentia were united in a group called Unguiculata (Linnaeus, 1758; Simpson, 1945) to the exclusion of Ferungulata (containing, among others, Carnivora and Artiodactyla). Shoshani and McKenna (1998) analyzed 260 morphological characters and found conflicting evidence, but slightly stronger support for this traditional clade than for the alternative, a Primate-Carnivora grouping. Molecular studies sometimes supported a Primate-Rodentia clade (called Euarchontoglires) (Murphy et al., 2001a; Kriegs et al., 2006), or a Primate-Carnivora group (Cannarozzi et al., 2007; Huttley et al., 2007). Since the rate of evolution in the rodent lineage is known to be faster (Lin et al., 2002), these results have been attributed to long-branch attraction. Also, the latter topology was often observed in studies with few taxa and it has been proposed that LBA is enhanced by insufficient taxon sampling (Lunter, 2007). This view is supported by the outgroup analysis with strong ($p_{OG}$ of 1.000) support of Euarchontoglires, but 51% of the gene trees supporting a Primate-Laurasiatheria clade (whereas 31% support Primate-Rodentia and 18% a Primate outgroup).

c) Position of Lagomorpha  

The position of Lagomorpha (here represented by rabbit and pika) has been long debated with the main hypotheses focusing on the relative positions of Lagomorpha, Rodentia and Primates. The original grouping by Linnaeus of Lagomorpha and Rodentia forming a superorder called Glires has been subject to a long debate among morphologists. Shoshani and McKenna (1998) wrote "like a pendulum, acceptance of Glires ... shifts in either direction, depending on available evidence". A similar observation can be made about early molecular studies, sometimes support-
ing Glires (Murphy et al., 2001a,b), sometimes supporting alternative groupings (Graur et al., 1996; Arnason et al., 2002b; Misawa and Janke, 2003). Since an extensive study by Douzery and Huchon (2004) and support from retroposon insertions (Nishihara et al., 2006) the concept of Glires seems to be generally accepted. Springer and Murphy (2007) attribute the main alternative topology (Rodentia basal) as long-branch attraction due to the faster rate of evolution in the mouse and rat lineage. The outgroup analysis ($pOG$ of 1.000) strongly supports Glires, despite a majority of 43% of the gene trees supporting a Rodentia outgroup and only 23% supporting Glires. This is another example where the primary method (WLS trees) is probably affected by LBA whereas the $pOG$ score identifies the correct phylogenetic signal.

d) Relationships among Rodentia  Many relationships among Rodentia are still unclear. The most comprehensive study of the intra-ordinal relationships in mammals has been performed by Huchon et al. (2002) using sequences from 22 rodent taxa and several outgroups. Because fully sequenced genomes are used in our study, only four rodents are available. They cover all three infraordinal clades as defined by Huchon et al. (2002): mouse and rat represent Myodonta, the squirrel represents Sciuroidea and the guinea pig is the only Ctenohystrica. The branching order of these three clades is not well defined, Huchon et al. (2002) obtain results that depended on the method used with none of the possible bifurcating trees having a significantly higher likelihood than another. A $pOG$ score of 0.950 speaks in favor of the guinea pig being more closely related to Myodonta than the squirrel. This is in agreement with the tree by Murphy et al. (2001b) and the consensus tree presented by Springer and Murphy (2007). Again, this result is obtained despite a clear majority (43%) of the gene trees supporting a guinea pig-squirrel clade, and only 23% support for the basal position of the squirrel.

e) Position of the tree shrew  The position of the tree shrew (order Scandentia) is not fully resolved. Current discussion focuses on the branching orders of Primates, Scandentia and Dermoptera, jointly forming a clade called Euarchonta, which in turn is seen as a sister group to Glires (Janečka et al., 2007). Nevertheless, there are also recent analyses, particularly on mitochondrial genomes, that place Scandentia basal in Euarchontoglires (Horner et al., 2007), or together with Lagomorpha (Lin et al., 2002). Studies of retroposon insertions could not resolve the position of Scandentia within Euarchontoglires (Nishihara et al., 2006). The $pOG$ score (0.890) favors a grouping of the tree shrew with Rodentia, rather than with Primates.
Intra-Laurasiatherian relationships  Within Laurasiatheria, the relative positions of Artiodactyla (here represented by the cow), Perissodactyla (horse), Chiroptera (bat) and Carnivora (dog and cat) remain unresolved. Particularly the positions of Perissodactyla and Chiroptera are ambiguous.

Morphological data places Perissodactyla together with Artiodactyla in Ungulata, a clade of hoofed and related animals (hyraxes, whales, aardvarks, elephants and Sirenia) (Shoshani and McKenna, 1998). Molecular studies, however, do not support the monophyly of Ungulata and even spread them among different superorders, Afrotheria and Laurasiatheria (Murphy et al., 2001a,b; Madsen et al., 2001). That Perissodactyla belong to Laurasiatheria is almost uncontested, but the exact position remains unclear. Current hypotheses place Perissodactyla either as a sister group to Cetartiodactyla (Murphy et al., 2001a; Lin et al., 2002) or as a sister group to Carnivora (Murphy et al., 2001b; Arnason et al., 2002b; Amrine-Madsen et al., 2003). Nishihara et al. (2006) found five retroposons supporting Pegasoferae (Perissodactyla + Carnivora + Chiroptera) but also one retroposon supporting Fereuungulata (Carnivora + Perissodactyla + Cetartiodactyla) excluding Chiroptera, a theory inconsistent with Pegasoferae. This inconsistency was attributed to incomplete lineage sorting.

Morphological studies proposed Chiroptera to be within Archonta (Primates + Dermoptera + Scandentia + Chiroptera) (Novacek, 1992b; Shoshani and McKenna, 1998), while molecular sequence analyses place Chiroptera within Laurasiatheria (Miyamoto, 1996; Murphy et al., 2001b). No consensus has been found on their position with Laurasiatheria. It has been placed as sister group to a Carnivora-Artiodactyla clade by the ENCODE analysis (Nikolaev et al., 2007) as well as by retroposon insertions (Kriegs et al., 2006). Nishihara et al. (2006) place Chiroptera within Pegasoferae, whereas Hallström and Janke (2008) propose a clade of Chiroptera, Artiodactyla and Carnivora, but are unable to resolve the exact relationships.

Figure 7.5 shows the analyses of Artiodactyla, Carnivora and Perissodactyla (a) and of Artiodactyla, Carnivora and Chiroptera (b). In the first case, Artiodactyla as sister group to Carnivora is supported by the outgroup analysis with a relatively low score of .645, while the alternative hypothesis of Artiodactyla with Perissodactyla has a score of .355. Even though .645 to .355 is a 2:1 ratio, this is a very weak support for the topology since in 35.5% of the Monte Carlo samples, the latter topology was the most decreasing. For the second problem (row b), a Chiroptera-Artiodactyla clade received moderate support with a pOG score of .814.

Combining these two highest scoring quartets would imply a Chiroptera-Artiodactyla clade with sister group Carnivora and a basal position of Perissodactyla. However,
since the scores suggest a high uncertainty and this topology is incongruent to the fully reconstructed tree (Figure 7.6), a more thorough analysis was performed by analyzing all rooted trees of the four species (Figure 7.5, part c). Furthermore, to reduce the variance of the \( pOG \) scores, the Monte Carlo samples were increased from 200 to 1000. For a rooted tree with four leaves there are fifteen possible topologies. Of these

Figure 7.5: Analysis of the evolutionary relationships of cow, horse, Carnivora (here dog and cat) and bat. The first two rows displays the three possible branching patterns a) for cow, horse and Carnivora and b) for cow, bat and Carnivora. Of the 15 possible trees containing all four groups, only nine are displayed: the three with the highest \( pOG \) scores (i–iii) as well the most common hypotheses from the literature (iv–ix).
fifteen topologies the three with the highest outgroup support (Figure 7.5c i–iii) as well as those found in the literature (Figure 7.5c iv–ix) are shown. Each tree has two internal branches and for each the \( pOG \) value was computed. The monophyly of the Carnivora, cat and dog, is unambiguous and thus not treated here. Two of these trees (i and ii) had relatively high scores for both internal branches. Both of them supported an Artiodactyla-Chiroptera clade, the highest scoring topology had then Perissodactyla as sister group (i) while the second highest scoring topology had Carnivora as sister group (ii). It is interesting that the unusual positioning of Chiroptera and Artiodactyla is always strongly supported, a topology also recovered by most data sets of Hallström and Janke (2008). The Chiroptera-Artiodactyla clade is also found in the next highest scoring tree (iii).

Trees (iv) to (ix) of Figure 7.5 display the \( pOG \) scores for trees found in the literature. Studies on mitochondrial DNA (Reyes et al., 2004), housekeeping genes (Kullberg et al., 2006) and mixed mitochondrial and nuclear genes (Murphy et al., 2001b) support Zooamata (Perissodactyla + Carnivora + Pholidota) with sister group Artiodactyla and Chiroptera in the basal position (v) while retroposons find evidence for Pegasoferae, Zooamata with with sister group Chiroptera and Artiodactyla in the basal position (vi) (Nishihara et al., 2006). Both of these hypotheses are soundly rejected by the \( pOG \) score analysis. Murphy et al. (2007) used a large nuclear and mitochondrial dataset to reestimate the placental tree and divergence times while constraining the monophyly of Atlantogenata because it was strongly supported by their analysis of coding indels. They found topology (vii) and estimated that Chiroptera, Perissodactyla and Carnivora diverged from each other in a remarkably short time frame. The point estimates from these branches were found to be less than two million years apart leading to the suggestion of incomplete lineage sorting as a possible mechanism for adding ambiguity to the branch resolution. In addition to evidence from morphology, Euungulata (Perissodactyla + Cetartiodactyla) proposing a single origin of hoofed animals was weakly supported by an ML tree constructed from genomic data (coding and non-coding) from 41 mammals and 3 other vertebrates (viii) (Prasad et al., 2008). They concluded that the branching within Laurasiatheria is difficult to resolve, susceptible to reconstruction artifacts and may be showing effects of incomplete lineage sorting. Euungulata was also supported with moderate bootstrap support but with Chiroptera rather than Carnivora in the basal position (ix) (Murphy et al., 2001b).
7.3.3 Tree Based on the pOG Score

Figure 7.6 shows the tree of 23 Placentalia that was computed such that the product over the pOG scores on all its branches is maximized. All branches have a pOG value greater than 0.5, and many of the branches have a score of 1.0, indicating a clear and stable pattern of decreasing support from more distant outgroups.

This tree agrees with the trees presented in Section 7.3.2 with only one exception, that the highest scoring topology of Figure 7.5a (Artiodactyla with Carnivora to the exclusion of Perissodactyla) is not consistent with the tree of all species. With a score of 0.645, however, this part of the tree was not well supported and inclusion of the bat changed the topology (Figure 7.5c).

The position of Eulipotyphla (here hedgehog and shrew) basal within Laurasiatheria is expected from Bayesian trees (Murphy et al., 2001b), as well from retroposon insertions (Nishihara et al., 2006).

In the analysis of the tree shrew in the previous section, the discussion was limited to the branching order of the three groups Scandentia, Primates and Rodentia. In the topology of all 23 taxa, the tree shrew is placed more precisely as a sister group to Glires.

Figure 7.6: Tree obtained by maximizing the product of the pOG values on all branches of a tree of 23 Placentalia. The pOG scores for each branch are shown in the figure. Branches not supported by the majority of the gene trees are marked with an asterisk.
Overall, the tree based on the pOG scores on the of 23 species confirms the monophyly of many mammalian orders and super-orders, including Euarchontoglires, Laurasiatheria, Afrotheria, Primates, Rodentia, Lagomorpha, Glires, Carnivora and Eulipotyphla. This demonstrates that the signal from the outgroup analysis is strong enough to build a Placentalian tree which is remarkably close to the current understanding of mammalian evolution. It is also noteworthy that 9 branches of this tree (marked with an asterisk in Figure 7.6) do not have maximum support from the WLS gene trees.

7.4 Conclusions

The pOG score is novel way of assessing the reliability of a branch without directly using the primary phylogenetic method. The advantage of the presented method is in its robustness against model violations. Assuming that if a proportion of the gene trees suffer from reconstruction biases, they are most likely incorrect independently of the outgroup used. If the majority of the gene trees is incorrect due the biased reconstructions, the primary result will also be incorrect. However, those groups not (or less) affected by the bias show the expected behavior of more correct reconstructions using close outgroups and fewer correct topologies with increasingly distant outgroups. Since the pOG score considers only the change of support using different outgroups, it also identifies the correct branch in these cases.

The pOG score should not be confused with a bootstrap value and it does not express the probability that a branch is correct. It quantifies the probability the outgroup support pattern for a branch is as expected. The simulations as well as the application to mammalian sequences presented in this chapter showed a strong correlation between a high pOG score and the correct branch. For the mammalian phylogenies analyzed, pOG values of 0.950 or higher always corresponded to widely accepted topologies, even when the majority of the gene trees supported different branchings. Lower pOG values imply some uncertainty in the outgroup support pattern and were often found in cases where the correct topology is still subject to debate. Nevertheless, the highest pOG value often proposed sensible solutions even in difficult situations such as in the presence of unequal rates among lineages and when only least-squares quartets were used.

To apply the pOG score to a phylogenetic problem, taxa with hundreds of shared sequences and with several known outgroups are required. The ideal application is therefore trees of species with sequenced genomes. As the number of sequenced genomes
is steadily growing, the number of evolutionary problems to which the \( pOG \) method can be applied is also increasing.

The \( pOG \) score’s robustness against model violation and long branch attraction, even when a simple primary method is used makes it both a useful tool for detecting branches that are incorrect due to reconstruction artifacts and a novel optimization criterion for use in heuristic tree searching. The importance of such methods is clearly shown by the many conflicting species trees published to date.
List of Tables

2.1 Summary of codon models ........................................ 24
3.1 Genome pairs used for estimating the codon matrix .......... 28
3.2 Statistics about the codon counts matrix ..................... 29
4.1 Example of the SynPAM calculation ............................ 48
4.2 Coefficients of variance for synonymous distances .......... 51
4.3 Percentages of correctly reconstructed splits ................ 56
4.4 CVs of species divergences .................................... 60
5.1 Overview of articles about the human, dog, mouse topology . 70
5.2 Results from distance trees of eight species ................. 71
5.3 Parsimony analysis of four species ............................ 72
5.4 Parsimony analysis after removing repetitive regions ....... 74
5.5 Results of ML analysis of four species ....................... 76
6.1 Example of a compatibility table ............................... 82
7.1 Example of pOG score computation ............................ 99
7.2 List of species used in pOG analysis .......................... 104
List of Figures

3.1 CodonPAM vs. PAM curve ................................................. 31
3.2 Alignment accuracy plot ................................................. 34
3.3 Score comparison on vertebrate sequences ......................... 36
3.4 Score comparison on cyanobacteria sequences ...................... 37

4.1 Illustration of counting methods ...................................... 42
4.2 Histogram of synonymous distance estimates ....................... 52
4.3 Histogram of synonymous distance estimates over longer distances 53
4.4 Reference species tree ................................................... 55
4.5 Plot of reconstruction percentages ................................... 57
4.6 Linearity of distance estimates with time ............................ 59

5.1 Rooted and unrooted trees of human, dog and mouse ............... 69
5.2 PAM distance tree of eight mammals ................................ 71
5.3 Distribution of $\alpha$ values ........................................... 76

6.1 Example of conflicting quartets ....................................... 81
6.2 Simulations of distance estimation under model violations .......... 87
6.3 Trees used for simulations .............................................. 88
6.4 Simulations of quartet reconstruction under model violations .... 89
6.5 Simulations of conflicting quartet reconstruction under model violations 90

7.1 Tree used for simulations ............................................... 101
7.2 Simulation results of $pOG$ score using gamma-distributed site rates 107
7.3 Simulation results of $pOG$ score using invariant sites . . . . . . . . . . . . . . 108
7.4 Outgroup analysis of five phylogenetic problems . . . . . . . . . . . . . . . . . . 110
7.5 Outgroup analysis of cow, horse, Carnivora and bat . . . . . . . . . . . . . . . . 114
7.6 Tree maximizing the global $pOG$ score . . . . . . . . . . . . . . . . . . . . . . . 116
Bibliography


Curriculum Vitae

- Born on 14 April 1979
- Citizen of Wetzikon and Volketswil, Switzerland

Education


Research Experience

- 2004–2009, PhD student in Gaston Gonnet’s group, Computer Science Department, ETH Zurich.
- 2004, Master thesis supervised by Gina Cannarozzi in Gaston Gonnet’s group at ETH Zurich entitled *Codon Transition Matrices and their Applications*.
- 2003, semester thesis supervised by Thomas Holenstein in Ueli Maurer’s group at ETH Zurich entitled *Attacken auf das McEliece Public-Key Kryptosystem*.

137
Awards and Honors

- UNICIBLE prix académique 2005 (for the highest average grade), Computer Science Department, ETH Zurich.
- Highest average grade for type C matura at Kantonsschule Oerlikon, 1999.

List of Peer-reviewed Publications

