universität wien

DISSERTATION

Titel der Dissertation

Maximum Likelihood Estimation of Gene Family Specific Duplication and Deletion Rates

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr. rer. nat.)

Verfasserin:	Andrea Führer
Matrikel-Nummer:	0647184
Dissertationsgebiet (lt. Studienblatt)	Molekulare Biologie (Bioinformatik)
Betreuer:	Univ.-Prof. Dr. Arndt von Haeseler

Wien, am 29. September 2007

Preface

This thesis summarizes my research at the bioinformatics group of the Department of Computer Science at the University of Düsseldorf and at the Center for Integrative Bioinformatics Vienna (CIBIV), where I was working as a research assistant since October 2003. It was supported by the Vienna Science and Technology Foundation.

First, I would like to thank my supervisor Arndt von Haeseler for giving my the opportunity to work at his chair. I appreciate his support and confidence in my work.

I would also like to thank Prof. Dr. Reinhard Bürger and Prof. Dr. Volmar Liebscher for accepting the task to read this thesis as referees.

Another word of thanks goes to my colleagues at the bioinformatics group in Düsseldorf and at the CIBIV, especially to Tanja Gesell, my fellow student since the time of my diploma thesis, Ingo Ebersberger for help with the assembly of the Inparanoid dataset and proofreading the manuscript, Ingo Paulsen for fruitful discussions in the beginning of this work, and Heiko Schmidt for technical support. Furthermore, I want to thank Matthew Hahn for making the mammalian dataset available and for fruitful discussions during my participation in a couple of conferences.

I want to extend my thanks to the members of the IGFZS for the pleasant time outside the university in Düsseldorf.

My special thanks go to my parents for their constant support on my way so far.

Last of all, and most of all, I want to thank Christopher for supporting me through all ups and downs and for making me want to leave at the end of each day.

Vienna, September 2007

Abstract

Genes can be divided into gene families, which are defined by homology, and let presume that the genes evolved from a common ancestral gene. In this content the size of a gene family is subject to evolutionary change. Variations in gene family size can be influenced by several biological processes. We are interested in the evolution of a gene family affected by gene duplication and deletion along a phylogenetic tree. To model the gain and loss of gene family members due to duplications and deletions respectively, we use a birth and death process. More specifically we model the change in the number of gene copies of a gene family along a phylogenetic tree. Based on the number of gene copies in a set of extant species and the appropriate species tree, a maximum likelihood approach has been applied to infer the birth rate and the death rate due to duplications and deletions, respectively. Furthermore the number of gene copies of the most recent common ancestor of the sample can be determined. In contrast to recently published methods, our method allows for the estimation of these parameters specific for every individual gene family.

To validate the method, simulation studies were performed. Assuming a fixed number of gene copies for the ancestor and specific rates for duplication and deletion of genes, the evolution of the number of gene copies along a phylogenetic tree can be simulated. Such simulated data showed the high variation of the process. Using our maximum likelihood framework, the rates and the ancestral gene number used for the simulation were estimated back subsequently. A collection of simulation studies using several phylogenetic trees with different combinations of duplication rate, deletion rate and ancestral gene number showed that the method can infer the parameters quite good. However, both rates were slightly underestimated, which is caused by the fact, that multiple repetitive duplications and deletions which do not change the absolute gene number cannot be detected in general.

We further applied our method to biological gene family data from vertebrates of the Inparanoid and the Ensembl databases. Compared to previous reported rates our estimates are about one magnitude lower. The data was also considered with regards to model violations, since it is assumed that e.g. large-scale duplication, like whole genome duplications, have occurred during evolution. Hence, extensions of our method and future work will be discussed.

CONTENTS

Introduction

Already 1970 Susumo Ohno recognized, that gene and genome duplications are the major forces by which the genetic raw material is provided for increasing complexity during evolution.

”It becomes quite clear that while allelic changes at already existing gene loci suffice for radial differentiation within species ..., they cannot account for large changes in evolution, because large changes are made possible by acquisition of new gene loci with previously non-existent functions. ... Thus, gene duplications emerges as a major force of evolution” (Ohno (1970))

More than thirty years after his well respected book ”Evolution by gene duplication” more research than ever is being carried out on the evolutionary significance of gene and genome duplications and the contribution of these mechanisms to the advances in genomic and organismal evolution. Besides the precise mechanisms, the rates at which gene duplications and moreover gene deletions have occurred in the evolution of different genomes are highly interesting. Studies on this topic requires the knowledge of the entire genomic information of the organisms under consideration. With the increasing availability of such genome-wide data different methods were proposed to infer gene duplication and deletion rates.

In this thesis a model is introduced to describe gene duplications and deletions along a phylogenetic tree. Using a maximum likelihood approach, we will shown how rates for these events can be inferred based on gene family data. To our knowledge this is the first method to estimate duplication and deletion rates simultaneously specific for individual gene families.

Chapter1. In the first chapter, background information about the evolution of genomes, phylogenetic trees as a medium to analyze the evolutionary history of genes, and mechanisms and consequences of gene duplications and deletions are presented. Furthermore, a short review of methods used for the estimation of rates for duplications and deletions so far, as well as explicit values for these rates from different studies are given. This introduction is necessarily a subjective and short summary, but should be sufficient to provide all information needed to place the rest of the thesis in context.

Chapter2. The second chapter includes the description of the birth and death process, which we will use to model gene duplications and deletions. Two different strategies for the estimation of model parameters e.g. duplication and deletion rates will be introduced and the application of these strategies for our model in connection with a phylogenetic tree demonstrated.

Chapter3. Due to the high complexity of the computations associated with the parameter estimation, much time was spent to ensure the accuracy of this computation. In chapter 3 we explain the number representation in the computer and the resulting difficulties for arithmetic operations. Based on selected examples, problems during the estimation of our model parameters are pointed out and solutions are presented.

Chapter4. The performance of the presented estimation methods was evaluated. Therefore we simulated gene family data for different phylogenetic trees assuming different values for the duplication and the deletion rates. Both estimation methods were subsequently used to infer the model parameters for the simulated data. The quality of the results of these estimation methods and the applicability for real data will be discussed.

Chapter5. Further, we applied our method to real data. Chapter 5 includes the studies on real datasets from the HOGENOM, the Inparanoid, and the Ensembl databases. Among other things, these studies show how important well-assembled data for the quality of the results can be. The estimated gene duplication and deletion rates of these studies will be compared to previous published rates and we will discuss the benefits and limits of our method.

Chapter6. In the last chapter some extensions of the model are discussed, since in reality there are more complex mechanisms of gene duplication, as we have considered in this study. Furthermore, we will give an outlook on possible future work.

CHAPTER ONE

Genome evolution in the light of gene duplication and deletion

All living organisms share a common history. Inferring these evolutionary relationships between different organisms has been and still is a major interest in biology. While originally solely morphological characteristics were used, the increasing availability of DNA and protein sequences, made molecular data become a powerful medium to reveal the evolutionary relationships. The entire hereditary material carried by an individual is called genome. Due to recent improvements in the DNA sequencing technology, projects to determine the sequence of entire genomes are now common in practice. Many of these genome projects are completed. Today 23 completely sequenced genomes of eukaryotes are available, among four from animals (*Caenorhabditis elegans* (roundworm), *Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse), and *Homo sapiens* (human)). Further 53 draft genomes from animals exist and 21 thereof are from mammals, like *Rattus norvegicus* (rat), *Pan troglodytes* (chimp), and *Canis familiaris* (dog) (Genome Project Database (2007), issue September 2007). This data does not only allow for studies on the relationship of organisms, but also for studies on the evolution of genomes. Segments of the genome that give rise to a functional product, either a functional RNA or a protein, are defined as genes. The entire genome, and thereby also all genes, is subject to evolutionary change through different biological processes.

Although, nowadays a lot is known about small changes in the DNA sequence of a gene, called nucleotide substitutions, and the influence of such changes on the evolution of organisms, still little is known about the creation of new genes. Understanding this phenomenon is essential, since it is not likely that an existing gene adopts a new function while keeping its old one. More probable is the development of a new functions through new genes that can perform them. But how do new genes arise? The most obvious way in which a new gene can be created is via duplication, where direct copies of a gene or groups of genes are made. These copies can then diverge in their function and genes with new functions can arise (Ohno (1970)). This is an important contribution to the evolution of genomes.

By this means, also the number of genes in the genome changes, which would lead to dif-

organism		estimated	estimated	average gene density	chromosome
		genome size	gene number	1 gene per	number
human	(*Homo sapiens*)	2,900 Mbp	30,000	100,000 bases	46
rat	(*Rattus norvegicus*)	2,750 Mbp	30,000	100,000 bases	42
mouse	(*Mus musculus*)	2,500 Mbp	30,000	100,000 bases	40
fruit fly	(*Drosophila melanogaster*)	180 Mbp	13,600	9,000 bases	8
plant	(*Arabidopsis thaliana*)	125 Mbp	25,500	4,000 bases	10
roundworm	(*Caenorhabditis elegans*)	97 Mbp	19,100	5,000 bases	12
yeast	(*Saccharomyces cerevisiae*)	12 Mbp	6,300	2,000 bases	32
bacterium	(*Escherichia coli*)	4.7 Mbp	3,200	1,400 bases	1
bacterium	(*Haemophilus influenzae*)	1.8 Mbp	1,700	1,000 bases	1

Mbp . . . million base pairs.

Table 1.1: Genome sizes and gene numbers of humans and other organisms from Human Genome Project (2007). Related publications: International Human Genome Sequencing Consortium (2001), Rat Genome Sequencing Project Consortium (2004), Mouse Genome Sequencing Consortium (2002), Adams *et al.* (2000), The Arabidopsis Genome Initiative (2000), The C. elegans Sequencing Consortium (1998), Goffeau *et al.* (1996), Blattner *et al.* (1997) and Fleischmann *et al.* (1995).

ferent numbers of genes for different organisms. A comparison of the numbers of genes for completely sequenced organisms reveals these differences, as shown in table 1.1. Eukaryotes have significantly more genes than prokaryotes, so there must have been an increase in the number of genes, given that eukaryotes have evolved from an prokaryote-like ancestor (Graur & Li (2000)). As we will see, gene duplications and gene deletions have played a major role in the evolution of genomes.

1.1 The evolution of gene families

1.1.1 Gene families

Most of the known genes belong to big and intensively studied gene families. A gene family is a group of genes that have evolved from a common ancestral gene through divergence and gene duplication. Members of a gene family are defined by *homology*, which is often concluded on the basis of sequence similarity. There are two different types of homology. If a speciation event, the divergence of a species into two separate species, was the reason for the divergence of genes, the genes are called *orthologs*. Thus, orthologs are genes from different species. On the other hand, if genes diverged as a result of a gene duplication, the genes are called *paralogs* (Graur & Li (2000)). Normally the function of the genes of a gene family will be similar, but in contrast to the required homology of the sequences, the functions of the gene family members are allowed to be different. Given the successive process of gene duplication and divergence over evolutionary time, the relatedness of genes within a family can vary, as can the number of gene family members for different species. For understanding the relationships among genes of a gene family and reconstructing evolutionary events, phylogentic analyzes can be useful. To represent evolutionary relationships phylogenetic trees are used. Therefore, in the following subsection a brief introduction into phylogenetic trees is given.

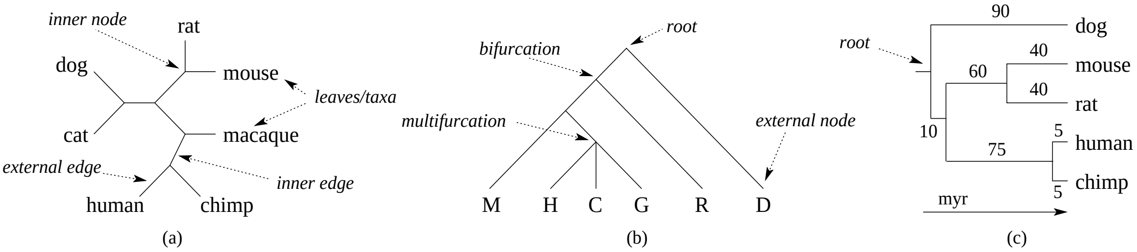


Figure 1.1: Notations on different phylogenetic leaf-labeled trees. (a) Unrooted binary tree. (b) Rooted multifurcating tree. In (a) and (b) edge lengths have no meaning. (c) Rooted binary tree, edge lengths correspond to times between speciation events in myr. If all leaves are contemporaries the lengths of all path from the root to the leaves should be equal.

1.1.2 Phylogenetic trees

A phylogenetic tree or also called evolutionary tree is a tree in the mathematical sense, composed of nodes and edges. The nodes are referred to as taxonomic units or taxa, which can be species, genes, proteins etc. In general, the external nodes represent contemporary taxa, like living species, whereas inner nodes represent hypothetical common ancestors of their descendants. Since we have no knowledge about the inner nodes, only the external nodes get labels and therefore phylogenetic trees are so called leaf-labeled trees. The edge lengths of the tree can have no meaning, serving only to illustrate the relatedness of the taxa, or they can correspond to the number of differences between them, as well as to time estimates (figure 1.1).

A distinction is drawn between rooted and unrooted phylogenetic trees. In a rooted tree one node is marked as root and all edges are directed away from the root. The root represents the most recent common ancestor (MRCA) of all entities at the leaves of the tree. An unrooted phylogenetic tree illustrates the relatedness of the leaves without making assumptions about the ancestry. Unrooted trees can be rooted by using an outgroup, which should be related to the taxonomic units at the leaves, but are known to have branched off before the evolution of the taxonomic units at the leaves has started.

All phylogenetic trees can be either bifurcating or multifurcating. In a bifurcating or binary rooted tree every inner node has exactly two descendants arising from it, which coincides to the concept that speciation occurs through splitting of one lineage into two. On the other hand a multifurcating rooted tree may have more than two descendants arising from each inner node.

The phylogenetic trees we will consider in this thesis are rooted leaf-labeled binary trees.

1.1.3 Species tree versus gene tree

Species trees are phylogenetic trees. The leaves correspond to different species and the tree presents the evolutionary relationships of the species. The root of a species tree represents the MRCA of all species in the tree and the edge lengths often correspond to time in myr.

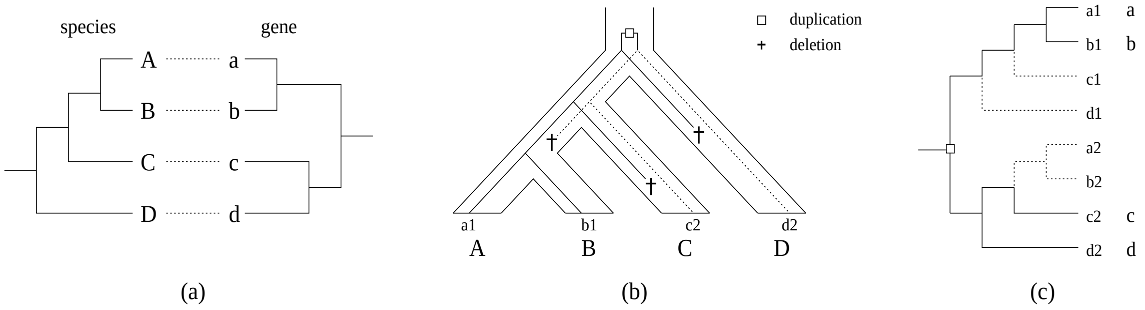


Figure 1.2: Gene duplication and deletion can introduce incongruence between gene trees and species tree. (a) Species tree for four species (A,B,C,D) and gene tree for four genes (a,b,c,d) from the species differ from each other. (b) Gene tree inside the species tree. The difference can be explained by one gene duplication (□) at the root of the gene tree and three gene deletions (†) (minimum number of events). More details can be found in the text. (c) Other representation: Reconciled tree for the species tree and gene tree of (a). Modified from Page & Charleston (1997).

The reconstruction of species trees is difficult. On the one hand, it is very hard to specify the times for speciation events. Therefore different types of molecular data, different statistical methods, and different calibration points are used (e.g. see Purvis (1995), Waddell *et al.* (1999), Wray (2001), Hedges (2002), Benton & Ayala (2003), Glazko & Nei (2003)). On the other hand, analyzes of different parts or genes of the genome can lead to different phylogenetic trees. These trees are called *gene trees*. With the reconstruction of more and more gene trees a big problem has arisen - the incongruence among gene trees and between gene trees and species tree (figure 1.2 (a)). A very well studied example in this context is the globin family (Strachan & Read (2003), Hardison (2006)).

Since there can only be one 'true' species tree, there should be a biological explanation for the occurrence of gene trees different from this species. In this context gene duplications and deletions may help to reconstruct the evolutionary history of a gene. In figure 1.2 an example for this problem is illustrated. We have a species tree (((A,B),C),D) which is different from the gene tree ((a,b),(c,d)) for a specific gene from these species. If we assume, that in the evolution of this gene, duplications and deletions have occurred, we can explain the difference between the species and the gene tree as follows: prior the the speciation event at the common ancestor of the species, one duplication occurred. In the following two independent copies of this gene existed, which evolved further according to the species tree (indicated by the solid and the dashed line in figure 1.2 (b)). One gene copy got lost in the branch leading to the ancestor of A and B. The other gene copy got lost in the branch leading to C and also in the branch leading to D. Since the two gene copies evolved independently of each other, genes which are descended from one of the copies share more similarity than genes from different copies. This in turn will lead to branching pattern of the gene tree ((a,b),(c,d)), although the gene has evolved along the species tree.

In this context, methods like tree reconcilation were developed (Page (1994), Eulenstein

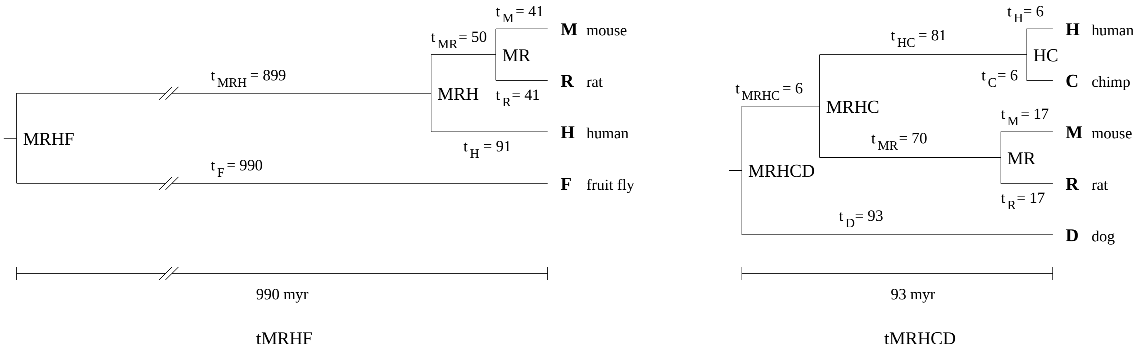


Figure 1.3: Species trees. Edges are denoted with t_X with X being the node where the edge goes in. The edge lengths are times estimates given in myr. tMRHF denotes the species tree for mouse, rat, human and fruit fly. Time estimates from Hedges (2002). tMRHCD denotes the species tree including mouse, rat, human, chimp, and dog. Time estimates from Demuth *et al.* (2006).

et al. (1997), Chen *et al.* (2000)). These methods try to minimize the number of events (duplication, deletion), which are necessary to explain the differences between species tree and gene tree. In figure 1.2 (c) the reconciled tree for the species and gene tree from (a) is shown. Reconciled trees can be used to find the species tree. Using a set of inferred gene trees and a set of potential species trees, the cost (number of duplications and deletions) for all reconciled trees can be computed. For each species tree the costs of all reconciled trees are added up and the species tree with the lowest cost is chosen (Page (2000), Cotton & Page (2002)).

In our studies mainly two different species trees were used (figure 1.3). One four taxa species tree including mouse, rat, human and fruit fly, whose common ancestor dates back to 990 myr (Hedges (2002)). The other one is a five taxa species tree including only mammalian species: mouse, rat, human, chimp and dog. The common ancestor for these species dates back to 93 myr before present (Demuth *et al.* (2006)).

1.2 Gene duplications and deletions

As briefly mentioned above, the duplication of a gene means, that the part of the DNA that comprises the gene is copied. So two equal duplicates of the gene arise, which evolve from there on independently of each other and underlie effects, like mutations, on their own. It is assumed that one copy retains the original function of the gene, whereas the function of the other copy is free to change (Ohno (1970)).

On the other hand, the deletion of a gene means either the loss of a part of the DNA, or the deactivation of the gene. Thereby the function of the gene is lost, if no other gene has the same function or can take over the function. Gene deletions are assumed to occur often after gene or genome duplications and possibly are subject to purifying selection, which is selection that prevents the fixation of deleterious mutations (Wagner (2002), Bertrand

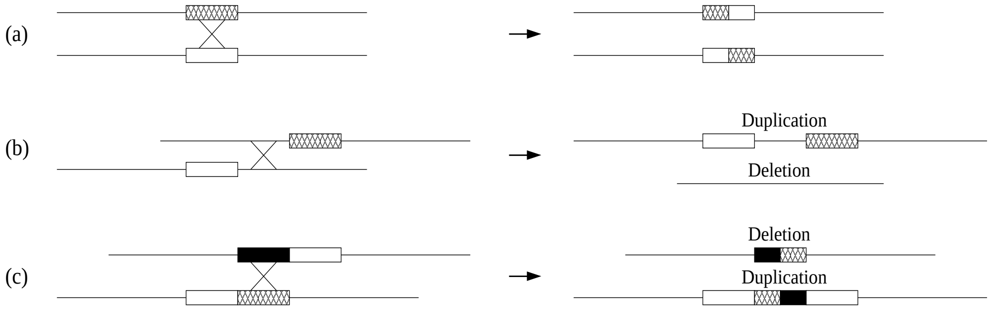


Figure 1.4: (a) Normal crossing-over which results in the exchange of some parts of the chromosome. (b) Unequal crossing-over resulting in the deletion of a DNA sequence on the lower strand and a duplication of the same part on the other strand. (c) With the increase of tandemly duplicated regions, the chance of mismatches increases as well and unequal crossing-over can occur more easily. Modified from Li (1997).

et al. (2004), Hughes (2005), Roth *et al.* (2006)).

1.2.1 Mechanisms of gene duplication and deletion

One distinguishes between partial or internal gene duplication, complete gene duplication, partial chromosomal duplication, complete chromosomal duplication, and polyploidy, or genome duplication (Graur & Li (2000)). The first four categories are also referred to as regional duplications because they only affect parts of the genome. On the other hand, genome duplications have an impact on the entire genome of an organism.

Unequal crossing-over. The main mechanism for gene duplication is assumed to be unequal crossing-over (Graur & Li (2000)). During meiosis two homologous chromosomes can exchange some parts of their DNA, which is called crossing-over. Normally crossing-over occurs between homologous regions of these chromosomes (figure 1.4 (a)), but if similarities in the DNA sequences results in an alignment of non-homologous regions of these chromosomes, unequal crossing-over can occur. That results in a tandemly duplicated region on one chromosome and a complementary deletion on the other one (figure 1.4 (b)). Repeats in DNA sequences make unequal crossing-over more likely to occur. As a consequence, tandemly duplicated regions have a higher probability to get duplicated or deleted again (figure 1.4 (c)). This mechanism can involve DNA segments up to Mb in length.

Replication slippage. Another mechanisms for duplications or deletions of DNA segments is replication slippage (Graur & Li (2000)). This type can occur during DNA replication in regions that contain short repeats. Thereby a mispairing of neighboring repeats results in rather small duplicated or deleted DNA segments (up to 20-30 nucleotides).
There are also other mechanisms for gene duplications or deletions, like intrastrand deletion and DNA transposition, as well as chromosomal non-disjunction, which lead to the

duplication of an entire chromosome (Graur & Li (2000)). But they will not be discussed further here.

Pseudogenization. Gene deletions can also happen independent of duplications. A lot of mutations in DNA sequences are deleterious, in fact far more mutations are deleterious than advantageous (Graur & Li (2000)). Some of these deleterious mutations might be able to be fixed, which means that these mutations are kept in the population. But most of the deleterious mutations will lead to a destruction of the relevant gene, so that it is not longer capable to produce a functional gene product. That does not necessarily reduce the fitness of an organism, if more than one copy of the gene exists, which can provide the function further on (Haldane (1932)). The inactivation of a gene due to deleterious mutations results in an, so called, unprocessed pseudogene. Most pseudogenes arised from a duplicated functional gene, that became nonfunctional afterwards (Graur & Li (2000)). Therefore they are frequently found in the neighborhood of the functional gene, from which they have been derived. Nevertheless, there are also pseudogenes found far away from its original copy due to rearrangements in the genome, e.g. in the globin family (Strachan & Read (2003), Hardison (2006)). Because of the lack of any selection pressure to conserve them, pseudogenes can be expected to be eliminated from the genome by various DNA turnover mechanisms after a while (Lynch & Conery (2000)).

1.2.2 Whole genome duplications - 2R hypothesis

Whole genome duplication (WGD) is a special type of gene duplication. In this case the entire genome is doubled. It can be caused e.g. by meiotic irregularities, which produce gametes with unreduced chromosome number (Graur & Li (2000)). Ohno (1970) described WGD as a more important mechanism compared to regional duplications, because the whole environment of a gene, like promoter sequences, regulatory elements etc. for each duplicated gene is presents afterwards. The lack of these parts of the genome in the duplicate can otherwise influence or inhibit the functionality of the gene.

Polyploidy. There two main types of polyploidy which are important with regard to WGD: allotetraploidy and autotetraploidy. The first type is based on the combination of genetically distinct genomes. So there are two similar but non-identical genomes present. This type is very often found in plants, especially in flowering plants. On the other hand, autotetraploidy is the doubling of an individual genome or the combination of genomes from two individuals of the same species. That leads to a symmetrical genome, where all genes are then present in double the number of the previous copies with exactly the same genetic information. Autotetraploidy could be detected in almost all organisms, from protists to mammals, but is very rare (Nagl (1990)).

2R hypothesis. As already mentioned, vertebrates have much more genes than invertebrates (table 1.1). WGDs seem to be a fast and easy way to increase the number of genes rapidly, and so might be responsible for this higher number of genes. With the discovery of four Hox gene cluster in vertebrates compared to one single cluster in most invertebrates (Garcia-Fernàndez & Holland (1994); Bailey *et al.* (1997); Holland (1997); Meyer & Málaga-Trillo (1999)), the idea arose that the larger genome results not from only one, but from two rounds of WGD during the origin of vertebrates (see figure 1.5 and Holland *et al.* (1994)). This is known as the '2R-hypothesis' in literature, whereas the 'R' stands for rounds. This basic tenet is still debated and every now and then new studies emerge supporting this hypothesis or refuting it (Skrabanek & Wolfe (1998), Hughes (1999), Wang & Gu (2000), Wolfe (2001) Gu & Huang (2002), Dehal & Boore (2005), Blomme *et al.* (2006)). Another highly discussed aspect is when these WGDs could have occurred (Skrabanek & Wolfe (1998), Gu *et al.* (2002a), Panopoulou & Poustka (2005)).

Effects. Following genome duplication, and a transient intermediate state, large-scale chromosome rearrangements could be a reason for chromosome divergences and the reestablishment of the number of chromosome sets with then twice the number of chromosomes. Furthermore, the pseudogenization and total loss of many unnecessary genes can be expected. If many of the duplicated genes get completely lost afterwards, the evidence for a previous WGD could be really sparse. Such consequences of WGDs, especially for the two rounds of WGD at the origin of vertebrates, will be discussed in more detail in chapter 6. Further information also can be found in the review of Panopoulou & Poustka (2005) and Roth *et al.* (2006).

1.2.3 Evolutionary fate and biological relevance of duplicated genes

Duplications occur in an individual and can be fixed or lost in the population. Two copies of a gene in one genome can be an advantage, a disadvantage or neither of them. If the duplicate is deleterious there will be selection against the fixation of the gene and it has a higher probability to get lost. That is also possible for neutral and advantageous duplicated genes, but there is no selection against them.

Pseudogenes. If a mutation occurs, the gene is likely to become a *pseudogene* (see also subsection 1.2.1), which is either no longer expressed or functionless. Relatively young pseudogenes are easy to detect, because of high sequence similarity to the original gene. In *C. elegans* e.g. 2168 pseudogenes were found, which is about one pseudogene per eight functional genes (Harisson *et al.* (2001), Zhang (2003)). Compared to human, that is very few, since one pseudogene per two functional genes were found in chromosomes 21 and 22 in humans (Harisson *et al.* (2002), Zhang (2003)). Most duplicated genes will become pseudogenes and will disappear within the first few myr after duplication (Lynch & Conery (2000)).

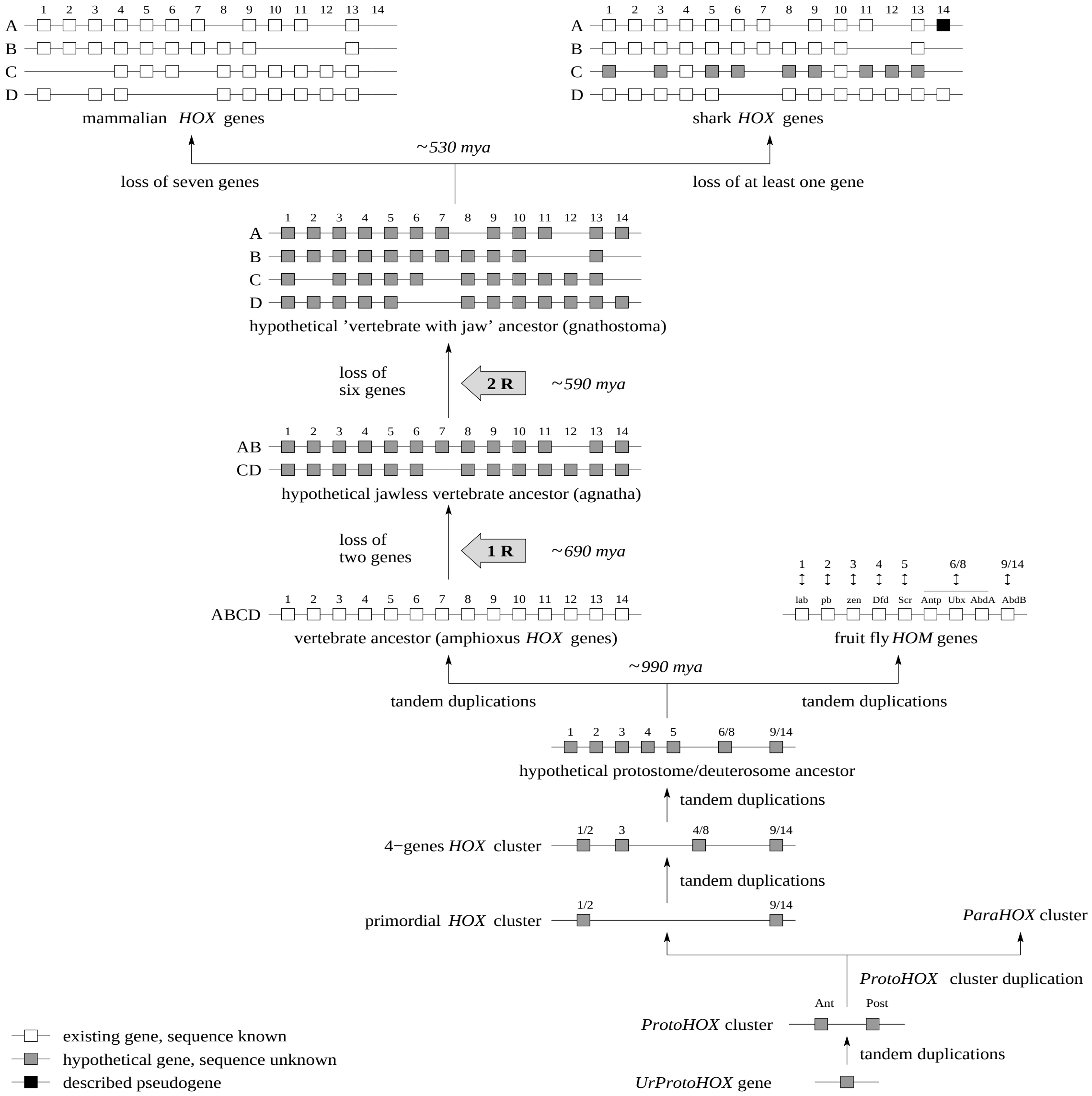


Figure 1.5: A hypothesis of the HOX cluster evolution, including results of Garcia-Fernàndez (2005) and Hoegg & Meyer (2005). Shown is the two-gene model (one Anterior and one Posterior) of the ProtoHOX cluster which gave rise to the primordial, two gene containing, HOX and ParaHOX clusters. By single tandem duplications the four-gene containing HOX cluster emerged which further expanded to the seven-genes containing cluster for the protostome/deuterotsome ancestor, again by tandem duplications. From this protostome/deuterostome ancestor the structure of the insect HOX (HOM) cluster arose, as well as the HOX cluster (ABCD) of the last vertebrate ancestor, due to tandem duplications independently in both clusters. One WGD (1R) around 690 mya and two gene losses gave rise to the two HOX gene clusters (AB and CD) in the jawless vertebrate ancestor. A second WGD (2R) around 590 mya in conjunction with six gene losses lead to the last ancestor of vertebrates with jaw. Sharks and mammals still maintained a four-cluster (A,B,C and D) state. Open squares indicate genes that have been previously described and black squares indicate pseudogenes. Grey squares are unknown ancestral genes or genes that have not been sequenced yet, but probably are present in the cluster. Approximate phylogenetic timing of the WGD are averaged from Panopoulou & Poustka (2005), speciation times are from Hedges (2002) and Hoegg & Meyer (2005) and are given in million years ago (mya). Modified from Málaga-Trillo & Meyer (2001), Garcia-Fernàndez (2005) and Hoegg & Meyer (2005).

Gene conversion and purifying selection. A duplicated gene can be advantageous by retaining its original function, simply because an extra amount of the gene product is provided. This can be very important for highly expressed genes. One way to preserve the original function of a duplicated gene is the so called *gene conversion*, which is a recombination process and leads to the replacement of one sequence by another (Graur & Li (2000)). Thereby both genes maintain the same sequence and the same function. Gene conversion between duplicated genes has been found in all analyzed species and in every part of the genome. Differences in the rate and the probability of occurrence could be detected in different parts of the genome.
However, Nei *et al.* (2000) and Piontkivska *et al.* (2002) suggest another process to be much more important in the maintenance of original functions of duplicated genes. That is, strong *purifying selection* against mutations that modify gene functions (Zhang (2003)).

Subfunctionalization. Although two genes with the same function can be advantageous, it can also happen that both genes, original and duplicate, develop slightly different functions. If each gene copy takes over a part of the functions of the original gene, it is more probable that both genes copies are retained (Nowak *et al.* (1997)). This process is called *subfunctionalization*. There were studies on the subdivision of functions following gene duplication, which gave reason to the existence of subfunctionalization (Hughes (2005)). But not until the work of Piatigorsky & Wistow (1991) there was a pertinent example for it. They found the phenomenon of 'gene sharing' for a single functional gene which can serve as an enzyme or as a crystallin depending on the location of the gene carrying cell in the organism. Michael Lynch and colleagues (Lynch & Force (2000), Lynch *et al.* (2001)) discussed possible mechanisms for subfunctionalization. If a gene with two functions gets duplicated, it may happen that one gene copy completely looses one function and the other copy completely looses the other function. In doing so, both genes become indispensable and will be protected against deleterious mutations. Since a number of genes have multiple functions (Hughes (2005)), gene duplication and specialization of the two gene copies by dividing the ancestral gene function might be a prevalent mode of gene evolution.

Neofunctionalization. So far non of the mentioned possibilities for the fate of a gene after duplication really give rise to genes with completely new functions. Susumu Ohno (Ohno (1970)) assumed that, after gene duplication, one of the gene copies would be entirely redundant and free to change in any directions. Such a gene could emerge and obtain a novel gene function.
Although it seems improbable that a gene would be changed so much, that an entirely new function emerges, examples for it were found (Zhang (2003)). In many cases a related function, similar to the original one, will evolve after gene duplication. This procedure is known e.g. for the red- and green-sensitive opsin genes in human. Mainly two changes in the DNA sequence are responsible for the sensitivity to the wide range of colors that human have (Yokoyama & Yokoyama (1989), Asenjo *et al.* (1994), Zhang (2003)). This

process is called *neofunctionalization*.

Discussion. Gene duplications and especially genome duplications seem indeed to be very important, because they allow the generation of large amounts of raw genetic data in a relatively short time. This genetic material can be changed by mutations, genetic drift and positive selection and by that genes with the same, specialized or new function can arise. Without gene duplications it seems difficult to imagine, how systems with many similar genes, like the vertebrate adaptive immune system, could have evolved (Zhang (2003)).

1.3 Rates of gene duplications and deletions

Although a lot of mechanisms leading to gene or genome duplication are known, the rates with which genes duplicates or get lost are uncertain. One big problem was the lack of genome wide information of species to be investigated in the past. Another more basal problem is that only duplications can be observed which have been preserved in contemporary genomes. Since many duplicated genes are likely to get lost after some time, inference about the 'true' rate of gene duplication as well as the rate of gene deletion is very difficult (Friedmann & Hughes (2003)). In principle two different approaches were used to estimate duplication and deletion rates.

Studies based on the age-distribution of gene duplicates. In the first approach, paralogous genes are identified based on a set of gene families. These genes are known to result from gene duplications. By comparing rates of nucleotide substitutions the ages of the duplicates and therewith the age-distribution of gene duplicates within a genome can be inferred. The age-distribution in conjunction with a model describing this distribution can then be used to estimate duplication and deletion rates, by fitting the model to the observed age-distribution. The models used depend on different parameters. In some studies only a rate for duplications is required, whereas other models depend on duplication and deletion rate. In this context simple birth and death models were applied. The rate estimates from Lynch & Conery (2000, 2001); Gu *et al.* (2002b,a); Lynch & Conery (2003); Rat Genome Sequencing Project Consortium (2004); Cotton & Page (2005) were inferred using this approach with different models and different gene family data. In the following their results are summarized.

The first study to estimate gene duplication and deletion rates from genomic data was conducted by Lynch & Conery (2000). They suggested an estimate of the average gene duplication rate on the order of 0.01 per gene per myr. More specifically they suggested a duplication rate of 0.0023 gene^{-1} myr^{-1} for *Drosophila melanogaster* (fruit fly), 0.0083 gene^{-1} myr^{-1} for *Saccharomyces cerevisiae* (yeast) and a much higher rate of 0.0208 gene^{-1} myr^{-1} for *Caenorhabditis elegans* (roundworm). They also analyzed mouse and

year	yeast	roundworm	fruit fly	human	rodent	mammals	vertebrates
	duplication rates in gene^{-1} myr^{-1}						
2000/2001	0.0083[1]	0.0208[1]	0.0023[1]	0.0071[2]	-	-	-
2002	0.028[3]	0.024[3]	0.0014[3]	-	-	-	-
2003/2004	0.004[4]	0.016[4]	0.001[4]	0.009[4]	0.00195[5]	-	-
2005/2006	-	-	0.002[8]	-	-	0.0016[9]	0.00097[6]/0.00115[7]
	deletion rates in gene^{-1} myr^{-1}						
2003/2005	-	-	0.002[8]	0.0924[4]	-	0.0016[9]	0.00048[6]/0.0074[7]

Table 1.2: Survey of already published duplication and deletion rates for yeast (*Saccharomyces cerevisiae*), roundworm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), human (*Homo sapiens*), rodent (mouse (*Mus musculus*) and rat (*Rattus norvegicus*)) and a group of mammals (mouse, rat, human, chimp (*Pan troglodytes*), dog (*Canis familiaris*)). All rates are given per gene per myr. [1]Lynch & Conery (2000), [2]Lynch & Conery (2001), [3]Gu *et al.* (2002b), [4]Lynch & Conery (2003), [5]Rat Genome Sequencing Project Consortium (2004), [6]Cotton & Page (2005) time window: ∼4700 mya to present, [7]Cotton & Page (2005) time window: ∼200 mya to present, [8]Hahn *et al.* (2005), [9]Demuth *et al.* (2006).

human, but gave no estimates (because the complete genomic sequences of both species were not available at that time), except a half-life of duplicated genes of 7.3 myr. This value was substantially higher than the half-life estimates for fruit fly and roundworm, which range around 2.9 myr. The work was criticized by Long & Thornton (2001) and Zhang *et al.* (2001) in terms of methods and data. Thus, Lynch & Conery revised their estimate for the half-life of duplicated human genes (about double of their previous estimate: 16 myr) and found the duplication rate to be about 0.0071 gene^{-1} myr^{-1} for human (Lynch & Conery (2001)).

Gu *et al.* (2002b) published additional estimates for fruit fly, yeast and roundworm one year later. They used different criteria and estimated recent duplication rates to be 0.0014 gene^{-1} myr^{-1} for fruit fly, 0.028 gene^{-1} myr^{-1} for yeast and 0.024 gene^{-1} myr^{-1} for roundworm.

In the same year Gu *et al.* (2002a) discussed the importance of small-scale (tandem or segmental) duplications versus large-scale duplications using vertebrate gene family data. This dataset included about one-quarter of the human gene families. Based on this vertebrate dataset two time windows, 750-900 mya and 80-430 mya, were determined in which only small-scale duplications occurred. For the first time window an average duplication rate of 0.79 myr^{-1} across the whole genome was estimated, while for the second time interval a duplication rate of 1.25 myr^{-1} was found. Since these rates are specific for the entire dataset not for single genes and the exact number of genes in their dataset is not given, it is not possible to compute a rate per gene per myr. Thus, these rates cannot be compared to previous estimates.

In 2003 Lynch & Conery published an update of their estimates, since nearly complete genomic sequences had emerged for several species. Based on the analyzes of seven species, they confirmed their average duplication rate for eukaryotic genes to be about 0.01 gene^{-1}

myr^{-1}. However, the duplication rates for the particular species changed slightly. A duplication rate of 0.009 gene^{-1} myr^{-1} was estimated for human, 0.001 gene^{-1} myr^{-1} for fruit fly and 0.004 gene^{-1} myr^{-1} for two yeast species.

The Rat Genome Sequencing Project Consortium (2004) conducted a genome-wide analysis for rat in comparison to the mouse and the human genome. They found a duplication rate between 0.0013 and 0.0026 gene^{-1} myr^{-1} for rodents, represented by mouse and rat.

In 2005 Cotton & Page reanalyzed the data from Gu *et al.* (2002a). They were the first who explicitly estimated the rate of gene loss in vertebrates. Using the entire data of Gu *et al.*, which include duplications estimated to date from ∼4700 mya to present day, Cotton & Page inferred a duplication rate of 0.00097 gene^{-1} myr^{-1} and a deletion rate of 0.00048 gene^{-1} myr^{-1} with a 95% confidence interval of 0.00089-0.00105 and 0.000153-0.000786, respectively. An assumption of the model they used for this analysis was that the rates are constant over time. They showed that this assumption was violated for the entire dataset, but could be accepted for smaller dataset which was restricted to duplications from the last 200 myr only. For this dataset Cotton & Page estimated a duplication rate of 0.00115 gene^{-1} myr^{-1} as well as a deletion rate of 0.0074 gene^{-1} myr^{-1} with a 95% confidence interval of 0.000902-0.00131 and 0.00409-0.00951, respectively. Their estimated duplication rate is an order of magnitude lower than the previous estimates for human genes (Lynch & Conery (2001), Lynch & Conery (2003)). For comparisons of their estimated deletion rate, Cotton & Page use the half-life time of human (7.5 myr) estimated by Lynch & Conery (2003) to calculate the corresponding deletion rate of 0.0924 gene^{-1} myr^{-1} for humans. Again their estimated deletion rate is even more than an order lower than the previous estimate of Lynch & Conery (2003).

Studies based on explicitly modeled gene duplication and deletions. In recent years stochastic models, where duplications and deletions are explicitly modeled, were applied. These models allows a more direct estimation of duplication and deletion rates (Roth *et al.* (2006)). The estimates from Hahn *et al.* (2005); Csűrös & Miklós (2006); Borenstein *et al.* (2006) are based on different stochastic models, due to different requirements. In all studies only the gene family size in different genomes, not the corresponding gene trees, served as data. It emerges that the estimation of multiple parameters in complex models is really a challenge. The results from these studies are given below.

Hahn *et al.* (2005) applied a stochastic birth and death model assuming equal duplication and deletion rates to describe the evolution of gene families under a given phylogeny. This birth and death model is a special case of our model, which will be described in the next chapter. For a dataset from five yeast species, they estimated the duplication-deletion rate as 0.002 gene^{-1} myr^{-1}. The method was also used to identify branches in the species tree and genes which evolved nonrandomly. One year later the same group published a study based on mammalian data (Demuth *et al.* (2006)). They adopted their methodology for this dataset and estimated the duplication-deletion rate to be 0.0016 gene^{-1} myr^{-1}. Furthermore the expected number of gene gain and loss for every branch in the tree was calculated.

All estimated rates of the previous studies are global rates. Either one duplication rate or one duplication rate and one deletion rate equal for the entire dataset was estimated. That means, that the rates are constant over time and equal for all considered genes and all considered species in the dataset. Table 1.2 summarizes the results from these studies.

Last year Csűrös & Miklós (2006) published a study on the estimation of gene duplication and gene deletion rates as well as additional the estimation of a rate for horizontal gene transfer. They applied the method to proteobacteria and used the gene families from the COG (Cluster of Orthologous Groups) database. The dataset was divided into 9 groups and for each group specific values for the three rates were inferred. For one group containing 19% of the data all rates were estimated to be nearly zero. Thus, there were no significant changes in the gene copy number. Further two groups containing 11% of the data stand out due to large horizontal transfer rates. In total, the deletion rates was for all groups greater than zero, whereas only for 4 of the 9 groups a duplication rates greater than zero could be estimated. For the remaining groups the duplication rate was always smaller than the deletion rate.

In the same year Borenstein *et al.* (2006) employed a single-parameter model to estimate gene-specific loss rates and applied it to data from 16 eukaryotic species. This is the first method where deletion rates can be estimated for every single gene family. So it is possible to look at the variability in evolutionary dynamics between different genes. They restricted their analyzes to eukaryotic species, in which horizontal gene transfer is unlikely, and on genes whose genomic copy number is one. They suggested that their method allows for the estimation of an optimal deletion rate for each gene family. Borenstein *et al.* gave no explicit deletion rates, but the number of gene deletions in each branch of the species tree of the analyzed species. A massive loss was e.g. found between chimp (*Pan troglodytes*) and its common ancestor with human (599 genes), whereas on the branch to human only 37 genes were lost.

This overview of already published rates is a good example for the dependency of the estimates on used method and chosen data for the analysis. Since it is impossible to find the 'true' rates for gene duplication and deletion, we have to check the robustness of the methods and the quality of the data carefully to minimize the sources of error. Estimates for the rates have to be handled with care and should be seen as a relative measure, not as fixed rates and can be compared to other estimates to that effect. For this reason the method presented in this thesis was tested in various simulation studies and on real data, which was partly already used in previous studies (data from Demuth *et al.* (2006)).

1.4 Conclusion

The mechanisms of gene duplication and deletion are manifold and so are their influences on the genome. So the question arose what kinds of gene duplication and deletion should be included in a mathematical model describing these processes. Simultaneously a mathematical model had to be found which is easy enough to use and allow for the estimation of model parameters in reasonable time.
It is obvious, that a model including all possible mechanisms for gene duplication and deletion is not practicable. Since many duplications are caused by unequal crossing-over which leads to tandemly duplicated regions, a model for single gene duplications and single gene deletions seemed reasonable. The simplest imaginable model would require the following assumptions: independence of genes, duplication or deletion of no more than one gene at the same time, a duplication rate and a deletion rate for each gene, which is constant over time. A model which fulfills these assumptions is the birth and death model described in the following chapter.
As we have seen, the knowledge of a gene family include besides the number of gene family members information about the corresponding gene tree, which can be reconstructed on the basis of the DNA sequences of the family members. But to keep the complexity of the model manageable, we only consider the information about the numbers of gene family members. Based on these numbers and the given species tree the rates for gene duplication and deletion will be estimated specific for each gene family.

CHAPTER TWO

Estimation of gene duplication and deletion rates

So far there were few probabilistic models describing the change of a gene family in size, but in the last few years some new methods came up try to model gene family evolution. Some of these methods were presented in the previous chapter (Hahn *et al.* (2005); Csűrös & Miklós (2006); Borenstein *et al.* (2006)). In this context, the stochastic birth and death model (BD model) (Feller (1950); Bailey (1964); Lahres (1964)) is very common. Hence, in Novozhilov *et al.* (2006) an overview about the biological applications in the theory of BD processes can be found. The BD process was first used in population evolution models but appealed nowadays also in the modeling of gene specific processes. A birth can be regarded as a duplication and the death as a deletion of a gene. Each of both events occurs with a certain rate which can be estimated.

2.1 Birth and death (BD) model

Our approach is also based on a stochastic birth and death model (Feller (1950); Bailey (1964); Lahres (1964)). It only allows for single gene duplications and single gene deletions. Large-scale duplications, like chromosomal or whole genome duplications, cannot be explained with this model. Furthermore all genes are assumed to be independent of each other. The duplication rate and deletion rate is equal for all considered species and over time, but can differ between gene families. To our knowledge, there is currently no method available where both rates, duplication rate and deletion rate, can specifically be estimated for single gene families.

2.1.1 Definition

Consider a family of genes whose total number at time t is given by the discrete random variable $X(t)$. The probability of $X(t) = i$ is denoted by $p_i(k,t) = P(X(t) = i|X(0) = k)$ whereas k is the number of gene copies at time $t = 0$. Every gene has the ability to give birth to a new gene while staying alive itself. This is considered a gene duplication and occurs at a specific birth rate λ_i. On the other hand, every gene can get lost, which corresponds to the deletion of a gene and occurs at a specific loss rate μ_i. A duplication of a gene leads to an increase by one in the gene family size, whereas a deletion leads to a decrease by one (figure 2.1).

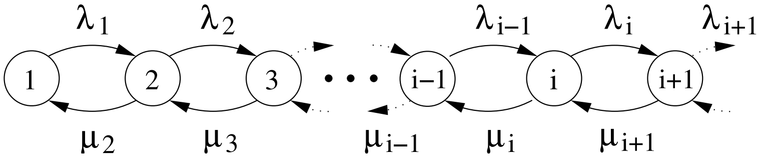


Figure 2.1: Schematic illustration of the transitions in a birth and death process.

In a small time interval Δt no more than one event is allowed. So the possible transitions for a gene family with size $X(t) = i$ can be classified as follows:

(1) transition from state i to state $i+1$ with probability $\lambda_i\,\Delta t + o(\Delta t)$

(2) transition from state i to state $i-1$ with probability $\mu_i\,\Delta t + o(\Delta t)$

(3) no transition with probability $1 - (\lambda_i\,\Delta t + \mu_i\,\Delta t) + o(\Delta t)$

(4) two or more transitions with probability $o(\Delta t)$

Because these transitions are independent of each other and mutually exclusive, their probabilities $p_i(t + \Delta t)$ add up to

$$p_i(t+\Delta t) = \underbrace{p_i(t)\{1-(\lambda_i\,\Delta t + \mu_i\,\Delta t)\}}_{\text{no transition}} + \underbrace{\lambda_{i-1}\,\Delta t\,p_{i-1}(t)}_{\text{transition } i-1\rightarrow i} + \underbrace{\mu_{i+1}\,\Delta t\,p_{i+1}(t)}_{\text{transition } i+1\rightarrow i} + \underbrace{0}_{\substack{\text{more}\\ \text{trans.}}} \quad (2.1)$$

By subtracting $p_i(t)$ from $p_i(t + \Delta t)$ and dividing the equation by Δt, the difference ratio of $p_i(t)$ is constructed. The limit for $\Delta t \to 0$ leads to the first derivative for t:

$$\begin{aligned} p_i'(t) &= \lim_{\Delta t\to 0} \tfrac{p_i(t+\Delta t)-p_i(t)}{\Delta t} \\ &= -(\lambda_i + \mu_i)\,p_i(t) + \lambda_{i-1}\,p_{i-1}(t) + \mu_{i+1}\,p_{i+1}(t) \quad \text{for } i \geq 1 \qquad (2.2) \\ &= -\lambda_0\,p_0(t) + \mu_1\,p_1(t) \quad \text{for } i = 0 \end{aligned}$$

Since $i = 0$ is an absorbing state no loss rate μ_0 exists. Assuming a linear BD process with no interactions among genes, the rates λ and μ become constant characteristics of the process and depend only on the present number of gene copies:

$$\lambda_i = i\,\lambda,\ \mu_i = i\,\mu \qquad (2.3)$$

Then the basic system of differential equations has the following form:

$$\begin{aligned} p_i'(t) &= -(\lambda+\mu)\, i\, p_i(t) + \lambda(i-1)\, p_{i-1}(t) + \mu(i+1)\, p_{i+1}(t) \\ p_0'(t) &= \mu\, p_1(t) \\ p_k(0) &= 1 \end{aligned} \tag{2.4}$$

where $p_k(0) = 1$ defines the initial state, which means that at $t = 0$ exactly k gene copies exist. The solution of this system results in the probability function $p_i(k,t)$. A detailed solution for the most general case $\lambda \neq \mu$ and $k > 1$ can be found in appendix B. All other cases can be derived from that general case.

2.1.2 Probability function of the BD process

The computation of the probability function depends on the parameters λ, μ, and k. Therefore we distinguish between several cases. See also Feller (1950); Bailey (1964); Lahres (1964).

Case 1: $\lambda \neq \mu$

To simplify matters we introduce two variables A and B which are defined as follows

$$A = \frac{\mu(e^{(\lambda-\mu)t}-1)}{\lambda e^{(\lambda-\mu)t}-\mu}, \quad B = \frac{\lambda(e^{(\lambda-\mu)t}-1)}{\lambda e^{(\lambda-\mu)t}-\mu} \tag{2.5}$$

For the case $k = 1$ the probability distribution is found to be

$$\begin{aligned} &p_0(1,t) = A \\ &p_i(1,t) = (1-A)(1-B)B^{i-1} \quad \forall i \geq 1 \end{aligned} \tag{2.6}$$

It is given by terms of a geometric series, except for the first term where $i = 0$. For the case $k > 1$ the probability distributions is expressed as

$$\begin{aligned} &p_0(k,t) = A^k \quad \forall k > 1 \\ &p_i(k,t) = \sum_{j=0}^{\min(k,i)} \tbinom{k}{j}\tbinom{k+i-j-1}{k-1}\, A^{k-j}\, B^{i-j}\, (1-A-B)^j \quad \forall i \geq 1, \forall k > 1 \end{aligned} \tag{2.7}$$

Case 2: $\lambda = \mu$

For equal birth and loss rates we introduce a variable C defined as

$$C = \frac{\lambda t}{\lambda t + 1} \tag{2.8}$$

Thus, the probability distribution for $k = 1$ becomes

$$\begin{aligned} &p_0(1,t) = C \\ &p_i(1,t) = \tfrac{(\lambda t)^{i-1}}{(\lambda t+1)^{i+1}} \quad \forall i \geq 1 \end{aligned} \tag{2.9}$$

For $k > 1$ the probability is computed using the following equations

$$\begin{aligned} p_0(k,t) &= C^k \quad \forall k > 1 \\ p_i(k,t) &= \sum_{j=0}^{\min(k,i)} \binom{k}{j}\binom{k+i-j-1}{k-1} C^{k+i-j-1}\,(1-2\,C)^j \quad \forall i \geq 1, \forall k > 1 \end{aligned} \tag{2.10}$$

This distinction of cases builds the foundation for the valid and efficient computation of the probability function.

2.1.3 Properties of the probability distribution

Two main properties of the probability function $p_i(k,t)$ are the mean and the variance. They can easily be obtained from the probability generating function (see appendix eq. B.10 or Bailey (1964)) or directly calculated from the basic system of differential equations eq. 2.4 (see Feller (1950)). For case 1 ($\lambda \neq \mu$) the mean $m(t)$ and the variance $\sigma^2(t)$ are defined by

$$\begin{aligned} m(t) &= ke^{(\lambda-\mu)t} \\ \sigma^2(t) &= \tfrac{k(\lambda+\mu)}{(\lambda-\mu)} e^{(\lambda-\mu)t}(e^{(\lambda-\mu)t}-1) \end{aligned} \tag{2.11}$$

Applying l'Hôpital's rule on eq. 2.11 with $\mu \to \lambda$, the mean and the variance for case 2 ($\lambda = \mu$) are found to be

$$\begin{aligned} m(t) &= k \\ \sigma^2(t) &= 2k\lambda t \end{aligned} \tag{2.12}$$

As we will show later, mean and variance can be used to estimate the parameters of the BD probability function and to evaluate the quality of simulated gene family data.

2.1.4 BD model for a phylogenetic tree

Since we want to use a phylogenetic tree to estimate the duplication and deletion rates, we introduce a specific notation for species trees to provide the probability function $p_i(k,t)$ of the BD model with the required information. This includes the current number of gene copies i, that emerged from the number of gene copies k after a specific time t, assuming the duplication and deletion rates λ and μ respectively.

We divide the set of nodes $\mathcal{N}$ of the tree into three subsets: leaves $\mathcal{N}^L$, internal nodes $\mathcal{N}^I$, and the root node n^R. Every leaf $n \in \mathcal{N}^L$ corresponds to a recent species and every internal node $n \in \mathcal{N}^I$ to an ancestor of several species. Each node $n \in \mathcal{N} \setminus \{n^R\}$ holds a certain number of gene copies i_n, depending on the number of gene copies of its parent k_n and the duplication rate and deletion rate, which correspond to the number of birth and death events on the branch between them in a certain time t_n (see figure 2.2). Since

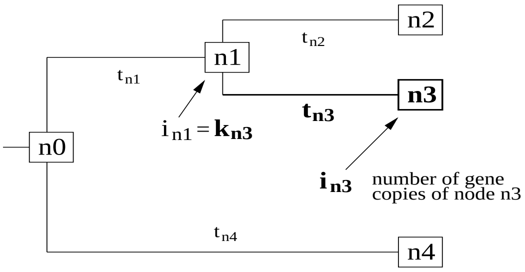


Figure 2.2: Required parameters for the BD model on a species trees. Set of nodes $\mathcal{N} = \{n0, n1, n2, n3, n4\}$ can be divided into root node $n^R = n0$, leaves $\mathcal{N}^L = \{n2, n3, n4\}$ and internal nodes $\mathcal{N}^I = \{n1\}$. To calculate the probability $p_{i_{n3}}(k_{n3}, t_{n3})$ for the node $n3$ the following parameters are necessary: the number of gene copies i_{n3}, the number of gene copies of the ancestor $k_{n3} = i_{n1}$ and the time between the ancestor and the node t_{n3}.

the root node n^R has no parent, it has an exceptional state within the species tree. The number of gene copies of the root node shall in the following be denoted as α.

If the duplication rate λ, the deletion rate μ, and α are known, the function of the BD model can be used to calculate the probability $p_{i_n}(k_n, t_n)$ for observing exactly i_n gene copies after time t_n, if the ancestor had k_n gene copies. In this case we can compute the most probable number of gene copies i_n for each leaf $n \in \mathcal{N}^L$.

In practice, the number of gene copies for the leaves can be obtained from gene family data, whereas λ, μ, and α are unknown. Then the function $p_{i_n}(k_n, t_n)$ can be used to infer values for λ, μ, and α which most probably led to the observed gene copy numbers.

In the next section we focus on strategies for the estimation of the BD model parameters λ, μ, and α.

2.2 Estimation of model parameters

A fundamental aim in statistics is to get information about the distribution of a set of observations. Generally for this purpose the characteristics of the distribution, called parameters, are determined. The estimation of a single parameter or sets of parameters from a set of observations is well established and extensively used in bioinformatics. In this section we give an overview over widely used estimation strategies based on Fahrmeir *et al.* (2004) and Ewens & Grant (2001) and discuss possible applications for our approach.

In the entire section we assume that X is a discrete random variable with the probability mass function $P(X = x \mid \theta)$ where θ is the unknown parameter. All of the following declarations also apply for continuous random variables. A single observed value x will usually not be sufficient to give a good estimate for θ, so a set of values resulting from multiply observations is required. These values can be assigned to i independent identically distributed (iid) random variables $X_1, X_2, \dots, X_i$, each with the probability mass function $P(X_j = x_j \mid \theta)$ identical to $P(X = x \mid \theta)$. Then an *estimator* of the parameter θ is a function of $X_1, X_2, \dots, X_i$ denoted as $\widehat{\theta}\,(X_1, X_2, \dots, X_i)$ or simply $\widehat{\theta}$. The value calculated from specific observed values $\widehat{\theta}\,(x_1, x_2, \dots, x_i)$ is called an *estimate* of θ.

In general we can choose between several estimators for the parameter θ. Consequently it is essential to have some criteria which can be used to evaluate the quality of an estimators, to decide which one to choose.

2.2.1 Quality of an estimator

Since the true value of the parameter θ is unknown we are not able to make statements about the accuracy of the estimate $\widehat{\theta}\,(x_1, x_2, \ldots, x_i)$. Thus, we have to make sure that the probability that $\widehat{\theta}\,(x_1, x_2, \ldots, x_i)$ differs from the true value of θ is as small as possible for the chosen estimator. Therefore, we analyze the distribution of the estimator and check its properties. There are several criteria to determine if a certain estimator is preferable to another.

One desirable property is that the estimator should be *unbiased*, which means that it should tend to result in the true value of θ and not to overestimate or underestimate it. This can be measured in terms of expectation or mean which is defined as:

$$\mathbb{E}(X) = \sum_x xP(X = x) \tag{2.13}$$

Then an estimator is unbiased if the following equation holds

$$\mathbb{E}(\,\widehat{\theta}\,(X_1, X_2, \ldots, X_i)\,) = \theta \tag{2.14}$$

Another desirable property is the *consistency* of an estimator. This property reflects the behavior of the estimator if the sample size is large. While the unbiasedness is only determined by the mean value, the consistency also takes the variance of the estimated parameters into account. The consistency can be measured using the mean squared error (MSE) which is defined as:

$$MSE = \mathbb{E}(\,(\widehat{\theta}\,(X_1, X_2, \ldots, X_i) - \theta)^2\,) \tag{2.15}$$

If the MSE goes to zero for increasing number of observations the estimator is called consistent. That means, if the sample size grows, the estimate is more and more likely to be the true value of θ. If several estimators are available, the best one in terms of unbiasedness and consistency should be used. An estimator is of good quality if it is unbiased, consistent, and has a small variance.

In the following two commonly used principles of constructing estimators for model parameters are described.

2.2.2 Method of moments

The method of moments (MOM) (Pearson (1894)) is a common way to estimate unknown parameters of a distribution and can be used for discrete and continuous probability distributions. Moments are characteristics of a distribution, describing its shape and scale. The MOM is based on the computation of theoretical moments of the assumed distribution as

well as the computation of the moments for the observed data. There are as many moments required as there are unknown parameters. The basic idea is to find values for the parameters such that the theoretical moments of the distribution are as close as possible to the one computed from the observed data.
For a random variable X with the probability mass function $P(X = x)$ and with z as a positive integer, the zth *moment* m_z of X is defined as (Grimmett & Stirzaker (2001)):

$$m_z = \mathbb{E}(X^z) \tag{2.16}$$

Then the zth *central moment* σ_z is defined as:

$$\sigma_z = \mathbb{E}((X - m_1)^z) \tag{2.17}$$

Two widely used moments are the first moment $m_1 = \mathbb{E}(X)$ for $z = 1$ which is equivalent to the mean and the second central moment $\sigma_2 = \mathbb{E}((X - \mathbb{E}X)^2)$ for $z = 2$ which is equivalent to the variance of X. The square root of σ_2 is known as standard deviation σ.
Suppose that $X_1, X_2, \ldots, X_i$ are iid discrete random variables each with the probability mass function $P(X_j = x_j \mid \theta)$ identical to $P(X = x \mid \theta)$ depending on the single unknown parameter θ. The first moment m_1 of X_j is then assumed to be a function g depending on θ, i.e.

$$m_1 = g(\theta) \tag{2.18}$$

The empirical first moment will be denoted with $\widehat{m}_1 = \widehat{m}_1(x_1, x_2, \ldots, x_i)$ and is computed as the average of the concrete sample $x_1, x_2, \ldots, x_i$. Then the estimator $\widehat{\theta}$ is given by equating the empirical first moment $\widehat{m}_1$ with the theoretical first moment m_1 which is the function g for $\widehat{\theta}$

$$\widehat{m}_1 = \frac{1}{i} \sum_j x_j == g(\widehat{\theta}) = m_1 \tag{2.19}$$

Solving this equation for $\widehat{\theta}$ depending on $\widehat{m}_1$, if possible, and inserting the concrete value for $\widehat{m}_1$, results in the estimate for θ.
For more than one parameter $\theta_1, \theta_2, \ldots, \theta_n$ the first moment m_1 and the second to nth central moment σ_z ($\forall\, 2 \leq z \leq n$) are used and assumed to be functions of the parameters $\theta_1, \theta_2, \ldots, \theta_n$. This leads to the following system of equations:

$$\begin{array}{lllllllll}
\widehat{m}_1 & = & \frac{1}{i}\sum_j x_j & == & g_1(\widehat{\theta}_1, \ldots, \widehat{\theta}_n) & = & \mathbb{E}(X) & = & m_1 \\
\widehat{\sigma}_2 & = & \frac{1}{i}\sum_j (x_j - \widehat{m}_1)^2 & == & g_2(\widehat{\theta}_1, \ldots, \widehat{\theta}_n) & = & \mathbb{E}((X - m_1)^2) & = & \sigma_2 \\
\ldots & & & & & & & & \\
\widehat{\sigma}_n & = & \frac{1}{i}\sum_j (x_j - \widehat{m}_1)^n & == & g_n(\widehat{\theta}_1, \ldots, \widehat{\theta}_n) & = & \mathbb{E}((X - m_1)^n) & = & \sigma_n
\end{array} \tag{2.20}$$

To find explicit solutions for $\widehat{\theta}_1, \ldots, \widehat{\theta}_n$ can be very difficult. If it is not possible, numerical root finding methods can be applied (subsection 2.2.4). In general, the MOM provides a simple way to estimate model parameters. Unfortunately, it is often not applicable and not always unbiased. In some cases the estimates given by the MOM are outside the parameter space, which happens more frequently with smaller samples than with larger ones. In that case the estimated parameters are useless.

MOM for the BD process

The application of the MOM on our BD model using a phylogenetic tree is in no way straightforward. Still, to apply the MOM for this purpose, we assume all leaves to be independent of each other and ignore the tree topology. In doing so, the gene copy numbers at the leaves become our sample and their distribution corresponds to the distribution described by the function $p_i(k,t)$ of our BD model. Then, t is the total time over the tree from the MRCA to the leaves. k is the number of gene copies of the root of the tree, which is α.
Assuming a fixed number of gene copies α at the root, we have to find MOM estimators for the duplication rate λ and the deletion rate μ. Therefore, we need the first moment m_1 and the second central moment σ_2 of the probability distribution $p_i(\alpha,t)$ (see eq.2.11):

$$\begin{array}{lclcl} m_1 & = & \mathbb{E}(X) & = & \alpha e^{(\lambda-\mu)t} \\ \sigma_2 & = & \mathbb{E}((X-m_1)^2) & = & \frac{\alpha\,(\lambda+\mu)}{(\lambda-\mu)} e^{(\lambda-\mu)t}(e^{(\lambda-\mu)t}-1) \end{array} \tag{2.21}$$

with X = number of gene copies. These theoretical formulas for the moments are equated with the empirical mean $\widehat{m_1}$ and variance $\widehat{\sigma_2}$ of the observed data:

$$\begin{array}{lcl} \widehat{m}_1 & == & \alpha e^{(\widehat{\lambda}-\widehat{\mu})t} \\ \widehat{\sigma}_2 & == & \frac{\alpha\,(\widehat{\lambda}+\widehat{\mu})}{(\widehat{\lambda}-\widehat{\mu})} e^{(\widehat{\lambda}-\widehat{\mu})t}(e^{(\widehat{\lambda}-\widehat{\mu})t}-1) \end{array} \tag{2.22}$$

This equation system can be solved and after some rearrangements we get the two estimators $\widehat{\lambda}(\alpha,t,\widehat{m}_1,\widehat{\sigma}_2)$ and $\widehat{\mu}(\alpha,t,\widehat{m}_1,\widehat{\sigma}_2)$ depending on α, the time t, and the two empirical moments $\widehat{m}_1$ and $\widehat{\sigma}_2$:

$$\begin{array}{lcl} \widehat{\lambda}(\alpha,t,\widehat{m}_1,\widehat{\sigma}_2) & = & -\frac{(\widehat{m}_1^2+\alpha\,(\widehat{\sigma}_2-\widehat{m}_1))\,ln\,(\frac{\widehat{m}_1}{\alpha})}{2\,(\alpha-\widehat{m}_1)\widehat{m}_1 t} \\ \widehat{\mu}(\alpha,t,\widehat{m}_1,\widehat{\sigma}_2) & = & \frac{(\widehat{m}_1^2-\alpha\,(\widehat{\sigma}_2+\widehat{m}_1))ln\,(\frac{\widehat{m}_1}{\alpha})}{2\,(\alpha-\widehat{m}_1)\widehat{m}_1 t} \end{array} \tag{2.23}$$

The estimates for λ and μ are calculated with the given estimators using the mean $\widehat{m}_1$ and the variance $\widehat{\sigma}_2$ from the observed data, i.e. the numbers of gene copies of the leaves. If we also want to estimate the ancestral number of gene copies α, we have to take the third central moment σ_3 into account

$$\sigma_3 = \frac{\alpha e^{(\lambda-\mu)t}}{(\lambda-\mu)^2}(\lambda^2+4\lambda\mu+\mu^2+e^{(\lambda-\mu)t}(2e^{(\lambda-\mu)t}(\lambda^2+\lambda\mu+\mu^2)-3(\lambda+\mu)^2)) \tag{2.24}$$

Solving the equation system composed of eq. 2.21 and eq. 2.24 and using $\varphi(\widehat{m}_1, \widehat{\sigma}_2, \widehat{\sigma}_3)$ with

$$\varphi(\widehat{m}_1, \widehat{\sigma}_2, \widehat{\sigma}_3) = \sqrt{\widehat{m}_1^2 - 3\widehat{\sigma}_2^2 + 2\widehat{m}_1\widehat{\sigma}_3} \tag{2.25}$$

as an auxiliary function we get the following estimators for α, λ, and μ:

$$\begin{aligned}
\widehat{\alpha}(\widehat{m}_1, \widehat{\sigma}_2, \widehat{\sigma}_3) &= \frac{\widehat{m}_1^2}{\varphi\,(\widehat{m}_1,\widehat{\sigma}_2,\widehat{\sigma}_3)} \\
\widehat{\lambda}(t, \widehat{m}_1, \widehat{\sigma}_2, \widehat{\sigma}_3) &= \frac{(-\widehat{m}_1+\widehat{\sigma}_2+\varphi\,(\widehat{m}_1,\widehat{\sigma}_2,\widehat{\sigma}_3))\; ln\,(\frac{\varphi\,(\widehat{m}_1,\widehat{\sigma}_2,\widehat{\sigma}_3)}{\widehat{m}_1})}{2\,(-\widehat{m}_1+\varphi\,(\widehat{m}_1,\widehat{\sigma}_2,\widehat{\sigma}_3))\; t} \\
\widehat{\mu}(t, \widehat{m}_1, \widehat{\sigma}_2, \widehat{\sigma}_3) &= \frac{(\widehat{m}_1+\widehat{\sigma}_2-\varphi\,(\widehat{m}_1,\widehat{\sigma}_2,\widehat{\sigma}_3))\; ln\,(\frac{\varphi\,(\widehat{m}_1,\widehat{\sigma}_2,\widehat{\sigma}_3)}{\widehat{m}_1})}{2\,(-\widehat{m}_1+\varphi\,(\widehat{m}_1,\widehat{\sigma}_2,\widehat{\sigma}_3))\; t}
\end{aligned} \tag{2.26}$$

The estimator for α is totally independent of the time t and the two estimator for λ and μ can be seen as rates divided by the time t. The empirical third central moment $\widehat{\sigma}_3$ can also be calculated from the observed data.
As we discuss later, these estimators are evaluated in several simulation studies. The results of these studies and the quality of the MOM for this application is discussed in subsection 4.3.2.

2.2.3 Maximum likelihood

Another estimation principle is maximum likelihood (ML). In many practical cases ML estimates seem to have a higher probability of being close to the true value than other methods, especially if the sample size increases. ML is widely used in statistical and bioinformatical fields and was first described by Fisher (1922). A historical overview and an extensive introduction can be found in Edwards (1992).
Suppose that $X_1, X_2, \ldots, X_i$ denoted with $\mathbf{X}$ are iid discrete random variables each with the probability mass function $P(X_j\!=\!x_j \,|\, \theta)$ identical to $P(X\!=\!x \,|\, \theta)$ depending on the single unknown parameter θ. The joint probability mass function for these random variables is then

$$P(X_1\!=\!x_1 \,|\, \theta) \times P(X_2\!=\!x_2 \,|\, \theta) \times \cdots \times P(X_i\!=\!x_i \,|\, \theta) \tag{2.27}$$

with $x_1, x_2, \ldots, x_i$ being arbitrary values for $X_1, X_2, \ldots, X_i$. Instead of interpreting the joint probability function at specific realizations $x_1, x_2, \ldots, x_i$, it is possible to analyze the function for given values $x_1, x_2, \ldots, x_i$ as a function of θ. This is called the *likelihood function* $L(\theta, \mathbf{X})$ which is defined as:

$$\begin{aligned}
L(\theta, \mathbf{X}) &= P(X_1\!=\!x_1 \,|\, \theta) \times P(X_2\!=\!x_2 \,|\, \theta) \times \cdots \times P(X_i\!=\!x_i \,|\, \theta) \\
&= \prod_{j=1}^{i} P(X_j\!=\!x_j \,|\, \theta)
\end{aligned} \tag{2.28}$$

The maximum likelihood theory is based on the maximization of this likelihood function. Therefore the parameter $\widehat{\theta} = \widehat{\theta}\,(X_1, X_2, \ldots, X_i)$ is chosen in a way so that the likelihood function $L(\theta, \mathbf{X})$ is maximized, i.e.

$$L(\widehat{\theta}, \mathbf{X}) = \underset{\theta}{max}\, L(\theta, \mathbf{X}) = \underset{\theta}{max} \prod_{j=1}^{i} P(X = x_j \mid \theta) \tag{2.29}$$

The value $\widehat{\theta}$ is then called *maximum likelihood estimator* (MLE). The estimator can be found by calculating the first derivative of $L(\widehat{\theta}, \mathbf{X})$ for θ and setting it to zero:

$$\frac{d}{d\theta}\, L(\theta, \mathbf{X}) = 0 \tag{2.30}$$

Since this procedure can lead to difficult terms in practice, it is more convenient to use the logarithm of the likelihood function $log\, L(\theta, \mathbf{X})$. This is called *log-likelihood function.* Because the logarithmic calculus is a monotonically increasing transformation, the maximization of $L(\theta, \mathbf{X})$ and $log\, L(\theta, \mathbf{X})$ leads to the same value for $\widehat{\theta}$. For the maximization of $log\, L(\theta, \mathbf{X})$ the following equation has to be solved

$$\frac{d}{d\theta}\, log\, L(\theta, \mathbf{X}) = \sum_{j=1}^{i} log\, P(X = x_j \mid \theta) = 0 \tag{2.31}$$

The logarithm of $L(\theta, \mathbf{X})$ is a sum rather than a product and the differentiation procedure is then almost always easier. By substituting $\mathbf{X} = X_1, X_2, \ldots, X_i$ with $\mathbf{x} = x_1, x_2, \ldots, x_i$ we get $L(\theta, \mathbf{x})$ and consequently the *maximum likelihood estimate* of θ for the observed values $x_1, x_2, \ldots, x_i$.

Often there are multiple parameter to estimate. In this case a general solution is given by replacing θ with the vector $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. The differentiation of eq. 2.31 becomes more complex and leads to a n-dimensional maximization problem resulting into a system of partial differential equations:

$$\frac{\partial\, log\, L(\theta_1, \ldots, \theta_n, \mathbf{X})}{\partial \theta_1} = 0\,,\ \ldots,\ \frac{\partial\, log\, L(\theta_1, \ldots, \theta_n, \mathbf{X})}{\partial \theta_n} = 0 \tag{2.32}$$

In practice, it is not always possible to solve this equation system. In that case optimization strategies have to be used to find the best estimate for $\widehat{\theta}$ for a given set of values $x_1, x_2, \ldots, x_i$ (see subsection 2.2.4).

Likelihood function for a phylogenetic tree under the BD process

Felsenstein (1981) introduced ML estimation in the field of phylogenetic tree reconstruction. In this context the parameters to be estimated are the branch lengths of the tree. An overview about phylogenetic tree reconstruction using ML can be found in Felsenstein (1981), Goldman (1990), and Swofford *et al.* (1996).

We define the likelihood function of a BD process on a tree different to the previous approach, since our aim is to estimate duplication and deletion rates for gene families and not the branch lengths of a tree. The number of gene copies for the considered recent species are known for different gene families. We assume that the number of gene copies of a gene family evolved along a phylogenetic tree under the influence of a BD process. This tree and its branch length are known.

The probability to find a certain number of gene copies at a node n of a tree $\mathcal{T}$ is calculated using the probability function of the BD process $p_{i_n}(k_n, t_n) = p_i(k, t)$ (see section 2.1.2), whereas i_n is the number of gene copies of the considered node n, k_n is the gene copy number of its ancestor, and t_n the branch length between node n and its ancestor. Then the likelihood function for a tree $L = L_{\mathcal{T}}$ is defined as the product of the probabilities $p_{i_n}(k_n, t_n)$ of all nodes $n \in \mathcal{N}$ of the tree T.

$$L_{\mathcal{T}} = \prod_{\forall n \in \mathcal{N}} p_{i_n}(k_n, t_n) \tag{2.33}$$

The likelihood function $L_{\mathcal{T}} = L_{\mathcal{T}}(\alpha, \lambda, \mu)$ depends on the tree $\mathcal{T}$ with branch lengths, the ancestral gene number α, the duplication rate λ, and the deletion rate μ. As defined in 2.1.4, the nodes $n \in \mathcal{N}$ of the tree can be divided into leaf nodes $n \in \mathcal{N}^L$ and internal nodes $n \in \mathcal{N}^I$. It is only possible to observe the number of gene copies at the leaves of the tree, because these nodes represent recent species, while all inner nodes represent unknown ancestors. Therefore, we have to sum over all possible numbers of gene copies at the inner nodes, excluding the number zero, since this represents an absorbing state:

$$L_{\mathcal{T}} = L_{\mathcal{T}}(\alpha, \lambda, \mu) = \sum_{i_{n_1^I}=1}^{\infty} \sum_{i_{n_2^I}=1}^{\infty} \cdots \sum_{i_{n_m^I}=1}^{\infty} \prod_{\forall n \in \mathcal{N}} p_{i_n}(k_n, t_n) \tag{2.34}$$

with $\mathcal{N}^I = \{n_1^I, n_2^I, \ldots, n_m^I\}$ representing the inner nodes of the tree $\mathcal{T}$. The probability of the root node results in $p_\alpha = 1$, assuming a specific number of gene copies α for the MRCA. This number α as a parameter of the likelihood function $L_{\mathcal{T}}(\alpha, \lambda, \mu)$ will also be inferred in the maximum likelihood estimation. The term for $L_{\mathcal{T}}(\alpha, \lambda, \mu)$ from eq. 2.34 can efficiently be converted by continuously factoring out probabilities, which are independent of the considered summation, that leads to a nesting of sums and products. This procedure is similar to the *Horner scheme* or Horner algorithm from William Horner (1819) developed for the evaluation of polynomials. Equation 2.35 shows the converted likelihood function for the tree $\mathcal{T}$ = tMRHF from figure 1.3.

$$\begin{aligned} L_{tMRHF}(\alpha, \lambda, \mu) &= p_{i_F}(\alpha, t_F) \sum_{i_{MRH}=1}^{\infty} p_{i_{MRH}}(\alpha, t_{MRH})\, p_{i_H}(i_{MRH}, t_H) \\ &\quad \cdot \sum_{i_{MR}=1}^{\infty} p_{i_{MR}}(i_{MRH}, t_{MR})\, p_{i_M}(i_{MR}, t_M)\, p_{i_R}(i_{MR}, t_R) \end{aligned} \tag{2.35}$$

In practice, the logarithm $log\, L_{\mathcal{T}}(\alpha, \lambda, \mu)$ of this function is used. Because of the high complexity of this log-likelihood function it is not possible to compute the partial derivatives for α, λ and μ as closed formulas and hence numerical strategies must be used. Since

α is an integer number, a discrete optimization strategy is needed. λ and μ on the other hand are real numbers and can be calculated by continuous optimization strategies. Both types are briefly discussed in the next section.

2.2.4 Optimization strategies

The optimization of a given function f depending on one or more variables is an essential task in many mathematical problems. During an optimization process the method determines values of the variables where f adopts an extremum, i.e. either a minimum or maximum. An optimization method should be quick and cheap, but if the costs of evaluating f are high, f should be evaluated as few times as possible.

For different requirements different optimization strategies were developed. Methods like the *bisection method* or *golden section method* do only evaluate the function itself, whereas other methods like the *Newton-Raphson method* also require evaluations of the derivatives of the function (Acton (1970), Ralston & Rabinowitz (1978)). A description of these methods as well as a couple of other methods can be found in Press *et al.* (1992) (chapter 9 and 10).

For the implementation of the ML estimation in our approach two different optimization strategies were used. The duplication rate λ and the deletion rate μ are optimized using a one-dimensional continuous optimization method, without the need of the derivatives. This method is knows as *Brent's algorithm* (Brent (1973)). For the estimation of the gene copy number α we applied a discrete optimization method, similar to the bisection method presented in Press *et al.* (1992) (section 9.1).

CHAPTER

THREE

Computation of the BD probability distribution

The computation of mathematical formulas using a computer is not necessarily easy, since the representation of numbers and the coherent workflow of arithmetic operations is not equal to the calculations by hand. Inconceivable errors can occur during such computations, which are often hard to explain and difficult to predict.

In this work, we have developed a software for the estimation of duplication and deletion rates as well as the ancestral number of gene copies of a certain gene family based on a BD model (see section 2.1). The optimization of these parameters requires many evaluations of the corresponding probability function. As we will show, the nature of this calculations demands for high accuracy computation to produce meaningful results.

To better understand the problem of high accuracy calculations, we start with a short overview of number representation, arithmetic operations, and possible errors in the computer. This summary is based on Bohn & Flik (2005), Hennessy & Patterson (1996), Goldberg (1991), and Press *et al.* (1992). Subsequently, we discuss specific problems in computing the probability function of our BD model as well as the maximum likelihood function and different solutions.

3.1 Computer arithmetic

3.1.1 Number representation in the computer

In mathematics there are different sets of numbers, like e.g. the natural numbers $\mathbb{N}$, integers $\mathbb{Z}$, or real numbers $\mathbb{R}$. These sets contain infinite many elements. In the computer, numbers can only be stored to a finite quantity and with finite precision. All numbers are stored in bits or bytes (8 bits), whereas a bit is either 1 or 0. Therefore, a representation of numbers in the decimal system is not suitable. Instead, the numbers are represented in the binary system. There are different representations or data types in a computer and a programmer usually has to choose between them. These data types differ in the number of bits which are used to store the value of a variable in the format of the specific representation. Two commonly used formats are the fixed point (`int` and `long`) and the floating point (`float` and `double`) representation.

Fixed point numbers. Natural numbers ($\mathbb{N}$) are represented in fixed point format. A natural number z can be displayed in the binary system using the numerics $a_i \in \{0,1\}$ by:

$$z = (a_{n-1}a_{n-2}\dots a_2a_1a_0)_{\mathbf{2}} = \sum_{i=0}^{n-1} a_i \mathbf{2}^i \tag{3.1}$$

For signed numbers, e.g. from $\mathbb{Z}$, there are two possibilities for the representation. One is an additional bit for the sign, another one is the two's complement notation, which is typically used in a computer.

In fixed point representation also non integer numbers can be displayed. Therefore, the binary point is always at the same location in the number format: $(a_{n-1}a_{n-2}\dots a_q.a_{q-1}\dots a_2a_1a_0)_{\mathbf{2}}$. The disadvantage of fixed point numbers is its fixed resolution. In general, it would be desirable to fit the resolution to the basis of the absolute size of the number, which can be done with the floating point representation.

Floating point numbers. A floating point number z is represented by its sign s, its mantissa m, and its exponent E (IEEE 754 standard):

$$z = (-1)^s \cdot m \cdot \mathbf{2}^E \tag{3.2}$$

If the sign bit s is 0, then the number is positive, if s is 1, then the number is negative. The mantissa m is a normalized binary fixed point number, which means that the binary point is shifted left or right until the mantissa starts with '1.'. Simultaneously the exponent E is decreased or increased. Then the range of the mantissa is $1.0 \leq m < 2.0$. In the computer, a specific number of bits is allocated for each part of this format. For a `float` variable in total 32 bits are reserved, for a `double` variable 64 bits. They are assigned as follows:

data type	sign	mantissa	exponent	sum
`float`	1 bit	23 bit	8 bit	32 bit
`double`	1 bit	52 bit	11 bit	64 bit

In this representation all numbers of a data type have roughly the same precision and no bit is wasted. The following table shows some important characteristics for the two data types `float` and `double`:

data type	`float`	`double`
greatest relative error	2^{-24}	2^{-53}
accuracy	7 decimal digits	16 decimal digits
bias	127	1023
smallest positive number	$2^{-126} \approx 1.2 \cdot 10^{-38}$	$2^{-1022} \approx 2.2 \cdot 10^{-308}$
largest positive number	$(2 - 2^{-23}) \cdot 2^{12} \approx 3.4 \cdot 10^{38}$	$(2 - 2^{-52}) \cdot 2^{1023} \approx 1.8 \cdot 10^{308}$

Although very small numbers as well as very big numbers can be stored in this representation, the precision of the numbers is limited and many numbers can only be stored rounded.

3.1.2 Computational pitfalls

The representation of numbers in the floating point format and the corresponding arithmetic operations can lead to many errors in the results. Here we will briefly discuss the most important sources of errors. More information can be found in Press *et al.* (1992) (section 1.3.) and Goldberg (1991).

Overflow and underflow. Common problems particular during multiplications are overflow and underflow of variables. An *overflow* occurs, if a positive number becomes larger than the largest positive number of the data type used. On the other hand, if the number becomes smaller than the smallest positive number, an *underflow* occurs. An underflow is normally set to zero automatically without any problem, whereas an overflow is mostly a disaster in the calculations. The number then becomes 'positive infinity' (∞) or 'not a number' (NaN) and the calculation becomes meaningless. The same holds for negative numbers respectively.

Roundoff errors and catastrophic cancellation. Most floating point numbers cannot be represented exactly. Those that can are called machine numbers. The decimal 0.25 e.g. is a machine number, whereas 0.2 on the other hand is not and cannot be represented exactly. The corresponding floating point representation yields to the decimal number 0.199218.... That means the desired decimal number 0.2 differs from its representation in the computer. The maximum deviation is related to the machine precision or machine accuracy ϵ_m. It is defined as the largest number ϵ for which $1+\epsilon = 1$ for a given data type. For example, the machine precision ϵ_m for `float` is normally about $3 \cdot 10^{-8}$. Arithmetic among floating point numbers is also not necessarily exact, even if the operands happen to be exact. In almost all operations an additional fractional error of at least ϵ_m can be introduced. This type of error is called *roundoff error*.

For example by adding two floating point numbers, the mantissa of the smaller one is shifted to the right and digits get lost. If the two numbers differ too much in the order of magnitude, the smaller one becomes zero because all significant digits are shifted out of the numerical range. In this case the sum would be equal to the larger number.

During the multiplication of floating point numbers it is also possible that significant digits of the mantissa of the numbers get lost. If the product of two numbers has more digits than bits available, some of the least significant digits get out of range and the number is saved rounded.

Repeated operations can accumulate roundoff errors which can lead to serious errors in the result. Roundoff errors are very frequent in computational analysis.

One of the worst errors in computer based computations is the *catastrophic cancellation*, which occurs usually in conjunction with rounding errors. It can happen during subtraction of nearby floating point numbers with equal sign. We illustrate this effect using a simple example for the decimal system:

Consider the term $b^2 - 4ac$ (part of the formula for the solution of a quadratic equation) to be computed with $a = 1.22$, $b = 3.34$, and $c = 2.28$ and the fixed number of digits $n = 3$ for their representation. We expect the result to be:

$$3.34 \cdot 3.34 - 4 \cdot 1.22 \cdot 2.28 = 11.1556 - 11.1264 = 0.0292 \tag{3.3}$$

In a computer, the computation including rounding to the given precision after each operation results to:

$$3.34 \cdot 3.34 - 4 \cdot 1.22 \cdot 2.28 = 11.2 - 11.1 = 0.1 \tag{3.4}$$

As can be seen, only the first digit from this result matches the exact result and further computations will not succeed to give the correct result. In literature, several, more complex examples for catastrophic cancellation can be found, which show the danger of this problem, since it might not be detected (Goldberg (1991), Cuyt *et al.* (2001)).

Conclusion. These problems show that numerical computation of even simple terms can lead to false results. Thus, results have to be checked carefully. In addition, identical outputs in different precisions do not necessarily imply the correctness of the computation (Cuyt *et al.* (2001)) and therefore it is sometimes very hard to verify the validity of results. Since these errors also depend on the computer hardware, it is important to know about possible errors and their characteristics on the underlying system. On the other hand, the formulas which should be computed have to be analyzed carefully, to measure numerical ranges of their particular components and detect critical parts in their computation.

3.2 Problems in the computation of the BD probability distribution

Our BD model has the nice property that the function for computing the probability distribution can be displayed in an explicit form. Here we will consider the most general case, that is birth and death rate are unequal ($\lambda \neq \mu$), the number of gene copies of the parent is greater than one ($k > 1$), and the actual number of gene copies is greater than zero ($i > 0$). The corresponding function looks like this:

$$\begin{aligned} p_i(k,t) &= \sum_{j=0}^{\min(k,i)} \binom{k}{j}\binom{k+i-j-1}{k-1} A^{k-j}\, B^{i-j}\, (1-A-B)^j \quad \forall i \geq 1, \forall k > 1 \\ \text{with } A &= \frac{\mu(e^{(\lambda-\mu)t}-1)}{\lambda e^{(\lambda-\mu)t}-\mu}, \quad B = \frac{\lambda(e^{(\lambda-\mu)t}-1)}{\lambda e^{(\lambda-\mu)t}-\mu} \end{aligned} \tag{3.5}$$

To avoid computational pitfalls, we will analyze the main components of the function $p_i(k,t)$ regarding its computability. Furthermore possible errors in the computation will be pointed out and the dependency on the precision is demonstrated.

3.2.1 Analysis of the components

First of all, the numerical ranges of the parameters of the function $p_i(k,t)$ have to be defined. That are the birth rate λ, the loss rate μ, the time t, the initial number of gene copies k, and the actual number of gene copies i. Since the number of gene copies is a positive integer, k and i are defined to be natural numbers and non-zero (eq. 3.6).

$$k, i \in \mathbb{N}^+ \equiv \mathbb{N}\backslash\{0\}\,; \qquad t \in \mathbb{R}^+\,; \qquad \lambda, \mu \in \mathbb{R}^+ \cup \{0\}\,,\ 0 \leq \lambda, \mu \leq 1 \tag{3.6}$$

The time t can adopt arbitrary positive numbers with the exception of zero. $t = 0$ would imply that the corresponding branch in the tree has the length zero which in turn would lead to another tree topology. Therefore $t = 0$ is not allowed. For the rest of this thesis t will always be given in million years (myr) or million years ago (mya). The two rates λ and μ can adopt arbitrary non-negative real numbers including zero, since it is possible that no duplication or deletion has occurred (eq. 3.6). From literature (Lynch & Conery (2003), Cotton & Page (2005), Hahn *et al.* (2005), Demuth *et al.* (2006)) and simulation studies it turns out, that λ and μ are between 0 and 1. So these numbers are defined to be the boundaries for both rates.

Looking at the two auxiliary variables A and B from eq. 3.5 more closely, it emerges that they only differ in the first factor of the numerator. It follows for A and B:

$$0 \leq A, B \leq 1 \tag{3.7}$$

To distinguish between the particular factors of the summands from the function $p_i(k,t)$, the following notation is used with j standing for the component of the jth summand respectively:

$$p_i(k,t) = \sum_{j=0}^{\min(k,i)} \underbrace{\underbrace{\binom{k}{j}}_{factor\,1_j} \quad \underbrace{\binom{k+i-j-1}{k-1}}_{factor\,2_j} \quad \underbrace{A^{k-j}}_{factor\,3_j} \quad \underbrace{B^{i-j}}_{factor\,4_j} \quad \underbrace{(1-A-B)^j}_{factor\,5_j}}_{prod_j} \tag{3.8}$$

Example parameter set. To point out problems in the computation, an example parameter set is used in the remaining analysis in this chapter. This example set is taken from one of our computations where fatal errors occurred and is based on a gene family of the HOGENOM database (Duret *et al.* (1999)) with the ID HBG126803 and the phylogenetic tree tMRHF from figure 1.3. The gene copy numbers in this family are 90 for mouse, 93 for rat, 101 for human and 10 for fruit fly. During the optimization procedure the probability function $p_i(k,t)$ received negative values as well as values bigger than one. In the following, the probability function is considered for the branch leading from the MRCA (MRHF) to the ancestor of mouse, rat and human (MRH). The time between them is 899.0 myr (Hedges (2002)) and critical values for the gene copy numbers are 150 for MRHF and 50 for MRH. The parameters for the BD function are summarized in eq. 3.9.

$$\begin{aligned} k &= 150 \\ i &= 50 \\ t &= 899.0 \\ \lambda &= 0.002049347886564 \\ \mu &= 0.002456230599999 \end{aligned} \tag{3.9}$$

The specific values for the factors using this parameter set are calculated with Mathematica (Wolfram Research, Inc. (2005) and section 3.4). Using a precision of 100 decimal digits, the values of the factors and the values of the whole function are computed correctly. The influence of the precision of the numbers in such calculations is discussed later in more detail (subsection 3.2.3 and section 3.4).

***factor*1 and *factor*2.** Analyzing the defined factors of the function $p_i(k,t)$ from eq. 3.8, we see that the binomial coefficients $factor\,1$ and $factor\,2$ and the product $factor\,1 \cdot factor\,2$ can get very large. For the chosen example for different j they are

$$\begin{array}{lll} j{=}0 & factor\,1_0 & = 1 \\ & factor\,2_0 & = 3403937834424345458005810427106511220714989953 96 \\ j{=}25 & factor\,1_{25} & = 1956464059530021643973123625 6 \\ & factor\,2_{25} & = 1087562120541667885404612090756 \\ & factor\,1_{25} \cdot factor\,2_{25} & = 21277762013460302927610454847901804472401816728260749649536 \\ j{=}45 & factor\,1_{45} & = 44166856121803108968165394209290808840 0 \\ & factor\,2_{45} & = 675993780 \\ & factor\,1_{45} \cdot factor\,2_{45} & = 29856520020493824047142024496728604987009015200 0 \end{array} \tag{3.10}$$

***factor*3 and *factor*4.** On the other hand, the two auxiliary variables A and B only adopt values between 0 and 1. If such a small number is raised to a power ($factor\,3$ and $factor\,4$), it gets even smaller. In our example, A and B take the values:

$$A = 0.72723087402374218676540 14\,, \quad B = 0.60676267721976652187665 87 \tag{3.11}$$

Then, $factor\,3$, $factor\,4$, and the product of both become:

$$\begin{array}{lll}
j{=}0 & factor\mathit{3}_0 & = 1.78174782038196321246021 4{\cdot}10^{-21} \\
 & factor\mathit{4}_0 & = 1.41560779440731876482212 9{\cdot}10^{-11} \\
 & factor\mathit{3}_0{\cdot}factor\mathit{4}_0 & = 2.52225610220095850200245861666743749367411312653 98{\cdot}10^{-32} \\
j{=}25 & factor\mathit{3}_{25} & = 5.11727233443939079939626 8{\cdot}10^{-18} \\
 & factor\mathit{4}_{25} & = 3.76245637105245217064319 5{\cdot}10^{-6} \\
 & factor\mathit{3}_{25}{\cdot}factor\mathit{4}_{25} & = 1.92535138971219406682609878729888745672991546352 41{\cdot}10^{-23} \\
j{=}45 & factor\mathit{3}_{45} & = 2.98945737430133270926485 3{\cdot}10^{-15} \\
 & factor\mathit{4}_{45} & = 0.08224211956077789798646 733 \\
 & factor\mathit{3}_{45}{\cdot}factor\mathit{4}_{45} & = 2.45859310799139369014797640266132653058905804378 71{\cdot}10^{-16}
\end{array} \tag{3.12}$$

***factor*5.** The last factor $factor\,5$ has a particular status. The sum $A + B$ can be larger than 1, since eq. 3.7 holds. In this case, $factor\,5$ gets a negative value which is raised to the power of j. That means, it cannot only get very small but also changes the sign for every odd j. This results in an alternating sum in the function $p_i(k, t)$. For the chosen example parameters (eq. 3.9) $factor\,5$ shows this behavior, as displayed below:

$$\begin{array}{llll}
j{=}44 & factor\,5_{44} & = & 1.10783628883685222747398 9{\cdot}10^{-21} \\
j{=}45 & factor\,5_{45} & = & -3.70010176305049719197771 5{\cdot}10^{-22}
\end{array} \tag{3.13}$$

Even if the precision of the values is limited, the computation of every single factor is not problematic. The number of the exact digits at the beginning corresponds to the accuracy of the particular data type (see subsection 3.1.1). Thus, the values for the different factors are as good as they can be in this context.

3.2.2 Possible errors

Nevertheless in the computation of the function $p_i(k, t)$ errors can occur, due to the accuracy of the commonly used variables in the computer. Thereby all possible sources of errors described in subsection 3.1.2 can emerge during the evaluation of $p_i(k, t)$.

prod. As already mentioned, the factors from eq. 3.8 can be computed with sufficient precision (mostly rounded), which apply also for the selected example (eq. 3.9). When multiplying all factors, starting with $((factor1 \cdot factor2) \cdot factor3) \cdot \ldots$, roundoff errors occur and accumulate. None the less, the single products of the five factors for each j are accurate in mostly 14 of the first digits, when using `double`. In the example for $j = 46$

the product became:

$$\begin{array}{lll} \texttt{double} & : & prod_{46} = \mathbf{1.523586725116547} \cdot 10^{9} \\ \text{high accuracy} & : & prod_{46} = \mathbf{1.5235867251165}18773245267 \cdot 10^{9} \end{array} \tag{3.14}$$

Here again Mathematica with precision 100 was used to compute the results with high accuracy.

Sum. In the probability function these products are added up. During the addition of two floating point numbers lots of information can get lost due to shifting (see subsection about roundoff errors (3.1.2)). In our example, an alternating sum arises. To further complicate matters, the summands (equivalent to the products $prod_j$) are in a range, where not only the summand but also the sum changes the sign in every step. If the summand is negative the value of the sum becomes negative and the other way round. That means, that every addition becomes a subtraction. As shown in subsection 3.1.2 a big problem in conjunction with subtraction is the catastrophic cancellation and exactly this error occurred in the computation of $p_i(k,t)$ for certain parameter sets. The particular sum s_{42}, where $s_j = \sum_{x=0}^{j} prod_x$, has barely three correct digits after adding the 42nd summand:

$$\begin{array}{lll} \texttt{double} & : & s_{42} = \mathbf{4.39}7061740304023 \cdot 10^{12} \\ \text{high accuracy} & : & s_{42} = \mathbf{4.39}6789636691625755834340 \cdot 10^{12} \end{array} \tag{3.15}$$

When adding the 46th summand no digit of the sum with normal `double` precision is correct and after adding the 47th summand even the sign of the sum changed:

$$\begin{array}{llrl} \texttt{double} & : & s_{46} = & 3.367233215185742 \cdot 10^{8} \\ \text{high accuracy} & : & s_{46} = & 6.461970920407727 \cdot 10^{7} \\ \texttt{double} & : & s_{47} = & 2.700091439904314 \cdot 10^{8} \\ \text{high accuracy} & : & s_{47} = & -2.094468324064392 \cdot 10^{6} \end{array} \tag{3.16}$$

After this addition, or in fact subtraction, the result computed with `double` precision is totally different from the correct result.

Entire probability. In the end the probability $p_i(k,t)$ for the parameters of the example (eq. 3.9) results in:

$$\begin{array}{lllll} \texttt{double} & : & p_i(k,t) & = & p_{50}(150, 899.) = 2.721036123146133 \cdot 10^{8} \\ \text{high accuracy} & : & & & p_{50}(150, 899.) = 0.000117497694394018 \end{array} \tag{3.17}$$

For this particular parameter set it is easy to recognize that the result is wrong, because probabilities have to be in the interval $[0,1]$ and the result for `double` is clearly bigger.

Putting it all together, using normal `double` variables in the computation of the probability distribution $p_i(k,t)$ of the BD process can easily lead to wrong results. Many of these incorrect results can be detected by checking the ranges for the probability, or through the appearance of 'NaN' values due to overflow or underflow, but most might not be detected, since they do not show such an extreme behavior. However, the detection of errors was the first important step, but the major task was to find a way to compute the probabilities correctly avoiding these errors. Therefore, data types with high accuracy can be used. In the following the dependency of the result on the precision of the values is demonstrated.

3.2.3 Dependency on the accuracy of numbers

Computational errors can be reduced by increasing the precision. But the computation with high precision does not come for free. A higher precision requires more resources and more computing time, since special data structures and operations are used. Hereupon the question arises how high the precision has to be to ensure that the probabilities are computed correctly, while keeping the expenses as low as possible. That is difficult to answer in general, but it is possible for our example.

precision	probability $p_{50}(150, 899.)$
10	$0. \cdot 10^{19}$
20	$0. \cdot 10^{9}$
30	$0. \cdot 10^{-1}$
33	$0. \cdot 10^{-4}$
34	0.0001
40	0.0001174977
50	0.000117497694394018
60	0.000117497694394017840832653403
80	0.00011749769439401784083265340286960616772848152778

With a precision of 50 digits the resulting probability has 15 correct digits, which corresponds almost to the accuracy of a `double` variable. But to compute the whole probability function $p_i(k, t)$ for all i and the remaining parameters from eq. 3.9, a precision of 80 is needed. In figure 3.1 the function $p_i(k, t)$ is plotted for normal `double` precision and for precision 80.

With normal `double` precision the probability function can be calculated up to $i = 35$ with an acceptable accuracy, after that the function gets out of hand and reaches values between $-2.6216 \cdot 10^{45}$ and $4.2819 \cdot 10^{46}$. With the high precision of 80 the functions looks fine and the area under the function for $i \in [0, 200]$ is almost 1 (0.99999299938920). Thus, one way to guarantee the correct computation of the probability function is the usage of high accuracy variables.

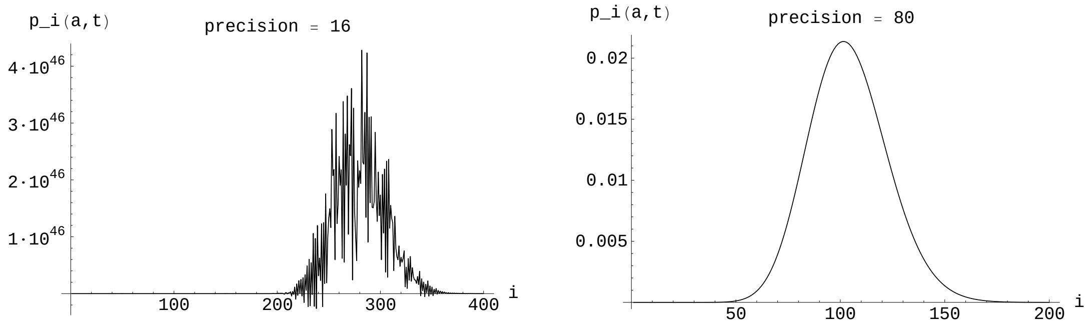


Figure 3.1: Probability function $p_i(k,t)$ from eq. 3.5 for the example parameters from eq. 3.9. The left plot shows the function for normal `double` precision (16 digits) and the right plot shows the function for precision 80. Both plots are generated with Mathematica.

For the implementation of the probability function in our software we had to find practicable solutions to ensure the correctness. For this purpose different approaches were evaluated. On the one hand, we examined whether it is possible to display the probability function $p_i(k,t)$ in another form as well as if changes in the order of the computation simplifies the entire calculations. On the other hand, it was sought-after computer-based solutions to provide the necessary accuracy of the numbers in the computation.

3.3 Mathematical solutions

The best solution to get the probability function under control would be on a mathematical level, since computer-based solutions mostly requires more running time and can also reach limits in the application. So the first idea was to change the order of the operations to minimize the possibilities of errors. We also analyzed if another representation of the probability function could simplify the problem.

3.3.1 Order of computations

The most critical arithmetic operation computing $p_i(k,t)$ is the summation, since roundoff errors and catastrophic cancellation can occur. To avoid unnecessary subtractions two sums can be used instead of one. The first is a positive sum where all positive summands are added up. The second is also a positive sum where all negative summands multiplied by (-1) are added up to avoid subtractions. In the end the second sum is subtracted from the first one to get the probability. That makes only one subtraction in the entire computation and might help to avoid catastrophic cancellation.

Unfortunately, the use of the two sums did not lead to the correct result in our example and also the intermediate results did not get better. After the last summations the two sums sum_{pos} and sum_{neg} result in

$$\begin{array}{lcl} \texttt{double} & : & sum_{pos} = 2.057269189586888 \cdot 10^{25} \\ & & sum_{neg} = 2.057269189586888 \cdot 10^{25} \\ \text{high accuracy} & : & sum_{pos} = 2.0572691895868411386069769709\mathbf{74953} \cdot 10^{25} \\ & & sum_{neg} = 2.0572691895868411386069769709\mathbf{63204} \cdot 10^{25} \\ \texttt{double} & : & sum_{pos} - sum_{neg} = 4.294967296 \cdot 10^{9} \\ \text{high accuracy} & : & sum_{pos} - sum_{neg} = 0.0001174976943940178408326534 \end{array} \tag{3.18}$$

The correct values for the sums are equal up to 29 digits, only then the two sums differ. Thus, the difference between sum_{pos} and sum_{neg} cannot be detected when using `double` variables, since the precision is only 16 decimal digits. Instead an arbitrary number is assigned to the difference, which has nothing in common with the correct solution. Consequently, for this parameter set it is impossible to get the right result by changing only the order of the computation.
However, since this approach basically tends to avoid errors, it is used in our implementation of the probability function.

3.3.2 Different representation of the probability function

It is also possible to search for another representation of the function $p_i(k,t)$. In this context, it can be tested if the function can be computed recursively or if it is possible to find a good approximation. Furthermore, it is possible to get the Taylor coefficients based on the generating function (see appendix B eq. B.10) and use these coefficients as separate functions $p_i(k,t)$ for every i. These possibilities were tested, but none of them led to a better way to compute $p_i(k,t)$ properly and so they are not explained further.
According to this, the only promising approach to get the correct results was the use of a computer-based solution which supports high accuracy computations. In the next section we will give an overview of some of these computer-based solutions.

3.4 High accuracy computation

In lots of applications, it is essential that numbers can be represented with high accuracy (Cowlishaw (2003)). On the one hand it is important for the representation of the numbers itself, because of the lack of representing even easy numbers, like 0.1 or 0.2, by binary fractions (see subsection 3.1.2). On the other hand, there are many applications, where arithmetic operations have to give exact results, in the sense of match exactly those results, which might be calculated by hand. That hold, for example, for financial applications, where the correctness of rounded results is really important and also the smallest numbers and the least significant digits of a decimal number have to be exact. Due to these demands there are several data types, libraries, and special programs which can deal with high accuracy numbers and arithmetics. Some of them are described below, whereas mainly solutions for the programming languages C and JAVA are introduced.

3.4.1 Overview

In this subsection a short survey of data types, libraries and special programs is given, which were evaluated for the implementation of our software. Some programs, like for example Mathematica, are especially used to compare results and to validate the correctness of computations.

Mathematica (Wolfram Research, Inc. (2005)) is one of the most powerful mathematical software system used for a wide field of applications, including numerical evaluations with an arbitrary precision of the numbers. Furthermore a huge amount of functions is provided for numerical calculations and for visualization of data and functions. It is the only commercial program used in this thesis and serves mainly as a tool to check the correctness of the results of the other programs. Furthermore it is used to generate plots for different requirements. Mathematica was nor chosen for the implementation of the parameter estimation due to its very long running time.

PARI is a computer algebra system developed for efficient and fast calculations for applications in number theory (factorization, algebraic number theory, elliptic curves, . . .) (The PARI-Group (2000)). It was originally developed from Henri Cohen et al. (Université Bordeaux I, France) and is now maintained by Karim Belabas. PARI is liable to the GNU General Public License and can be used as a library, integrated in a program written in C. PARI is rather simple, compared to other more sophisticated computer algebra systems, like Mathematica, but it is supposed to be faster and provides various modules for algebraic number theory.

ARIBAS is an interactive interpreter for big integer arithmetic and multi-precision floating point arithmetic (Forster (2004)). It was developed by Otto Forster (LMU Munich, Germany) in 1996 for applications in algorithmic number theory and has several built-in functions, like Jacobi symbol, Rabin probabilistic prime test and factorization algorithms. ARIBAS is written in C and there are versions for UNIX/LINUX, Windows, and MS-DOS available. It is also distributed under the GNU General Public License.

Decimal floating point arithmetic. Another approach, that is widely used in computer science, is the implementation of decimal floating point arithmetic on the software and on the hardware level. This makes sense regarding the spread of numbers in decimal representation. For instance, a study from Cowlishaw (2003) shows that data from commercial databases contain mainly decimal data. The analyzed databases cover a wide range of applications, including airline systems, banking, financial analysis, insurance, inventory control, management reporting, marketing services, order entry, order processing, pharmaceutical applications, and retail sales.

Today many programming languages support decimal floating point arithmetic, including C (decNumber C package), and Java (BigDecimal class) (Cowlishaw (2003)). Since the advantages of this arithmetic became apparent and due to the reason that processing time in decimal arithmetics is limited computer-bound, developments of decimal floating point hardware has started. One decimal-encoded format proposed by the IEEE 754 revision committee, is already implemented in the IBM System z9 (mainframe) processor.

3.4.2 Application for the BD distribution

Our software for the estimation of gene family specific gene duplication and deletion rates and the ancestral gene copy number, was initially written in the programming language C. Therefore is was essential to find a software solution in C to compute the probability function $p_i(k, t)$ correctly.

`double` and PARI-library. During the implementation of our program, it was soon obvious that normal `double` variables with a precision of 16 digits were not suitable for the computation of the probability function. Only for some cases the accuracy in the arithmetic operations was sufficient. Hence, in the first implementation, PARI was used as a C library to support high accuracy variables. The precision can be set to an arbitrary number and was chosen to be 100 decimal digits. The computation seemed to work fine, until real datasets were used. There, probabilities appeared, which were negative or much bigger than 1 (see e.g. gene family HBG126803, subsection 3.2.1). By analyzing the computation of the incorrect probabilities, catastrophic cancellation errors were found. This could be ascribed to the internal precision of the PARI system, which is fixed to 38 digits and is not automatically set to the precision chosen from the programmer. It was not possible to change this internal precision and so the computations could only be performed using the default internal precision of 38 significant digits, independent of the chosen external precision. For the simulated gene families by then, this precision was apparently sufficient, but real datasets and other following simulation studies could not be processed and therefore this solution was discarded.

For a meaningful visualization of the observed behavior of the probability function using different data types, another parameter set was taken and the corresponding images can be found in figure 3.2. While for `double` variables (figure 3.2: DOUBLE) only the first few probabilities are computed correctly before the function starts to oscillate, the function could be evaluated correctly up to $i = 50$ using the PARI data types (figure 3.2: PARI). After that the functions also starts to oscillate.

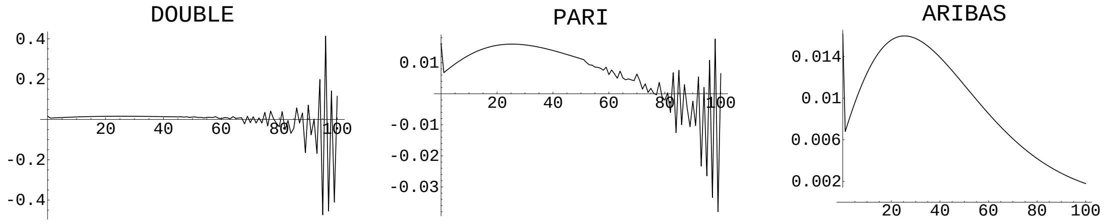


Figure 3.2: Probability function $p_i(k,t)$ from eq. 3.5 for the parameters: $\lambda = 0.06$, $\mu = 0.05$, $t = 100$ and $k = 15$. The first plot shows the function for normal `double` precision (16 digits), the second function is computed using data types of the PARI library (precision 38) and the last plot shows the function for precision 154 , computed with ARIBAS. On the x-axis the number of gene copies i is given and on the y-axis the corresponding probabilities $p_i(k,t)$. All plots are generated with Mathematica.

ARIBAS is a self-contained program with the possibility to choose the desired precision of the numbers between 32 and 4096 bits. 32 bits correspond to a normal `float` variable with 9-10 decimal digits accuracy, while 4096 bits correspond to a variable with 1232 decimal digits accuracy. This precision applies not only to the given parameters and the result but also to all variables used in the computation and in all arithmetic operations. Hence, a procedure was written in ARIBAS for computing the probability function $p_i(k,t)$ and a procedure to calculate the log-likelihood function for a tree. The precision was set to 512 bits, which is an accuracy of 154 decimal digits. These procedures provided the correct results for all tested parameter sets.

The rightmost plot in figure 3.2 shows the probability function computed with ARIBAS. In contrast to the other two plots the function shows no oscillations. The results of the ARIBAS procedures were checked against the results computed with Mathematica and no differences could be found, so they were assumed to be right.
Thereupon the C program was changed to the effect, that the ARIBAS procedure for computing the log-likelihood function was called as an external program. This approach worked very well and was used for many studies in this thesis. But one disadvantage still remains, which is the overall running time. The running time for estimating all three parameters depends on the used tree and turns out to be about five hours on average for the tree tMRHF (figure 1.3). If a dataset is very big, for example the HOGENOM dataset containing about 1700 gene families, the running time is very long.

Java - BigDecimal. One big problem for reducing the runtime in the C program is the usage of the external ARIBAS program. Every call of the procedure costs time, which cannot be reduced due to technical limitations. Furthermore, it was not possible to include other features, like lookup tables, to save computing time.
Therefore, the C program was rewritten in another programming language, which provides a possibility to include high accuracy variables directly in the program. We have chosen *Java* and its special classes *BigDecimal* and *BigInteger*. In the new Java program the probability function $p_i(k,t)$ is computed using these classes as well as the log-likelihood

function, as in the C-ARIBAS program. As for most programs or packages, the precision can be specified at the beginning, and was set to 154 decimal digits in the Java program, which is the same precision used in the ARIBAS procedures. Comparisons of the resulting values for the probability and the log-likelihood function with the C-ARIBAS program and Mathematica showed, that the results were identical. Hence, the rightmost plot in figure 3.2 is equal to the one for Java (not shown).

3.4.3 Running times

To compare the mentioned software solutions with respect to the running times, some tests were made. On the one hand, we evaluated the running time for computing a single probability and the probability distribution of the BD model. On the other hand, the runtime of the computation of the log-likelihood function for a specific tree as well as the entire estimation of the duplication rate λ, deletion rate μ, and the ancestral gene copy number α was measured. The results are given in the following.

BD Probability function

Test setting. Three example parameter sets were chosen, for which $p_i(k,t)$ was calculated several times. These parameters can be found in table 3.1. The third example *Ex3* corresponds to our error example from subsection 3.2.1.

parameter	*Ex1*	*Ex2*	*Ex3*
λ	0.4	0.06	0.002049347886564641
μ	0.3	0.05	0.002456230600000000
t	5.0	100.0	889.0
k	5	10	150

Table 3.1: Parameter sets for three different examples *Ex1*-*Ex3* used to compare the running times for different programming languages and libraries.

We measured the running time for different programs with the *time*-function of Linux. The used PC was an AMD Athlon 64 Processor 3500+ with a clock rate of 2,2 GHz and a memory of 2 GB (DDR2 RAM, 667 MHz). We tested two C programs, one with `double` variables and another one with variables from the PARI library. The third candidate was our Java program including the *BigDecimal* and *BigInteger* classes. It was run with Java version 1.6 (released December 2006). The last test candidate was written in ARIBAS.

For all parameter sets the probability function $p_i(k,t)$ was computed until 99% of the distibution was achieved. Therefore the interval for i and consequential the number of calls of $p_i(k,t)$ was different for each parameter set. Since in most cases the running time was very short, the computation of the whole distribution was repeated. To measure the times for the multiple computations of the probability distributions the whole computation was once more repeated 1000 times to get an average value for each example.

Ex1. For the first example *Ex1* the interval for i was found to be between $i = 0$ and $i = 27$ and the computation of the distribution was repeated 40 times. This corresponds to 1080 calls of the probability function in total. For the high accuracy variables of the different programs the precision was set to 16 decimal digits which is equal to `double` precision.

Ex2. In *Ex2* the probabilities had to be computed untill $i = 99$ to get 99% of the probability distribution and this was repeated 10 times, resulting in 990 calls of the probability function in total. Here the precision was set to 38 decimal digits, which corresponds to the internal precision of the PARI library.

Ex3. For the last example *Ex3*, i went up to 151. Since here it took much more time to compute the whole distribution, it was only computed once. 1000 repetitions gave the runtime average. The precision in this case was set to 100 decimal digits for the high accuracy variables, since in subsection 3.2.3 was shown, that this precision is sufficient for this specific example.

Resulting running times. Table 3.2 contains the running times in milliseconds for the three example parameter sets and the different programs. For *Ex1* the running times are in the first row. The fastest program was the C-`double` program with 0.07 ms, followed by the C-PARI with 112 ms, ARIBAS (480 ms), and the last Java (1,339 ms).
In the second row of table 3.2 the running times for *Ex2* are given. Since the accuracy of the C-`double` program was not sufficient, the fastest program here was the C-PARI program with 170 ms, then ARIBAS (1,403 ms), and again last Java (3,581 ms).
For *Ex3* ARIBAS appeared to be the fastest program with 16,843 ms, followed by Java with 259,924.
If the distribution for a parameter set could not be calculated due computational errors, the corresponding values in table 3.2 are not specified (indicated with a dash).

Discussion. The C-`double` program was only applicable for the first example parameter set, while the C-PARI program worked for the first and the second, but not for the third parameter set. In contrast, the Java program as well as the ARIBAS program were applicable for all examples. Although the C-`double` program and the C-PARI program are preferable with respect to running times, they were discarded because it could not be guaranteed to get the correct results. For some cases they might work well, but for others they will not.
The probability computation with the ARIBAS program was the fastest of the two remaining programs, followed by the Java. Thereby the ARIBAS computation was about

	C		Java - BigDecimal	ARIBAS
	`double`	PARI library		
Ex1	0.07	112	1,339	480
Ex2	-	170	3,581	1,403
Ex3	-	-	259,924	16,843

Table 3.2: Running times measured over 1000 repetitions in milliseconds. Parameters used for the computation are from table 3.1. The probability function $p_i(k,t)$ was computed for increasing i until 99% of the distibution were achieved. Further details in the text.

2.5 times faster than the Java computation for the first two examples, and incredible 15 times faster in the third example. This last example is indeed a special case because the gene copy numbers in the gene family HBG126803 from the HOGENOM database (subsection 3.2.1) are exceptional high (10-101 copy numbers). Gene copy numbers in this range are very rare and most gene families have much lower copy numbers. If only these running times would be taken into account, the ARIBAS program would be the best choice.

Log-likelihood function and entire estimation procedure

Nevertheless, the runtime of a single probability call is only one part that influences the entire running time of the programs. Hence, the log-likelihood function, which depends on the chosen tree, was also analyzed with respect to the runtime. The remaining programs, the Java and the ARIBAS program, were compared for two example cases.

Test setting. In both cases the log-likelihood function for the tree tMRHF (figure 1.3) was calculated, which can be found in eq. 2.35. In the first example *Loglik1* the log-likelihood function was computed for five simulated gene families, named a,b,c,d, and e. The parameters λ, μ and α, used for the simulation of example *Loglik1* are given in table 3.3 as well as the gene copy numbers i_X for every species X of the gene families. In contrast, the gene copy numbers in the second example *Loglik2* are taken from real data (gene family HBG126803). This example is similar to the third example *Ex3* from table 3.1 and again acts as an example with extreme values.

The Java program was tested again for the Java version 1.6. Moreover two different versions of the Java program were tested. In the first basic version the probability function and log-likelihood function are used without any further features (denoted by 'without lookup' in the table). In the second Java program version lookup tables for binomial coefficients and for values of the probability function have been implemented (denoted by 'with lookup'). These lookup tables are initialized at the start of the program, remain until its termination and will be filled up during the execution of the program. Thus, binomial coefficients have to be calculated only once, as well as the probability for a specific parameter set.

	parameter			gene families				
	λ	μ	α	ID	i_M	i_R	i_H	i_F
Loglik1	0.0008	0.0002	5	a	7	7	11	9
				b	8	8	8	5
				c	12	11	12	8
				d	13	14	14	7
				e	11	11	11	15
Loglik2	0.002049347886564641	0.0024562306	150	HBG126803	90	93	101	10

Table 3.3: Parameter sets for two different examples *Loglik1* and *Loglik2* used to compare the running times of the log-likelihood function for the Java program and the ARIBAS program. In the first example *Loglik1* the log-likelihood function was evaluated for five given simulated gene families successively, named a,b,c,d,e. i_X denotes the number of gene copies for the considered species X for each gene family. Species are M = mouse, R = rat, H = human and F = fruit fly. In the second *Loglik2* the gene copy numbers i_X from the gene family HBG126803 are given.

Resulting running times. The resulting running times are shown in table 3.4. If the log-likelihood is computed only once, the advantage of the lookup tables is negligible, but if the function is evaluated many times, the computation is speed up significantly. This is reflected in the results for the first example *Loglik1*. The runtime of the Java program 'without lookup' is about 3.4 times higher than the one for ARIBAS, while the runtime for the Java program 'with lookup' is only about 1.3 times higher. The lookup tables in the Java program create a speedup of about 2.5 in comparison to the Java program without lookup tables. Therefore only the Java program 'with lookup' was used for further comparisons.

In example *Loglik2*, where only one log-likelihood value is calculated and the gene copy numbers are very big, the ARIBAS program is clearly the faster program (5.6 times faster). But also here the effect of lookup tables can be seen, since the factor is relatively smaller as for the computation of the probability distribution with the same parameter set in *Ex3* (15 times faster).

		Java - BigDecimal		
example	tree	without lookup	with lookup	ARIBAS
Loglik1	tMRHF	21,478	8,538	6,404
Loglik2	tMRHF	-	4,660,350	830,840
Estim1	tMRHF	-	32,047,670	16,127,660
Estim2	tMRHCD	-	14,466,550	9,010,830

Table 3.4: *Loglik1-Loglik2*: Running times for the computation of the log-likelihood function for the tree tMRHF measured over 100 repetitions in milliseconds. Details of the used parameters and gene copy numbers of the species are given in the text and in table 3.3. *Estim1-Estim2*: Running times for the whole estimation procedure of duplication rate λ, deletion rate μ and the ancestral gene copy number k for the tree tMRHF (*Estim1*) and tMRHCD (*Estim2*) also in milliseconds. Further details in the text.

Test setting. Finally, the runtime for a whole estimation procedure was measured for two examples. One for the tree tMRHF and another one for the tree tMRHCD. In both cases simulated gene families were used with the following gene copy numbers:

$$\begin{array}{llllll} \textit{Estim1}: & i_{mouse} = 7, & i_{rat} = 7, & i_{human} = 11, & i_{fly} = 9 & \\ \textit{Estim2}: & i_{mouse} = 16, & i_{rat} = 12, & i_{human} = 8, & i_{chimp} = 8, & i_{dog} = 7 \end{array}$$

All input values of the programs, e.g. the start and end values for λ, μ and α, the precision etc., were identical in both programs, but differed for the two examples.

Resulting running times. In the first example *Estim1* the runtime for the Java program was measured to be 32,047,670 ms, which is almost 9 hours, while the C-ARIBAS program needed 16,127,660 ms (4.5 hours), which is only half of the running time of the Java program. The runtime of the C-ARIBAS program in *Estim2* was 9,010,830 ms (2.5 hours), while the estimation of the parameters for the same example gene family with the Java program took 14,466,550 ms (4h). Here the differences in the runtime, with about 1.5, is smaller than in the previous example. In summary the runtime of the C-ARIBAS program is almost every time less than the one of the Java program. Nevertheless there are computations, where both program require nearly the same time to finish and the Java program might be preferred.

3.5 Conclusion

After it turned out that normal `double` variables and the first tested C library PARI were not sufficient for the necessary high accuracy computation, two programs were developed using two different programming languages. The C program uses an external program (ARIBAS) to perform the computation of the log-likelihood function with high accuracy, while the Java program uses the Java classes *BigDecimal* and *BigInteger* for this computation. Both programs can handle the required precision and produce correct results for all analyzed datasets - simulated and real ones.

While the Java program can easily be used on all platforms (Unix, Macintosh, and Windows), the C-ARIBAS program has to be coded specifically for a certain environment. Furthermore it is essential, that the up-to-date ARIBAS version is present and correctly installed. The Java program can easily be enhanced with a graphical user, to improve usability.

In terms of runtime the C-ARIBAS program is the better program, even though the external ARIBAS procedure has to be called for every log-likelihood computation. Although the inclusion of lookup tables in the Java program did not bring the desired reduction in running time, it showed, that it gives a large speedup compared to the Java program without lookup tables. If advantages and disadvantages of the programs are weighted up, both can be applied as the case may be.

CHAPTER
FOUR

Simulation studies

In chapter 2 different estimation methods and the application for modeling gene family evolution were introduced. With the knowledge how to to compute the corresponding estimators with the required accuracy, we can start to validate the quality and performance of these estimation procedures.

For that purpose, we performed several simulation studies. First, we started simulating the evolution of a gene family along a given species tree. We assumed that the gene family has a specific number α of gene copies for the MRCA. Further, we assumed that the number of gene copies evolved according to the BD model described in section 2.1.1 with specific values for the duplication rate λ and the deletion rate μ.

For all studies in this chapter two trees were used. The first tree tMRHF includes the species mouse, rat, human, and fruit fly and goes back in time to 990 mya B.C. The second tree tMRHCD including the mammals mouse, rat, human, chimp, and dog goes back in time to just 93 mya B.C. These trees were already introduced in section 1.1 and the corresponding illustrations can be found in figure 1.3.

4.1 Simulation of gene family data

Starting at the root of a phylogenetic tree for every node a random number is drawn from the BD probability function (from 2.1.2) recursively. This function depends on the rates λ and μ, the number of gene copies of the parent of the node and the time t between the node and its parent. This procedure leads to a specific number of gene copies for each leaf in the tree and represents the current state of the gene family. To test the procedure, simulations for different parameter sets were performed with 100 repetitions per set. The specific parameter combinations are given in table 4.1.

Table 4.2 displayes the mean and standard deviation of the simulated gene copy numbers for the different species. The results are subdivided for the tree tMRHF and the tree tMRHCD. The first column contains the identifier of the parameter set used for the

tree	tMRHF				tMRHCD		
labeling	(a)	(b)	(c)	(d)	(e)	(f)	(g)
α	2	5	10	20	5	10	20
λ	0.0009	0.0008	0.0002	0.0004	0.008	0.002	0.001
μ	0.0005	0.0002	0.0008	0.0006	0.002	0.008	0.009

Table 4.1: Parameter combinations for the simulation of gene family data. α is the ancestral number of genes, λ denotes the duplication rate and μ the deletion rate.

simulation (see table 4.1). Since both trees are clocklike, the time between the MRCA and each leaf is the same. Consequentially the theoretical mean m and the theoretical standard deviation σ are equal for all species in the same tree, because mean and standard deviation depend only on the overall time of the tree, besides α, λ and μ (eq. 2.11, 2.12). Furthermore, the observed mean $\widehat{m}$ and standard deviation $\widehat{\sigma}$ of the gene numbers for each species for the 100 repetitions.

Summary. The observed means are very close to the theoretical values. Actually, for the tree tMRHF the observed means are at average slightly smaller. The mean error $ME = \frac{1}{n}\sum_n (\widehat{m} - m)$ for parameter set (b) is equal to $ME_{(b)} = -0.168$, for (c) and (d) accordingly $ME_{(c)} = -0.398$ and $ME_{(d)} = -0.098$. While the observed means are at average also slightly smaller for parameter set (f) ($ME_{(f)} = -0.018$), the observed means for the two other parameter sets for the tree tMRHCD are at average slightly bigger, $ME_{(e)} = 0.016$ and $ME_{(g)} = 0.228$.

Furthermore a correlation in the gene copy number can be found between the species close together in the trees, e.g. mouse and rat, as well as human and chimp. Between these species the means differ not that much as between other species.

The observed standard deviations are also close to the theoretical ones. For parameter sets where the duplication rate λ is much greater than the deletion rate μ (first set for both trees) the standard deviations tend to differ more from the theoretical value as in the other cases.

To get more information about the resulting simulated gene copy numbers, their distributions at the leaves of the tree were analyzed in more detail.

4.2 Analysis of the simulated data - leaf distribution

The analysis of the frequency distributions of the number of gene copies at the leaves, shortly leaf distributions, was only performed for the tree tMRHF, whereas three different parameter combinations have been chosen:

Ex1: $\alpha = 1$, $\lambda = 0.0008$, $\mu = 0.0002$
Ex2: $\alpha = 5$, $\lambda = 0.0002$, $\mu = 0.0008$
Ex3: $\alpha = 5$, $\lambda = 0.0008$, $\mu = 0.0002$

tMRHF

pc	theoretical		mouse		rat		human		fruit fly	
	m	σ	$\widehat{m}$	$\widehat{\sigma}$	$\widehat{m}$	$\widehat{\sigma}$	$\widehat{m}$	$\widehat{\sigma}$	$\widehat{m}$	$\widehat{\sigma}$
(b)	9.06	3.50	8.79	3.30	8.76	3.22	8.89	3.30	9.13	3.29
(c)	5.52	2.03	5.07	2.02	5.05	1.97	5.06	2.00	5.31	1.95
(d)	16.41	3.84	16.1	3.83	16.31	4.10	16.48	3.93	16.36	4.37

tMRHCD

pc	theoretical		mouse		rat		human		chimp		dog	
	m	σ	$\widehat{m}$	$\widehat{\sigma}$	$\widehat{m}$	$\widehat{\sigma}$	$\widehat{m}$	$\widehat{\sigma}$	$\widehat{m}$	$\widehat{\sigma}$	$\widehat{m}$	$\widehat{\sigma}$
(e)	8.74	3.30	8.96	3.74	8.95	3.59	8.56	2.77	8.47	2.64	8.84	3.06
(f)	5.72	2.02	5.69	2.05	5.68	2.06	5.79	2.00	5.75	2.04	5.6	1.89
(g)	9.5	2.50	9.6	2.27	9.75	2.22	9.54	2.67	9.56	2.72	10.19	2.90

Table 4.2: Statistic for simulated gene families for the trees tMRHF and tMRHCD and several parameter sets. In the first column the identifier for the parameter combination is given, denoted with 'pc'. The corresponding values of the parameters can be found in table 4.1. Specified are the mean m and the standard deviation σ computed theoretical (see 2.1.3, equations 2.11, 2.12) and the observed mean $\widehat{m}$ and standard deviation $\widehat{\sigma}$ of the gene copy numbers for each extant species of the tree.

For each parameter set, we repeated the simulation 10,000 times independently. Furthermore, for every species at the leaves at least one gene copy was required. Hence, to make sure to get at least 10,000 valid gene families all with gene copy numbers ≥ 1, the simulation was repeated 20.000 times for each parameter set. After that, we excluded those gene families where one or more species had zero gene copies and chose the first 10,000 valid gene families for the analysis. In the examples *Ex1* and *Ex2* we sorted out 5,057 and 1,774 of the 20,000 gene families respectively, whereas in the third example *Ex3* all simulated gene copy numbers were at least one. For a detailed description of the gene families excluded we refer to section 4.2.4.

4.2.1 Observed leaf distributions

We started our analysis with the frequencies of the observed gene copy numbers depending on the different species, which are shown in table 4.3 for all three examples.

The frequencies for a defined number of gene copies do not vary much between the four species. But for some gene copy numbers it seems that the frequency for the fly is more different from the others than the frequency among the others.

However, if these frequencies are plotted in a histogram for each species separately, the four histograms look very similar. Figure 4.1 displays all leaf distributions for example *Ex3*. Here it is not possible to detect differences between the distributions. Therefore, in figure 4.2 the leaf distributions for the other examples are only displayed for mouse. This figure show the different shapes the distribution might take. It can look like a exponential distribution, a gamma distribution or nearly a normal distribution.

(a) example *Ex1*

gcn	mouse	rat	human	fly
1	4731	4731	4732	4920
2	2556	2591	2567	2474
3	1294	1265	1309	1251
4	715	697	669	619
5	346	358	360	338
6	190	189	185	204
7	95	90	102	86
8	42	47	38	52
9	20	22	26	36
10	16	20	12	12
11	3	3	6	7
12	4	2	3	6
13	1	2	1	4
14	1	1	2	3
15	0	1	0	0
16	1	0	0	0
17	1	1	0	0

(b) example *Ex2*

gcn	mouse	rat	human	fly
1	1488	1517	1492	1545
2	2801	2739	2788	2764
3	2816	2824	2751	2791
4	1718	1752	1735	1751
5	808	811	857	776
6	281	275	291	291
7	70	62	65	60
8	15	15	16	19
9	3	5	5	3

(c) example *Ex3*

gcn	mouse	rat	human	fly
1	5	4	3	9
2	46	51	46	41
3	167	166	162	149
4	403	397	398	440
5	708	704	760	732
6	1024	1043	1014	994
7	1227	1221	1217	1228
8	1249	1220	1236	1277
9	1140	1176	1154	1160
10	1062	1024	969	974
11	803	844	798	830
12	633	608	673	626
13	426	430	483	490
14	363	345	327	348
15	243	271	249	251
16	174	175	185	161
17	106	94	125	108
18	80	86	79	69
19	53	55	41	45
20	33	27	33	26
21	24	27	20	15
22	15	13	9	12
23	7	7	10	7
24	5	6	4	3
25	1	3	3	2
26	0	0	0	2
27	1	0	0	0
28	0	1	0	0
29	1	1	1	1
30	0	0	1	0
31	0	0	0	0
32	1	1	0	0

Table 4.3: Frequencies of the gene copy numbers (gcn) of the different species of the tree tMRHF for the examples *Ex1*-*Ex3*.

4.2.2 Theoretical leaf distributions

Using the probability function for the BD process (section 2.1.2) the theoretical leaf distributions can be computed. Thereby it was important to keep the exclusion of the gene families in mind, where one or more species had no gene copies. To get a fair comparison

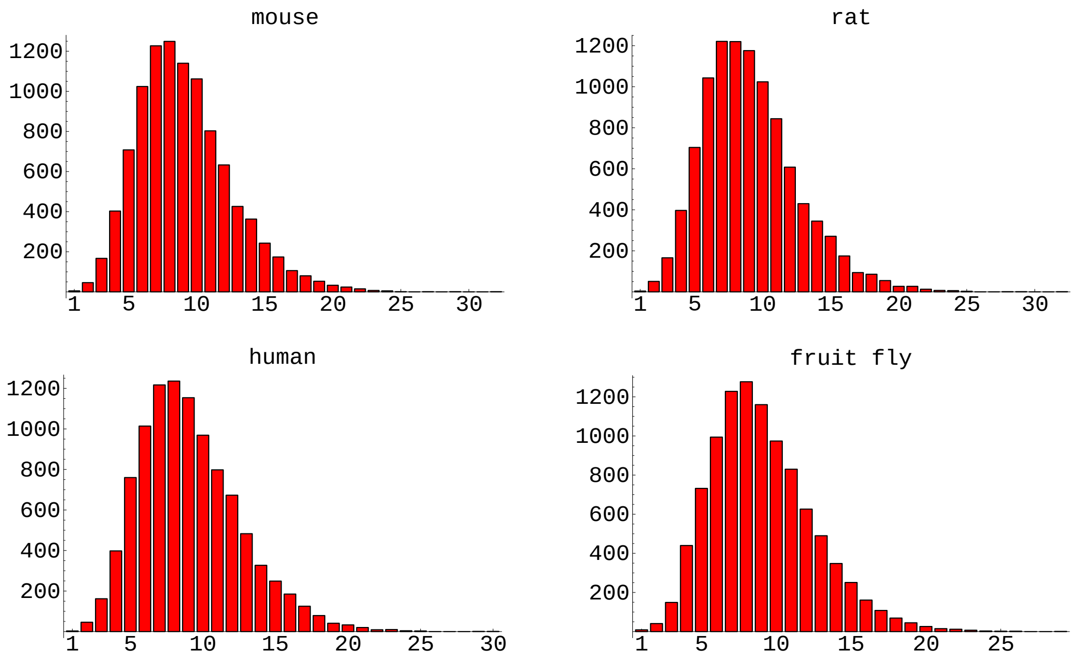


Figure 4.1: Leaf distributions separately for all four species of the tree tMRHF for example *Ex3*. On the x-axis the number of gene copies is given and on the y-axis the corresponding frequency.

between theoretical and observed leaf distributions the probabilities have to be conditional on survival. That means the probabilities have to be divided by $1 - p_0(\alpha, t)$ where $p_0(\alpha, t)$ is the probability for the total loss of gene copies after time t assuming α gene copies for the MRCA. This probability can be computed using the formulas from section 2.1.2 for all three examples:

$$\begin{aligned} &\textit{Ex1}\text{: } p_0(\alpha, t, \lambda, \mu) = p_0(1, 990., 0.0008, 0.0002) = 0.129902 \\ &\textit{Ex2}\text{: } p_0(\alpha, t, \lambda, \mu) = p_0(5, 990., 0.0002, 0.0008) = 0.0378766 \\ &\textit{Ex3}\text{: } p_0(\alpha, t, \lambda, \mu) = p_0(5, 990., 0.0008, 0.0002) = 0.0000369 \end{aligned} \tag{4.1}$$

Figure 4.2 shows the conditional theoretical distribution (solid line) in relation to the observed distribution for mouse (bars). It is easy to see that both match very well.
For further investigations the observed means were compared to the expected means. Again the constraint of surviving had to be taken into account for computing the expected means. The expectation is computed using the following formula:

$$\mathbb{E}(X \mid X > 0) = \frac{\mathbb{E}(X)}{1 - P(X = 0)} \overset{Eq.2.11}{=} \frac{\alpha e^{(\lambda - \mu)t}}{p_0(\alpha, t)} \tag{4.2}$$

Since the total time from the MRCA to every leaf in the tree tMRHF is the same, the expected mean of the leaf distribution of every species is also the same and must be computed only once (see table 4.4). The conditional expected means are consistent with

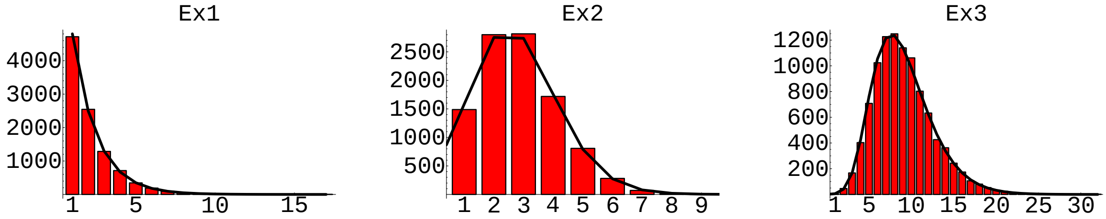


Figure 4.2: Leaf distributions for all examples (*Ex1-Ex3*) solely for mouse: Bars display observed distribution, solid line display theoretical distribution. On the x-axis the number of gene copies is given and on the y-axis the corresponding frequency.

the observed values of the means, which is expected facing the good match of the leaf distributions.

These results support the correctness of the simulation workflow, which means that the simulated gene copy numbers for the species follow the distribution of the corresponding BD process. for given α, λ, and μ. All leaf distributions of the species show the same behavior, since the total time to the MRCA is always the same. To obtain the dependencies between the species due to the tree structure another analysis was needed. Therefore a pairwise comparison was applied.

4.2.3 Pairwise comparison of leaf distributions

If the simulation of gene duplication and deletions is done along a phylogenetic tree, we expect to see the influence of the tree in the resulting datasets. To measure the effect of the tree we need to pairwise compare the leaf distributions for the involved species. Therefore, the number of gene copies of the species were plotted against one another. In doing so, we expected to find similar gene copy numbers for closely related species and relatively distant gene copy numbers for distant species.

Description of the pairwise plot. The plots were done for all three examples, but since they all showed the same result, only the pairwise plots for the third example *Ex3* are described in detail (figure 4.3).

example	mean observed				mean expected	
	mouse	rat	human	fruit fly	$\mathbb{E}(X)$	$\mathbb{E}(X\|X>0)$
Ex1	2.1043	2.1016	2.1040	2.0691	1.8112	2.0816
Ex2	2.8782	2.8788	2.8929	2.8693	2.7606	2.8693
Ex3	9.0970	9.1019	9.1010	9.0535	9.0561	9.0564

Table 4.4: Observed and theoretical means for the examples *Ex1-Ex3*. The observed means are given for each species of the tree tMRHF, the expected mean $\mathbb{E}(X)$ and the expected mean conditional on survival $\mathbb{E}(X|X>0)$ are equal for every species.

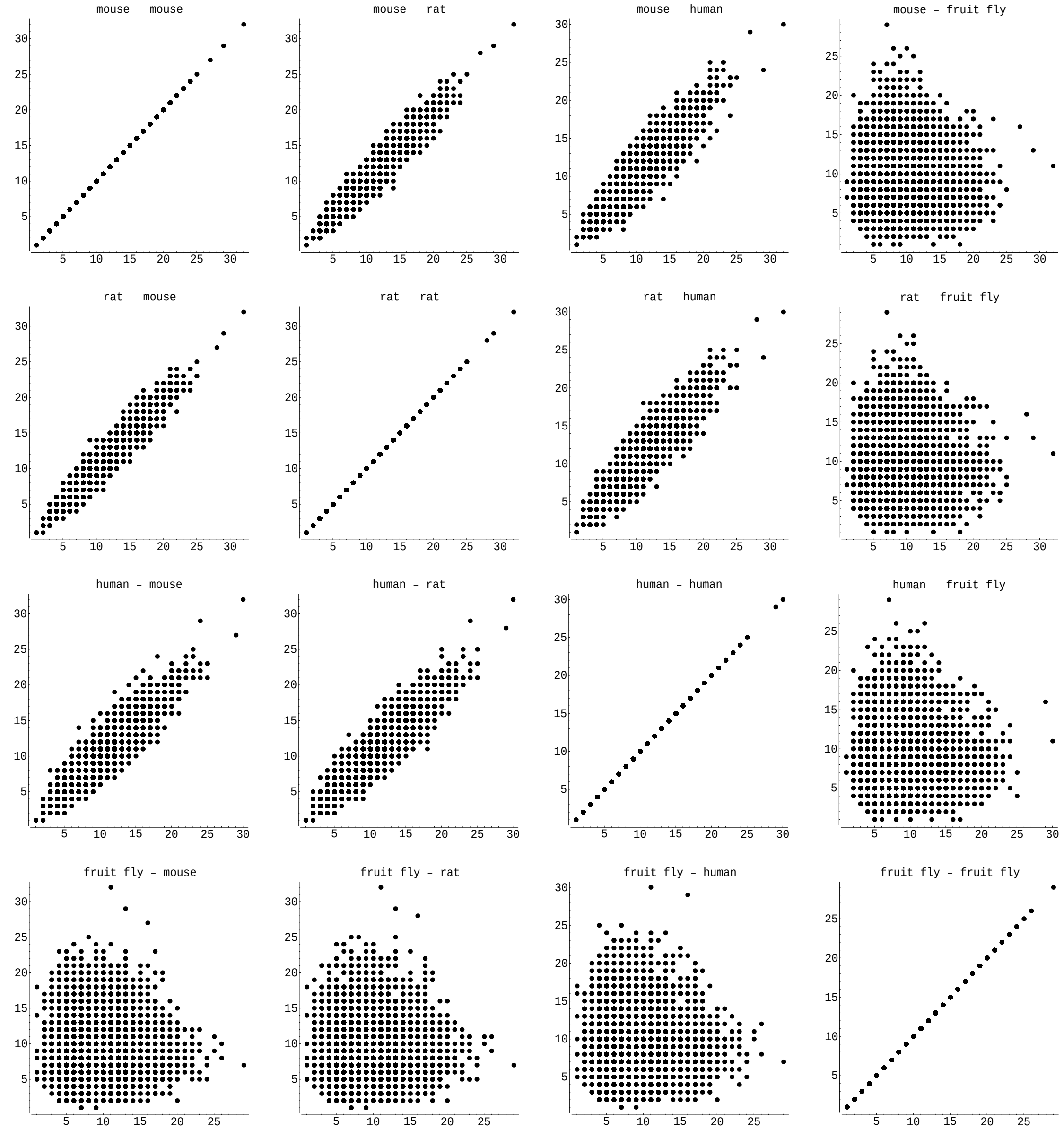


Figure 4.3: Pairwise plots of the gene copy numbers for all species of tree tMRHF for the data of example *Ex3*. Further explanation of the plots can be found in the text.

Each plot is arranged in a matrix and has the labeling 'species1 - species2'. On the x-axis the number of gene copies of species1 is given and on the y-axis the number of gene copies of species2. Thus, in every plot in a row of the matrix, the species on the x-axis is the same and on the other hand in every plot in a column of the matrix the species on the y-axis is the same. That means, if species1 has a gene copies and species2 has b gene copies in the same simulation, a point is drawn at position (a, b) in the plot. Since 10,000 simulations are taken into account, each plot contain 10,000 points. The less the differences between a and b the more closely the point (a, b) is to the bisecting line. The plots in the diagonal of the matrix act as a control. There every species is plotted against itself and therefore all points lie on the bisecting line as expected.

Results. The highest correlation is found between mouse and rat with a correlation coefficient of 0.97. Also highly correlated with almost the same correlation coefficient of 0.93 are mouse and human, as well as rat and human. As expected, for the fruit fly no correlation to another species was found (see table 4.5). The calculated correlation coefficients reflect the relations of the species in the tree as well as the times between them. The fruit fly has no common history with the three other species except the common ancestor. Moreover there was a time of 990 mya from this common ancestor to the recent state for changing in number of gene copies, which is absolutely independent from the other species. Hence, there can be found no correlation to the other species. On the other hand the species mouse, rat and human had 899 mya of common evolution before they first split up. The human branched off and had 91 mya until its current state. So it is obvious that the number of gene copies of the human is similar to mouse and rat due to the relatively short time to change in copy number size. The same holds for mouse and rat.

ϱ	mouse	rat	human	fruit fly
mouse	1	0.9676	0.9296	-0.0083
rat	-	1	0.9293	-0.0112
human	-	-	1	-0.0046
fruit fly	-	-	-	1

Table 4.5: Pearson correlation coefficients for the pairwise plots of the data from example *Ex3*.

Unfortunately, the high correlation between mouse, rat and human makes the estimation of the parameter difficult since only two instead of four data points are given effectively.

4.2.4 Excluded gene families

Let us come back to the excluded gene families where for one or more species all gene copies were deleted in the simulation. An analysis regarding the pattern of those deleted genes was made. In this context, the pattern denotes the occurrence of zero or non-zero gene copies for the considered species and is denoted with 'MRHF' whereas M, R, H and F are either 'x' for non-zero or '0' for zero gene copies. M stands for mouse, R for rat, etc. For four species there are $2^4 = 16$ possible patterns. Since for the third parameter combination *Ex3* no gene families were excluded, only example *Ex1* and *Ex2* were taken into account. Table 4.6 shows the possible patterns together with their occurrence for the two examples.

The most frequent pattern is 'xxxx' is included, which means that all gene copy numbers are non-zero. The second most frequent patterns are 'xxx0' and '000x' in both examples.

Ex1			*Ex2*		
occur	pattern	%	occur	pattern	%
14943	xxxx	74.72	18226	xxxx	91.13
2280	xxx0	11.40	684	xxx0	3.42
2115	000x	10.58	556	000x	2.78
301	0000	1.51	214	xx0x	1.07
119	xx0x	0.60	102	00xx	0.51
84	00xx	0.42	86	x0xx	0.43
60	0xxx	0.30	85	0xxx	0.43
50	x0xx	0.25	31	0000	0.16
18	xx00	0.09	6	xx00	0.03
11	0xx0	0.06	4	0x0x	0.02
9	00x0	0.05	2	0xx0	0.01
8	x0x0	0.04	2	x00x	0.01
1	x00x	0.01	1	x0x0	0.01
1	x000	0.01	1	0x00	0.01
0	0x0x	0	0	x000	0
0	0x00	0	0	00x0	0

Table 4.6: Pattern of gene copy numbers. '0' denote zero and 'x' denote a non-zero number of gene copies. The first place in the pattern represent the number of gene copies of the species mouse, the second for rat, third human and the for last fruit fly. Furthermore the occurrence (occur) of every pattern and the corresponding percentage (%) are given.

The order of these pattern is also the same, although the differences in their occurrence are not substantial. The first of these patterns 'xxx0' represents the extinction of the gene in fruit fly whereas the second pattern '000x' stand for the extinction of a gene in mouse, rat, and human while the gene in the fruit fly survives. Both patterns reflect the history of the species where mouse, rat and human evolved very closely together and the fruit fly almost independently of them. The next frequent patterns 'xx0x' and '00xx' reflect the correlation between mouse and rat, since these two species have often the same number of gene copies. These results show the influence of the tree on the simulated datasets and support the correctness of our simulated program.

Summary. Putting it all together, the study of the leaf distributions for the tree tMRHF confirms the assumed properties. The simulation of gene families with a specific gene copy number for every species works well and the simulated gene copies are distributed according to the BD process. The influence of the tree can be evaluated by looking at pairwise plots of the leaf distribution of the species and correlation analysis. Using such simulated data, we can now evaluate the correctness and performance of our program for the parameter estimation.

4.3 Estimation for individual gene families

As shown in the previous section the procedure to simulate the evolution of gene families according to the BD process works well. Therewith a lot of example data could be generated to test different estimation procedures and their performance.

4.3.1 Example datasets

For seven different parameter combinations we simulated datasets with 100 gene families each as described in section 4.1. For four combinations we used the tree tMRHF (table 4.1 (a)-(d)) and for three combinations the tree tMRHCD (table 4.1 (e)-(g)) was used. The detailed gene copy numbers for the species of all gene families of the simulated datasets can be found in appendix C. Some statistics for the resulting gene copy numbers for the parameter combinations (b)-(g) were already shown in table 4.2.

4.3.2 Method of moments

The first method we tested was the method of moments (MOM) described in section 2.2.2. Using the BD process to model the change in gene family size, it was possible to compute the MOM estimators for the three parameters α, λ and μ in an explicit form (eq. 2.26). These equations make it easy and also quick to estimate the parameters. On the other hand this method has a big disadvantage, since it is not possible to restrict the range of the values for the estimated parameters. But there are several restrictions that have to be considered: The rates λ and μ have to be positive. Further the ancestral number of gene copies α is only allowed to be a non negative integer. Since these restrictions for the parameters cannot be included in the MOM estimation, we found many estimates which are not valid. Especially the occurrence of many complex numbers for all estimates were conspicuous. In these cases it did not make sense to rely on the estimates.

Setup. The MOM was applied for six simulated datasets with the parameter combinations (b)-(g) from table 4.1. Three of these dataset were based on the tree tMRHF and the other three on tree tMRHCD. In appendix C.1 the detailed results for these six datasets are given. For each simulated gene family the ancestral number of gene copies α, the duplication rate λ, and the deletion rate μ are listed there. Additionally, in table 4.7 the mean and standard deviation of all parameters for the different datasets can be found. For this statistic only gene families were taken into account, when the estimates of all parameters are real and non negative. Although it was also not allowed to have a non integer for α, these estimates are still in the statistic, since almost all estimates for α were real. In the fifth column of table 4.7 with notation # the number of gene families is given where the estimates fulfilled the requirements. In total 100 gene families were analyzed in a dataset, so that the numbers for # correspond to percentages.

Valid results. On average valid estimates could only be found for 56% of the simulated gene families. To most of the other datasets complex estimates were assigned (25%). Other parameters turned out to be negative (17%) or computational errors occured (only tree tMRHF: 3%), like a division by zero (marked with 'indeterminate' in appendix C.1).

	tree		tMRHF			tMRHCD		
simu.	labeling		(b)	(c)	(d)	(e)	(f)	(g)
	α		5	10	20	5	10	20
	λ		0.0008	0.0002	0.0004	0.008	0.002	0.001
	μ		0.0002	0.0008	0.0006	0.002	0.008	0.009
estimation	#		61	70	50	41	68	50
	$\widehat{\alpha}$	m	8.0	4.8	15.8	8.2	6.2	9.9
		σ	2.57	1.82	4.04	2.28	2.29	2.05
	$\widehat{\lambda}$	m	0.000150	0.000181	0.000081	0.002018	0.001471	0.001442
		σ	0.000203	0.000254	0.000170	0.002338	0.001439	0.001653
	$\widehat{\mu}$	m	0.000063	0.000146	0.000052	0.001324	0.002829	0.001463
		σ	0.000075	0.000229	0.000076	0.001210	0.005171	0.001472

Table 4.7: Results from the method of moments estimation. The tree is given in the first row and the parameters used for the simulations are given in the next four rows. In the sixth row # the number of gene families of 100 with valid estimates is given. Mean m and standard deviation σ of the estimated number of gene copies of the MRCA $\widehat{\alpha}$, the estimated duplication rates $\widehat{\lambda}$, and deletion rates $\widehat{\mu}$ are shown.

Estimated ancestral number. Comparisons of mean and standard deviation of the estimated parameters $\widehat{\alpha}$, $\widehat{\lambda}$ and $\widehat{\mu}$ to the parameters used for the simulation showed that parameters could not be inferred with acceptable accuracy. The mean of the estimated ancestral number of gene copies $m(\widehat{\alpha})$ for the parameter sets (b) and (e) is 1.6 times higher than the α used for the simulation for both trees. On the other side, the means $m(\widehat{\alpha})$ for parameter set (c) and (g) are about 2 times smaller than the 'true' value. Also for parameter set (f) it is 1.6 times smaller and for the remaining parameter set (d) $m(\widehat{\alpha})$ is 1.3 times smaller than the value of the simulation.
If we compare the values for the estimated means $m(\widehat{\alpha})$ to the theoretical means from table 4.2, it stands out that they are very similar. The theoretical means from table 4.2 represent the expected number of gene copies for the recent species dependent on the tree, α, λ, and μ. That means, that the MOM for this purpose does not find the ancestral gene copy number α used in the simulation, but the expected mean for the gene copy number at the leaves under the BD process.
The estimate for α depends on the difference $\lambda - \mu$. If $\lambda - \mu > 0$ it happens that α will be overestimated, while $\lambda - \mu < 0$ lead to an underestimation of α. Besides, the distinction between the 'true' α and the mean of the estimates $m(\widehat{\alpha})$ depend on the absolute value of the difference of λ and μ. With increasing values for $|\lambda - \mu|$ also the relative distinction between α and $m(\widehat{\alpha})$ grows.

Estimated rates. The means of the estimated rates $\widehat{\lambda}$ and $\widehat{\mu}$ are in most cases almost equal and do not reflect the absolute values of the rates λ and μ used in the simulation. For half of the datasets even the relative proportions of the rates could not be found. For

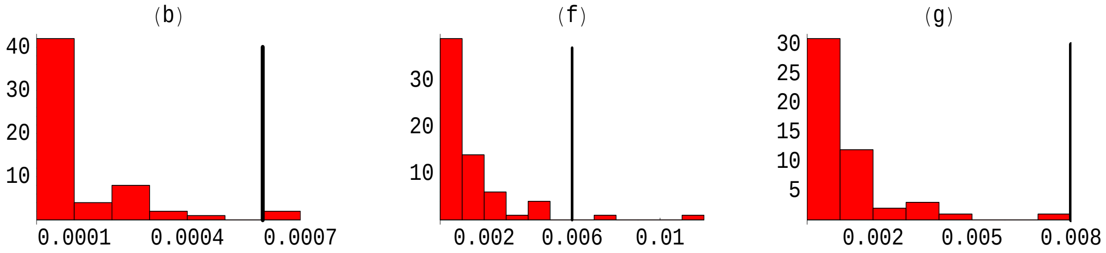


Figure 4.4: Results from the method of moments estimation. Shown are the histograms for the absolute differences between estimated duplication rates $\widehat{\lambda}$ and the deletion rates $\widehat{\mu}$ for selected datasets. The vertical line in the plots marks the 'true' value of the difference. The labeling (b),(f),(g) corresponds to that used in table 4.1.

example in the first parameter set (b) the duplication rate λ was bigger than the deletion rate μ in the simulation, but for the estimates it is the other way round. That is to say, the mean of $\widehat{\lambda}$ is smaller than the mean of $\widehat{\mu}$ in this case.

In table 4.8 the mean of the absolute differences between the estimated duplication rates and the deletion rates $m(|\widehat{\lambda} - \widehat{\mu}|)$ is given for each parameter combination. For further comparison, the 'true' absolute differences of λ and μ are given there. With the exception of (d), the mean $m(|\widehat{\lambda} - \widehat{\mu}|)$ for all parameter combination is much smaller than the 'true' difference $|\lambda - \mu|$. If we consider the distributions of the absolute difference of the estimated rates, we find no or very small differences as the most frequent ones. For three parameter sets this distribution is shown in figure 4.4. It is remarkable, that in all of these distributions the 'true' value, indicated by the vertical line, is on the right hand side of the distribution.

Summary. Summing up, it can be said that the estimates for all three parameters α, λ, and μ are strongly biased. Further a high failure rate of about 44% was found for the MOM. As already realized in section 2.2.2 the estimator for α (eq. 2.26) is independent of the time. This fact and furthermore, that it is not possible to take the species tree into account are strong limitations of MOM, as this information cannot be used. Especially the information about the topology and the branch lengths of the species tree is very essential and for many species available. Thus, the MOM seems to be not suitable for the estimation of the parameters α, λ, and μ in this content.

labeling	(b)	(c)	(d)	(e)	(f)	(g)
$\lvert\lambda - \mu\rvert$	0.0006	0.0006	0.0002	0.006	0.006	0.008
$m(\lvert\widehat{\lambda} - \widehat{\mu}\rvert)$	0.00011	0.00012	0.00054	0.0015	0.0014	0.0011

Table 4.8: Results from the method of moments estimation. Mean of the absolute differences between the estimated duplication rates $\widehat{\lambda}$ and the deletion rates $\widehat{\mu}$ for the analyzed dataset. Labeling see table 4.1.

4.3.3 Maximum likelihood estimation

The maximum likelihood (ML) method is another estimation method used to determine model parameters. It was introduced in subsection 2.2.3. The likelihood function $L_{\mathcal{T}}(\alpha, \lambda, \mu)$ of the BD process for a specific species tree $\mathcal{T}$ is given in eq. 2.34. Again both trees tMRHF and tMRHCD are used for the simulation studies. We further used the same data as in the MOM estimation before ((b)-(g) from table 4.1). Additionally a dataset for parameter combination (a) from table 4.1 for the tree tMRHF was used.

With the ML method it is possible to overcome the difficulties with the restriction of the parameter space, by using first a constrained optimization method for λ and μ and second a discrete optimization method for the parameter α. The estimation of the parameters by maximizing the likelihood function $L_{\mathcal{T}}(\alpha, \lambda, \mu)$ was implemented in our computer software (appendix A). Using this software, it was possible to estimate the parameters for every single gene family according to the required restrictions.

	tree		tMRHF				tMRHCD		
simu.	labeling		(a)	(b)	(c)	(d)	(e)	(f)	(g)
	α		2	5	10	20	5	10	20
	λ		0.0009	0.0008	0.0002	0.0004	0.008	0.002	0.001
	μ		0.0005	0.0002	0.0008	0.0006	0.002	0.008	0.009
estimation	#		81	85	83	71	91	89	90
	$\widehat{\alpha}$	m	3.8	10.9	10.0	19.3	10.8	10.2	18.6
		σ	3.75	7.28	9.75	9.16	8.90	8.01	10.45
		CI_l	1	3	2	7	2	3	3
		CI_u	15	32	39	38	37	27	45
	$\widehat{\lambda}$	m	0.000354	0.000228	0.000143	0.000191	0.003763	0.001541	0.001940
		σ	0.000493	0.000343	0.000245	0.000285	0.004414	0.003046	0.003469
		CI_u	0.001291	0.000901	0.000672	0.000771	0.011973	0.007207	0.010699
		zero	44 (54%)	42 (49%)	51 (61%)	32 (45%)	35 (38%)	55 (62%)	55 (61%)
	$\widehat{\mu}$	m	0.000306	0.000272	0.000468	0.000260	0.003342	0.005699	0.005929
		σ	0.000592	0.000456	0.000671	0.000361	0.004543	0.006392	0.006403
		CI_u	0.001638	0.001234	0.002096	0.000909	0.013886	0.018282	0.012802
		zero	42 (52%)	41 (48%)	32 (39%)	31 (44%)	45 (49%)	20 (22%)	24 (27%)
	$\widehat{\lambda} > 0$ & $\widehat{\mu} > 0$		5 (6%)	5 (6%)	5 (6%)	9 (13%)	10 (11%)	13 (15%)	11 (12%)

Table 4.9: Results from the maximum likelihood estimation. The tree used is given in the first row and the parameters used for the simulations are given in the next four rows. In the sixth row # the number of gene families of 100 with valid estimates is given. Mean m and standard deviation σ of the estimated number of gene copies of the MRCA $\widehat{\alpha}$, the estimated duplication rates $\widehat{\lambda}$ and deletion rates $\widehat{\mu}$ are summarized. Furthermore the two-sided 95% confidence interval CI is given for estimates of α. Thereby CI_l denotes the lower boundary of the interval and CI_u the upper boundary. In contrast, for the estimated duplication rates λ and the estimated deletion rates μ the upper boundary CI_u of the 95% one-sided confidence interval is given. Furthermore the number of estimates equal to zero (1.0E-15) for both rates is given (denoted with 'zero'), as well as the corresponding percentages. In the last row the number of gene families is given, where both estimated rates were greater than zero (1.0E-15).

Setup. Since the rates λ and μ could not obtain values outside the required parameter space, the method was classified as 'failed' for a certain gene family, if the estimation of the ancestral number of gene copies α failed. In the discrete optimization method used to estimate α, an interval must be assigned, inside which the value of α is presumed. The lower bound of this interval is in most cases one and cannot be smaller. If this boundary was reached by the method as the optimum, it was accepted. On the other hand, the upper boundary can be chosen arbitrarily and should always be relatively big compared to the gene copy numbers of the recent species. Thus, if the upper boundary was reached the estimate was not accepted. For example, if the numbers of gene copies for the species of a gene family have values of about 2-4 and the upper bound is chosen to be 50, an estimate for $\alpha > 50$ is presumed to be unrealistic. Furthermore, we could show that in these cases the value of the likelihood function did not change much even for higher values for α. Which means, that in these cases many almost equal solutions exist.

Valid results. The detailed results can be found again in the appendix C.2. The high variation of the BD process is reflected in these results and unfortunately there were also gene families were the method failed. In table 4.9 the results are summarized. The mean, standard deviation and confidence interval for each estimate are given, conditioned by the different datasets and used trees. The fifth column, denoted with #, contains the number of gene families, where the entire estimation was successful. That was on average 85 of 100 gene families and results in a mean failure rate of 15%.

Estimated ancestral number. If the means of the estimates of α are compared with the values used in the simulations (table 4.9), significant differences could be found for the parameter combinations (a),(b) and (e). The mean of the estimates is about 2 times higher than the 'true' value of the parameter. But for the remaining parameter sets the mean of the estimates of α are very close to the 'true' values. The standard deviations σ for the estimates of α are high, which can also be seen in the wide 95% confidence interval. For (a)-(c),(e) and (f) the standard deviation is almost as big as the mean $m(\widehat{\alpha})$ itself, while for (d) and (g) it is only half as big as the mean $m(\widehat{\alpha})$.

Artifact of the method. It is obvious that the estimates for the duplication rate λ and the deletion rate μ depend on the estimate $\widehat{\alpha}$. If the estimated $\widehat{\alpha}$ is relatively high compared to the numbers of gene copies of the recent species, the corresponding duplication rate $\widehat{\lambda}$ is often found to be near to zero. That happened on average for 47% of the valid gene families. Note that 1.0E-15 was the smallest possible value fixed in the program. That is because a lot of gene copies had to get lost during evolution and further duplications, which would increase the gene number again, are not probable. In this case the BD process becomes a pure death process. Otherwise, if $\widehat{\alpha}$ is very small compared to the given gene copy numbers at the leaves, the estimated deletion rate $\widehat{\mu}$ tends to be zero and the BD process becomes a pure birth process. A deletion rate μ equal to zero was estimated for 39% of the valid gene families. Depending on the tree, only for a few gene families estimates with both rates greater than zero (1.0E-15) were found. That is for the tree tMRHF on average 8% and for the tree tMRHCD 13% of all valid gene families. Thus, for most simulated gene families one of these both cases occurred, which is an artifact of the method.

Comparison to true values for rates. If the means of the estimates $\widehat{\lambda}$ and $\widehat{\mu}$ are compared to the 'true' values, it stands out that in general the rates are underestimated. Only for three cases the estimated rate is bigger than the 'true' value. That is for the deletion rate $\widehat{\mu}$ of parameter combination (b) and (e), and for the duplication rate $\widehat{\lambda}$ of combination (g). The underestimation of the rates is very likely to be due to the occurrence of multiple duplication and deletion events. If the duplication of one gene is followed by a loss of one gene, which means the number of gene copies do not change in total, it is hardly possible to track these events.

Ratio of the estimated rates. For the parameter combinations (a),(b), and (e) the rates are not only underestimated but also wrong in their proportion. That is understandable given the overestimated ancestral number $\widehat{\alpha}$, but obviously do not lead to the 'true' values of λ and μ. In these cases the means of the estimated rates $\widehat{\lambda}$ and $\widehat{\mu}$ are almost equal.
In the other cases the ratios of the rates to each other are nearly the same as for the 'true' values. For example, the ratio λ/μ in the simulation of parameter combination (d) is 0.67, whereas the ratio of the means $m(\widehat{\lambda})/m(\widehat{\mu})$ is 0.73. For combination (f) it is even better, since the ratio in the simulation λ/μ is 0.25 and the ration of the estimated means $m(\widehat{\lambda})/m(\widehat{\mu})$ is equal to 0.27.

Confidence intervals. For the rates a one-sided 95% confidence interval was calculated. That was done because the most frequent value was 1.0E-15 and that results in a skew distribution for λ and μ. The confidence intervals for the rates are wide, as it was already observed for the ancestral gene number α. That corresponds to the high standard deviation, which is bigger than the mean itself for both rates and all parameter combinations.

Histograms. To get an impression of the distribution of the estimated parameters $\widehat{\alpha}$, $\widehat{\lambda}$ and $\widehat{\mu}$, histograms for the estimates are drawn for the tree tMRHCD with the parameter combination (f) used in the simulation: $\alpha = 10$, $\lambda = 0.0002$ and $\mu = 0.0008$ (see figure 4.5). The histograms for $\widehat{\lambda}$ and $\widehat{\mu}$ (figure 4.5 (f2),(f3) show the distribution of the estimated values. The value 1.0E-15 occurred 55 times for $\widehat{\lambda}$ and 20 times for $\widehat{\mu}$ and thus is the most frequent value for both rates, as already mentioned. Here also the underestimation of the duplication and deletion rate can be seen, since most values are located left of the 'true'

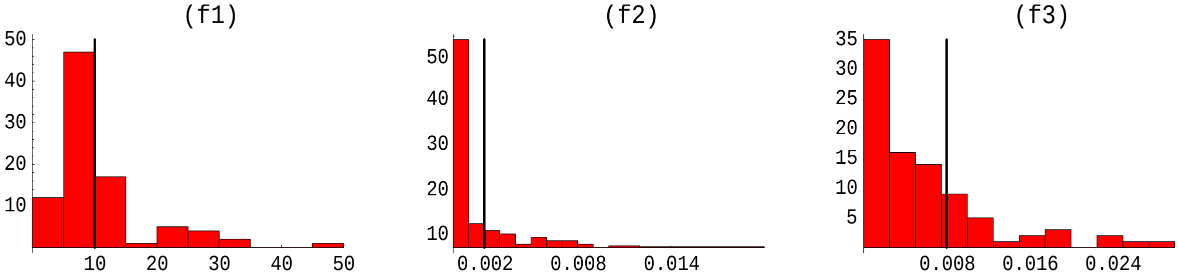


Figure 4.5: Results from the maximum likelihood estimation. Shown are the histograms for the estimated parameters $\widehat{\alpha}$ (histogram f1), $\widehat{\lambda}$ (histogram f2), and $\widehat{\mu}$ (histogram f3) for the tree tMRHCD and the parameters $\alpha = 10$, $\lambda = 0.002$ and $\mu = 0.008$ used in the simulation of the dataset. The vertical line in the plots marks the 'true' value of the parameter.

value (indicated with the vertical line). In contrast, the distribution of the estimates of α (histogram f1) is more or less as expected with the most frequent value being 10.

Summary. Altogether the simulation studies showed that the estimates are not very stable. If the duplication rate λ was much bigger than the deletion rate μ in the simulations, all estimates $\widehat{\alpha}$, $\widehat{\lambda}$ and $\widehat{\mu}$ were biased and the ratio between the rates near to one, regardless of the 'true' ratio from the simulation. For the other case, where μ was bigger than λ, the estimates of α seem to be unbiased. Nevertheless the estimates of the rates $\widehat{\lambda}$ and $\widehat{\mu}$ were again biased, but appear to be in the right ratio to each other.

4.3.4 Discussion

Since the method of moments had a very high failure rate (about 44%), the parameter space could not be defined and MOM estimates did not reflect the values used for the simulation, this method is considered to be insufficient for the analyses.

The problem of estimates being outside the parameter space never arises in the method of ML and furthermore the information of the species tree can be used. The quality of the estimates depend on the ratio of the 'true' duplication rate and 'true' deletion rate λ/μ. If $\lambda/\mu > 1$ the estimates for α, λ, and μ were biased. For $\lambda/\mu < 1$, the ML estimates for α were found to be unbiased, while the rates λ and μ were underestimated in the same extent. Two things have to be pointed out. One is the restriction of the parameter space. The ancestral number of gene copies has to be equal or bigger than 1 ($\alpha \geq 1$). If the value for α in the simulation is close to one, like e.g. $\alpha = 2$ or $\alpha = 5$, the resulting distribution of the estimates is very near to the boundary $\alpha = 1$ and might be influenced. If estimates can only receive values from one side of the distribution, properties, like the mean, might not be determined precisely enough. Another important point is the number of leaves in the species tree. The usage of very small species trees (4-5 leaves) does probably not provide sufficient information for a stable estimation of the parameters. For testing, much larger trees should also be analyzed. That is problematic for two reasons: First, the running time of the program for large trees is too long to perform a sufficient number of simulation studies and second there is hardly enough real data, because gene family data from fully sequenced species is rare. Results should always be seen critical, when using the ML estimation method.

4.4 Estimation for sets of gene families

There is also another possibility to increase the amount of information for the parameter estimation. Since the number of data points (species) in a single gene family is limited, the idea arose to consider not only single gene families, but multiple gene families in a set together. Then the estimation would not be done independently for every single gene family, but for the set of families. That means that there is one ancestral gene number α, one duplication rate λ and one deletion rate μ for the entire set. In doing so, it might be possible to overcome artifacts which were found using single gene families. Since it was completely unclear how many gene families should be grouped together, three studies concerning sets have been carried out. All of them were performed for the tree tMRHF.

4.4.1 Generation of datasets

To generate sets of gene families, first of all single gene families were simulated again, using different parameter combinations (table 4.10). For study A and study B in total 100 gene families were simulated, each. Then the 100 gene families for study A were evenly divided into 10 sets each with 10 gene families and the same for the data for study B. For study C 105 gene families with $\alpha = 1$ and 100 gene families with $\alpha = 5$ were simulated. The 105 gene families with $\alpha = 1$ were divided into sets with 5, 10, 15, 25, and 50 families. The remaining 100 gene families with $\alpha = 5$ formed one single set.

labeling	α	λ	μ	n
study A	1	0.0008	0.0002	10
study B	5	0.0002	0.0008	10
study C	1/5	0.0008	0.0002	varying

Table 4.10: Parameter combinations for the simulation of gene families. The gene families were combined into sets each with n gene families. In study C both ancestral gene numbers $\alpha = 1$ and $\alpha = 5$ were used.

4.4.2 Maximum likelihood estimation

To get the maximum likelihood estimator, the likelihood functions for all gene families in a set had to be multiplied. That is equal to the summation of the log-likelihood functions and is implemented in our software. Thereby the likelihood function for each gene family $L_{\mathcal{T}}(\alpha, \lambda, \mu)$ is the same as in simulation studies before for single gene families (see eq. 2.34). For the tree tMRHF the ML estimator for a set $\mathcal{S}$ of gene families L^{S}_{tMRHF} is given by

$$\begin{array}{rclcl} L^{\mathcal{S}}_{tMRHF} & = & L^{\mathcal{S}}_{tMRHF}(\alpha,\lambda,\mu) & = & \prod\limits_{\forall\,\mathrm{gf}\,\in\,\mathcal{S}} L_{tMRHF}(\alpha,\lambda,\mu) \\ & & log\,L^{\mathcal{S}}_{tMRHF}(\alpha,\lambda,\mu) & = & \sum\limits_{\forall\,\mathrm{gf}\,\in\,\mathcal{S}} log\,L_{tMRHF}(\alpha,\lambda,\mu) \end{array} \tag{4.3}$$

whereas gf denotes 'gene family' and the specific likelihood function $L_{tMRHF}(\alpha,\lambda,\mu)$ is used from eq. 2.35. Using this ML estimator the number of ancestral genes α, the duplication rate λ and the deletion rate μ are computed as global parameters for all gene families of a set. The estimation of the parameters by maximizing the likelihood function L^{S}_{tMRHF} is also implemented in our program (appendix A).

The results of the studies are given in the tables 4.11, 4.12, and 4.13, respectively. In the first two studies the performance of the estimation was analyzed when using relatively small sets of gene families. Whereas in the last study the influence of the number of gene families in a set was considered.

Study A. Thus, the individual estimates for each set of gene families for the first study, as well as the mean m, and the standard deviation σ of these estimates are shown in table 4.11. Study A represents a special case, since the ancestral gene number adopts the lowest value allowed: $\alpha = 1$. It stands out, that all estimated deletion rates μ, except one (no. 8), are almost zero. It should be noted that in the program the smallest value for λ and μ was 0.0000000001 = 1.0E-10 (at the time of these analyzes). That leads to the fact, that for almost all sets a pure birth process can be assumed. The estimates of the duplication rate $\widehat{\lambda}$ on the other hand, are quite good, but again underestimated. The differences bet-

no.	n	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	$logL$	running time
1	10	1	0.0010185185	0.0000000001	−22.9794087479	8.46 h
2	10	1	0.0007953473	0.0000000001	−19.8088607570	6.65 h
3	10	1	0.0007242607	0.0000000001	−16.1317619701	5.94 h
4	10	1	0.0007073825	0.0000000001	−15.9552417210	6.38 h
5	10	1	0.0007390006	0.0000000001	−16.1246865872	5.58 h
6	10	1	0.0004185029	0.0000000001	−10.1796965986	1.91 h
7	10	1	0.0007625233	0.0000000001	−16.3159035093	3.66 h
8	10	2	0.0004220521	0.0003246660	−21.2278997613	3.18 h
9	10	1	0.0006041346	0.0000000001	−13.4921919960	2.36 h
10	10	1	0.0005230294	0.0000000001	−13.6745058121	2.13 h
m		1.1	0.0006714750	0.0000324667	−16.5890157460	4.63 h
σ		0.3	0.0001841195	0.0001029563		

Table 4.11: Results for study A (parameters in table 4.10). The no. indicates a simple numbering of the sets of gene families and n is the number of gene families in the set. In this study all sets contained the same number of gene families $n = 10$. Furthermore the estimates for α, λ and μ are given as well as the corresponding log-likelihood value $logL = log(L_{tMRHF}(\widehat{\alpha},\widehat{\lambda},\widehat{\mu}))$ and the running time in hours h.

no.	n	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	$logL$	running time
1	10	5	0.0002763168	0.0006118859	−23.8267893650	4.83 h
2	10	8	0.0000000001	0.0011618871	−21.0210290789	2.99 h
3	10	3	0.0002758491	0.0003679453	−17.3134722108	3.39 h
4	10	6	0.0001755582	0.0007468157	−23.8414533419	4.11 h
5	10	14	0.0000000001	0.0017118055	−22.8967425198	3.83 h
6	10	3	0.0003158235	0.0002889580	−19.6509535006	4.08 h
7	10	7	0.0000000001	0.0009678228	−20.9468759162	3.15 h
8	10	16	0.0000000001	0.0014535706	−26.9358540165	4.97 h
9	10	2	0.0003395219	0.0000742370	−16.5999418638	2.95 h
10	10	7	0.0001176793	0.0010988683	−22.4059140418	4.14 h
m		7.1	0.0001500750	0.0008483800	−21.5439025855	3.84 h
σ		4.6	0.0001442221	0.0005266878		

Table 4.12: Results for study B (parameters in table 4.10). The no. indicates a simple numbering of the sets of gene families and n is the number of gene families in the set. In this study all sets contained the same number of gene families $n = 10$. Furthermore the estimates for α, λ and μ are given as well as the corresponding log-likelihood value $logL = log(L_{tMRHF}(\widehat{\alpha}, \widehat{\lambda}, \widehat{\mu}))$ and the running time in hours h.

ween the mean of the estimates $\widehat{\lambda} = 0.00067$ and the original value used in the simulation $\lambda = 0.0008$ is not as big as in previous studies on single gene families.
In comparison to the study on single gene families, see for example parameter combination (a) in table 4.9, an improvement in the quality of the estimates is noticeable. In cases where λ was much bigger than μ the estimation of the rates for single gene families was poorly, whereas in the case of sets, the 'true' ratio of the rates could be found in the estimates.

The estimation of the ancestral gene number α was also much better than in the single gene family case. All values estimated were equal to the 'true' value, except one (no. 8). Furthermore the tendency to get estimates twice as big as the 'true' one, which was found for single gene families, could not be detected.

One explanation for the estimation of μ to be near zero could be the specialty of this parameter combination. Since no node in the tree is allowed to have zero gene copies it is improbable to get a high deletion rate. The gene families generated for this study are also restricted to have at least one gene copy for each species. So there might be a little bias to smaller deletion rates due to this restriction. If the estimated deletion rate $\widehat{\mu}$ rarely adopts values greater 1.0E-10, the duplication rate λ does not have to be as big as the original one to explain the data.

Study B. In the second study also the performance of the strategy for rather small sets of gene families was analyzed. Therefore another parameter combination was used, where the duplication rate λ was much smaller than the deletion rate μ. In this study

		simulation			estimation			
no.	n	α	λ	μ	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	$logL$
1	5	1	0.0008	0.0002	1	0.0003862527	0.0000000001	−5.2140704206
2	10	1	0.0008	0.0002	1	0.0010185197	0.0000000001	−22.9791782412
3	15	1	0.0008	0.0002	1	0.0009014272	0.0000000001	−28.3758601357
4	25	1	0.0008	0.0002	1	0.0006572090	0.0000000001	−38.2592765049
5	50	1	0.0008	0.0002	1	0.0007556500	0.0000000001	−87.8798052650
6	100	5	0.0008	0.0002	5	0.0008251194	0.0001975704	−355.188678731

Table 4.13: Results for study C (parameters in table 4.10). The no. indicates a simple numbering of the sets of gene families and n is the number of gene families in the set. In this study these numbers varied from $n = 5$ to $n = 100$. For each set the parameters used in the simulation are given, the estimates $\widehat{\alpha}$, $\widehat{\lambda}$ and $\widehat{\mu}$ and the corresponding log-likelihood value $logL = log(L_{tMRHF}(\widehat{\alpha}, \widehat{\lambda}, \widehat{\mu}))$.

the variation in the estimates was higher than in the previous study, due to the higher ancestral number of gene copies $\alpha = 5$.

In turned out, that the number of sets where one of the rates are estimated to be zero, became less frequent. Whereas for 92% of the gene families in single gene family studies one rate became zero, it happened only for 40% of the gene families in study B. Furthermore the absolute values of the rates could be estimated very well in contrast to previous single gene family studies. Actually the mean of the estimates of μ is slightly bigger than the 'true' value and so the underestimation of the rates is not a problem any more. Although the estimates of the ancestral gene number α could be more accurate, the estimates for α are more stable, than in the single gene family estimation. That can be seen in the smaller standard deviation of $\widehat{\alpha}$.

Study C. In the last study (table 4.13) the case of one rate (here the deletion rate) being close to zero occurred again for all cases where the value of α was 1 in the simulation (no. 1-5). So, this problem can be associated with the ancestral gene number being 1 in the simulation. For these cases the estimated deletion rate was always 1.0E-10. On the other hand, it seems that the estimates of the duplication rate λ get better with increasing number of gene families in a set. Furthermore, the estimation of α is for every set excellent, since in all cases the estimate is equal to the 'true' value.

In the last set no. 6, where α used to be 5 in the simulation, the estimates are very good. The estimated ancestral gene number $\widehat{\alpha}$ is equal to the 'true' value and both rates are found with high accuracy, no appreciable under- or overestimation.

4.4.3 Discussion

In contrast to single gene families, the usage of sets of gene families seems to give more precise and more stable estimates. For only 10 gene families in a set, we found a noticeable improvement in the quality of the estimates. Thereby the sets, where the 'true' ancestral number of genes was equal to one, are special cases. There it seems very hard to determine the 'true' deletion rate.

If the datasets were simulated with an ancestral gene number bigger than one ($\alpha > 1$), we found much less estimates for the rates being almost zero (study B, study C no. 6). Thus, this artifact of the method can be associated with the amount of information the data provide. Study C showed, that the estimates got better with increasing number of gene families in a set. For 100 gene families the 'true' values were found with high accuracy. Furthermore, it should be noticed that for none of the sets the method failed.

Nevertheless the application of this approach to real data is complicated, since it is not clear how to divide a real dataset with hundreds of gene families into sets of which size. A possible criterion for the division could be a similar pattern of gene copy numbers of the species between gene families. That means, that gene families are grouped together if the proportions of the gene copy numbers of the species are nearly equal. According to this, it might be assumed that these gene families had undergone the same evolution history. But this is a very theoretical approach and it will be hard to explain this kind of division with regards to the biological history. Other criteria could be the division according to the position of the genes in the genome or the division according to the biological function of the genes. But this information is often hard to obtain and can also be contradicting for example, if the members of a gene family are located on totally different positions in the genome. Furthermore the functions of many genes are not known, uncertain or multiple. Then it will also be difficult to find a reasonable division into sets.

4.5 Conclusion

We have shown that the simulation of the changes in gene family size according to the specified BD process using our program works well. Such simulated gene family data was used to analyze the proposed estimation methods.

The first method tested was the method of moments. The estimates of this method were biased and inconsistent. Furthermore a lot of non-valid estimates were found with the MOM. That can be ascribed to the fact, that no parameter restrictions can be included in the method. Besides, not all available information, like the shape of the species tree, can be used. Due to the poor performance of the MOM, we decided that it is not suitable for our purpose.

The maximum likelihood method was also tested for simulated single gene family data. Compared to the MOM, the ML estimation led to better results. In total the parameters were underestimated. The quality of the estimates depend on the proportion of the duplication rate λ and the deletion rate μ. If the gene family data was simulated with λ greater than μ, the estimation of the parameters was poor. In these cases the estimated ancestral gene number α was about twice as high as the 'true' value and both rates were estimated to be almost equal. For the other case, λ smaller than μ, the estimates were overall quite good. Again both rates were underestimated, but in the same extent. An artifact of the method is the estimation of one rate being almost zero for many simulated gene families (on average 90%).

Studies on the usage of sets of gene families instead of single gene families showed an improvement in the estimation. The estimates were found with higher accuracy and higher consistency. But a meaningful application of this approach to real data is complicated. Since our main goal is the estimation of the duplication and the deletion rate for individual gene families, this approach was not tested on real data.

CHAPTER

FIVE

Application to biological data

In this chapter the application of the maximum likelihood method for single gene families to biological data is described. For the two taxon sets MRHF and MRHCD used in the simulation studies, we extracted nucleotide and protein data from three data sources. In this chapter, we will discuss the results for the estimated parameters, especially in comparison to already published studies. The assembly of the data and problems in the data quality are described.

5.1 HOGENOM database

HOGENOM is a database of HOmologous sequences from complete GENOMes (`http://pbil.univ-lyon1.fr/databases/hogenom.html`). It contains protein sequences from the European Bioinformatics Institute (EBI) (`http://www.ebi.ac.uk/proteome/`) and the corresponding nucleotide sequences from EMBL Nucleotid Database (`http://www.ebi.ac.uk/embl/index.html`).
In the current version (release 03, October 2005) fully sequenced genomes from 263 organisms are included, 25 of them belong to the eukaryotes. The total amount of protein sequences is 950,216 at present. Homologous protein sequences are divided into families, and for each family the multiple sequence alignment and the phylogentic tree is provided. Currently there are 262,877 families in the database.
A distinction is made between the HOGENPROT database, which contains the protein sequences and the HOGENNUCL database containing the nucleotide sequences.

5.1.1 Data assembly

Using the Cross-Taxa system of the HOGENOM database, it is possible to search for families which include a specific set of species. Therefor, one of the two databases, HOGENPROT or HOGENNUCL, must be chosen (`http://pbil.univ-lyon1.fr/search/cross_fam.php`).

A total of 1661 families containing at least one sequence from mouse (*Mus musculus*), rat (*Rattus norvegicus*), human (*Homo sapiens*), and fruit fly (*Drosophila melanogaster*) were found in the HOGENNUCL database. The information of these families, including the family ID, the number of sequences for each species, and the description of the family, was extracted using a php-script.

In 9 of the 1661 families exactly one sequence for each species was found. In 991 families the number of sequences was between 1 and 9 for all species. These families will be called one-digit families. For 628 families (double-digit families) at least one species was represented by 10 to 99 sequences, and for 33 families the number of sequences for at least one species exceeds 100. These families are called three-digit families. No family with more than 999 sequences for one species was observed.

Because only the number of sequences of the considered species are used for the estimation of the ancestral number of gene family members α, the duplication rate λ, and the deletion rate μ, the dataset was scanned for equal patterns. Thereby a pattern 'M-R-H-F' represent the number of sequences for mouse(M), rat(R), human(H), and fly (F) in this order. If two families are represented by the same number of sequences for each species, the families have the same pattern. In the estimation only families with different patterns had to be taken into account. In the 1661 families 1150 different patterns were found.

The same procedure was repeated for the HOGENPROT database and 1688 families were extracted. Of these, 425 families have exactly one sequence for each species, 1137 one-digit families, 121 double-digit families, and 5 three-digit families. Further 481 different patterns were detected.

5.1.2 ML method for individual gene families

For the maximum likelihood estimation (subsection 2.2.3) of the ancestral number α, the duplication rate λ, and the deletion rate μ for the four species in our set, the tree tMRHF from figure 1.3 was used. Estimates were calculated from both, the protein and the nucleotide dataset.

Every single pattern of sequence numbers was denoted with an identifier (ID) from one of the corresponding families. Then the parameters for each pattern were inferred using our software package (appendix A).

In table 5.1 an extract of the results from the HOGENNUCL data is given. It stands out, that all estimates for α have the same value, $\widehat{\alpha} = 50$, which was the upper boundary in these estimations. Thus, the method failed for all shown families. Since this behavior continued further in the procedure and additional problems (5.1.3) with the data came up, the analysis of the HOGENNUCL dataset was stopped.

In the analysis of the HOGENPROT dataset more reasonable results were obtained, since only for 9 of 53 processed patterns the upper boundary $\widehat{\alpha} = 50$ was reached (table 5.2). That is equal to a failure rate of 17%. Also for this dataset the computations were stopped due to difficulties in the data.

ID	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	$logL$
HBG004831	6	1	6	6	50	0.0100000000	0.0120721447	-5.402241481
HBG013194	20	5	34	9	50	0.0359185364	0.0368344588	-7.937746161
HBG016729	11	1	7	6	50	0.0204736831	0.0222562729	-6.552275197
HBG019644	4	1	12	6	50	0.0135433474	0.0154656241	-5.741676504
HBG025837	3	2	9	4	50	0.0058146269	0.0081105305	-4.702726611
HBG112829	8	1	5	3	50	0.0125784952	0.0147544792	-5.713160821
HBG030751	10	1	4	2	50	0.0164409072	0.0186657964	-6.098400229
HBG032298	6	1	11	4	50	0.0137904346	0.0157961655	-5.875263520
HBG032854	5	1	4	1	50	0.0051696551	0.0080387267	-4.533508450
HBG033225	10	3	4	3	50	0.0092546135	0.0115172938	-5.438410575
...		...				...		

Table 5.1: Part of the results for the dataset from the HOGENNUCL database. The first column contains the family identifier (ID). In the second to the fifth column the number of nucleotide sequences for each species is given. M = mouse, R = rat, H = human, and F = fruit fly. $\widehat{\alpha}$ is the estimated ancestral number, $\widehat{\lambda}$ the estimated duplication rate, $\widehat{\mu}$ the estimated deletion rate, and $logL = log(L_{tMRHF}(\widehat{\alpha}, \widehat{\lambda}, \widehat{\mu}))$ the corresponding log-likelihood value.

5.1.3 Problems with the data

One problem associated to the data from the the HOGENOM database is that the apparent size of a gene family in a given species varies whether HOGENNUCL or HOGENPROT is used. In table 5.3 some families with the corresponding number of sequences for the four species are given. It is easy to see, that the number of sequences differ remarkable between the nucleotide and protein data sets.

One possible explanation for these differences could be pseudogenes. Homologous sequences from pseudogenes would be found on the nucleotide level, if they are still homologous to the sequence of the considered gene. When analyzing protein sequences, pseudogenes are not considered, since they are not translated into proteins any more. According to our definition a pseudogene is considered as a deleted gene and should therefore not add to the number of genes in a family of a species.

Further investigations showed that for the HOGENOM data the differences in family size did not occur because of pseudogenes but due to redundancies in the data. That means, that homologous sequences from different resources might be stored as two independent sequences. Also isoforms or parts of the sequences are included in a family. Consequently the nucleotide sequences in the HOGENNUCL database do not reflect the number of genes of a species in a particular gene family.

Although the data from the HOGENPROT database obtained more reasonable results, also in this dataset redundant sequences were found. A good example is the family HBG000062. This gene family codes for the ATPase 6 in the mitochondrion. The family HBG000062 in humans contains 72 sequences in the HOGENNUCL database and two

ID	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	$logL$
HBG004452	2	1	2	2	2	1.0E-15	0.0002358791	-2.1441057012
HBG019735	5	5	6	3	3	0.0003970357	1.0E-15	-2.1830030037
HBG009835	3	3	3	4	4	1.0E-15	0.0001290479	-0.7678848270
HBG016812	3	1	4	4	50	0.0000029622	0.0027224864	-3.8010707888
HBG112829	1	1	1	1	1	1.0E-15	1.0E-15	-1.834729E-12
HBG012539	1	1	1	2	2	1.0E-15	0.0002810651	-0.7276794184
HBG018085	1	1	1	7	1	0.0013605019	1.0E-15	-2.0326646224
HBG018487	2	2	3	4	9	1.0E-15	0.0010454721	-2.2438932039
HBG018740	4	3	4	3	5	1.0E-15	0.0004560745	-2.2488222691
HBG034336	3	1	4	3	50	0.0000480224	0.0029094518	-3.7011690259
...		...					...	

Table 5.2: Part of the results for the dataset from the HOGENPROT database. Notation of the columns as in table 5.1.

sequences in the HOGENPROT database. However only one single gene without any further duplicates is annotated in the human mitochondrial genome.

5.1.4 Discussion

Due to the problems in the HOGENOM data the analysis of the entire dataset was stopped. The usage of adequate data is a main part in the analyzes, since inaccurate data might lead to incorrect results.

For the family HBG112829, e.g., the numbers of sequences from the HOGENNUCL database are different to the ones from the HOGENPROT database, which results in contradicting estimates for α. While for the first pattern the ancestral number reached the upper boundary $\widehat{\alpha} = 50$ in the estimation and the method failed, for the second pattern the ancestral number received the lower boundary $\widehat{\alpha} = 1$ (see HBG112829 in table 5.1 and table 5.2 highlighted in gray).

This is an impressive demonstration of how bioinformatic analyzes and their results rely on well assembled data. It further shows, that it is essential to know about the ideas how datasets have been compiled in order to use them properly.

5.2 Inparanoid database

The Inparanoid database is a comprehensive database of eukaryotic orthologous proteins. It consists of a collection of orthologous groups obtained from the pairwise comparison of proteins from species whose entire genome has been sequenced (`http://inparanoid.sbc.su.se/`).

Inparanoid was developed to identify truly orthologous proteins without redundancies.

ID	HOGENNUCL				HOGENPROT				ID	HOGENNUCL				HOGENPROT			
	M	R	H	F	M	R	H	F		M	R	H	F	M	R	H	F
HBG000012	5	3	5	3	3	2	2	2	HBG000246	2	4	5	5	2	2	3	2
HBG000042	2	1	4	4	2	1	1	1	HBG000251	11	9	14	9	6	5	8	5
HBG000049	1	1	4	2	1	1	1	1	HBG000258	16	12	34	5	5	3	4	1
HBG000052	2	1	6	6	1	1	1	5	HBG000269	5	6	13	4	2	2	3	2
HBG000062	17	4	72	5	2	2	2	1	HBG000304	13	6	25	10	4	3	7	4
HBG000066	2	1	6	2	1	1	1	1	HBG000315	25	6	25	5	7	2	6	2
HBG000069	1	1	9	3	1	1	1	1	HBG000321	9	3	13	8	4	3	5	5
HBG000123	6	2	5	4	3	1	1	2	HBG000330	15	6	7	8	6	3	2	3
HBG000146	2	1	5	2	1	1	1	2	HBG000333	2	1	5	23	1	1	1	1
HBG000163	8	6	24	7	2	2	1	2	HBG000335	3	1	13	9	1	1	1	1
HBG000166	17	2	412	2	2	2	2	1	HBG000339	4	2	5	3	2	2	1	1
HBG000189	1	1	26	1	1	1	1	1	HBG000341	1	1	2	1	1	1	1	1
HBG000222	37	10	43	25	9	7	9	9	HBG000355	9	2	6	2	2	2	3	1
HBG000223	46	20	58	112	13	10	13	14	HBG000362	7	2	16	2	3	2	2	1
...	...				...				...	...				...			

Table 5.3: Comparison of the species specific number of sequences from the HOGENNUCL database and the HOGENPROT database of some HOGENOM families. ID is the identifier of the family, M denotes the species mouse, R rat, H human, and F fruit fly.

Furthermore, paralogous genes, which originate from gene duplications can be identified. The inparanoid approach distinguishes between 'inparalogs' and 'outparalogs'. Inparalogs emerge from a gene duplication after speciation, while outparalogs result from a gene duplication prior to speciation. So inparalogs from one species can form a group of genes that is orthologous to a gene in another species. They are, therefore, also sometimes referred to as 'co-orthologs'.

A special interest of the Inparanoid database was to reduce the redundancy in the data. Protein sequences which were equal or almost equal to other protein sequences were removed. That results in a dataset where a gene is represented only by a single sequence and the number of genes is equal to the number of sequences. For more information about the reassessment of the data see O'Brien *et al.* (2005). The aim to avoid redundancies was the reason to choose the Inparanoid database to obtain data for the study on the tree tMRHF.

5.2.1 Data assembly

The data used in this study is based on the version 5.1 (January 2007) of the Inparanoid database. At this time the Inparanoid database contained 26 organisms and 511,758 sequences. 23,234 sequences are from mouse; 21,952 from rat; 22,218 from human; and 14,037 from fruit fly. These numbers are close to the number of genes known for these species.

To generate a dataset for these four species an outgroup was needed to ensure that the resulting gene families were already present in the MRCA of the four species with their appropriate gene copy numbers. For this purpose yeast (*Saccharomyces cerevisiae*) was

chosen. By comparing yeast and human, pairs of orthologs and associated inparalogs to these genes in both species were identified, that presumably emerged by duplication and diversification in the two lineages after the human and yeast lineage split. These genes are combined in an Inparanoid cluster. In total 2137 clusters could be detected. Genes found in these clusters can be assumed to have already been present in the ancestor of yeast and human and so were also present in the ancestor of mouse, rat, human, and fly. For each cluster the yeast ortolog was chosen as reference.
Then the set of these reference genes from yeast, one from each human-yeast-cluster, was compared to mouse, rat, human, and rat. Each of the resulting clusters included the yeast gene and all orthologs and inparalogs from mouse, rat, human, and fly. For each of the four species at least one gene was required for a valid cluster. These clusters can be defined as gene families, since these genes trace back to a single gene in the ancestor of yeast and the four species considered. In the following the clusters will referred to as gene families each with an ID equal to the corresponding yeast ortholog.

This procedure obtained a total of 1626 gene families. 758 gene families were represented only by a single gene in each species. All these gene families have the same pattern 'M-R-H-F' = '1-1-1-1'. For the remaining 868 gene families 273 different patterns were found. That led to a total of 274 patterns to analyze.

5.2.2 ML method for individual gene families

For each pattern the maximum likelihood estimation procedure was applied using the tree tMRHF as in the HOGENOM study before. Pattern specific ancestral gene copy numbers $\widehat{\alpha}$, duplication rates $\widehat{\lambda}$, and deletion rates $\widehat{\mu}$ were inferred. For 164 of 274 (60%) patterns the method failed because the upper boundary for α was reached. The results for the remaining 108 patterns are summarized in table 5.4.
283 gene families belong to the 164 patterns where the method failed, whereas 1341 gene families belong to patterns where the parameters could be inferred. Accordingly, on the level of gene families the failure rate reduces to 18%.
It should be noted that most of the gene families, for which the method succeeded, belong to the big group of gene families, where the number of genes for all four species are the same (M=R=H=F=x). Altogether 781 gene families had this kind of pattern. Thereof 758 gene families with x=1, 19 with x=2, 2 with x=3, 1 with x=4, and 1 with x=6. Since there is no difference in the number of copies between the species, the most likely scenario is, that no change had occurred during the evolution of these species. Thus, the estimated ancestral gene number is equal to the number of copies of the recent species $\widehat{\alpha} = x$, in most cases one. Further both estimated rates turned out to be zero (1.0E-15).

Mean of the estimates. When all gene families are considered together, the influence of gene families with equal gene copy numbers for all four species is apparent. Including these gene families in the analysis results in a change of the absolute values of the means

	gf.		M	R	H	F		$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	$logL$
patterns	108	m	4.56	5.38	3.98	3.88	m	7.24	0.002282	0.002363	-2.823976
		σ	4.46	6.86	2.43	5.81	σ	8.36	0.007234	0.007266	-
		$Q_{.5}$	4	4	4	2	$Q_{.5}$	3	0.000407	0.000167	-
		$\sum$	493	581	430	419	CI_u	26	0.009650	0.011201	-
no equals	560	m	2.42	2.59	2.29	1.79	m	3.68	0.000626	0.000719	-1.739761
		σ	2.34	3.39	1.55	2.79	σ	5.3	0.003277	0.003320	-
		$Q_{.5}$	2	2	2	1	$Q_{.5}$	2	1.0E-15	0.000269	-
		$\sum$	1356	1450	1282	1001	CI_u	17	0.001153	0.002333	-
all gf.	1341	m	1.62	1.69	1.56	1.35	m	2.14	0.000261	0.000301	-0.726522
		σ	1.67	2.33	1.2	1.85	σ	3.67	0.002139	0.002173	-
		$Q_{.5}$	1	1	1	1	$Q_{.5}$	1	1.0E-15	1.0E-15	-
		$\sum$	2168	2262	2094	1813	CI_u	10	0.000728	0.001358	-

Table 5.4: Results for the Inparanoid dataset. In the first five rows the results for the analysis of the 'patterns' are given. In the next five rows the results for all gene families without gene families where the gene copy numbers of all four species are equal (denoted with 'no equals') are summarized. The last five rows contain the results for all gene families (denoted with 'all gf.'). The column 'gf.' contains the number of gene families in the analyzes. Then for each species, namely mouse(M), rat(R), human(H) and fly(F), the mean m, standard deviation σ, median $Q_{.5}$ and sum $\sum$ of the gene copy numbers are given. Further the same characteristics and additional the upper boundary CI_u of the 95% one-sided confidence interval are given for the estimated ancestral gene numbers $\widehat{\alpha}$, the duplication rates $\widehat{\lambda}$, the deletion rates $\widehat{\mu}$, and the log-likelihood values $logL = log(L_{tMRHF}(\widehat{\alpha}, \widehat{\lambda}, \widehat{\mu}))$.

of the estimated parameters, but not in the relations of the means to each other. Thus, the averaged estimated ancestral number $m(\widehat{\alpha})$ excluding gene families with equal gene numbers is 3.7, whereas $m(\widehat{\alpha})$ for all gene families together is 2.1. The averaged estimated duplication rate is 0.00062 gene^{-1} myr^{-1} without equal-copy families and less than half of that, 0.00026 gene^{-1} myr^{-1}, when all gene families are included. The averaged estimated deletion rate is slightly higher than the duplication rate. Without equal-copy families, it is about 0.00072 gene^{-1} myr^{-1} and for all gene families it is 0.0003 gene^{-1} myr^{-1}. These and other characteristics of the estimates are also summarized in table 5.4.

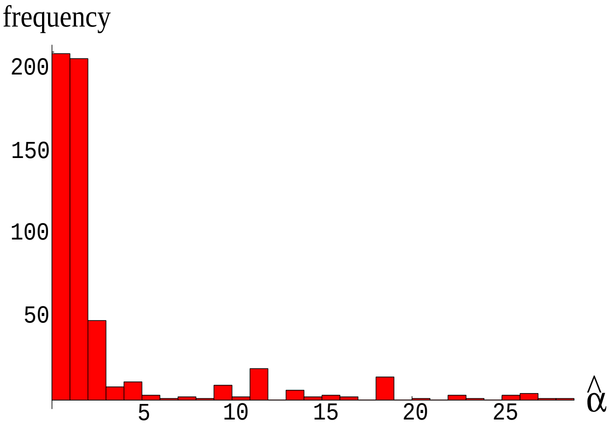


Figure 5.1: Inparanoid dataset: Values estimated for the ancestral number of gene copies α. On the x-axis the specific values for $\widehat{\alpha}$ can be found, while on the y-axis the frequencies are given. Gene families were excluded when the gene copy numbers for all species were equal. The remaining 560 gene families are included.

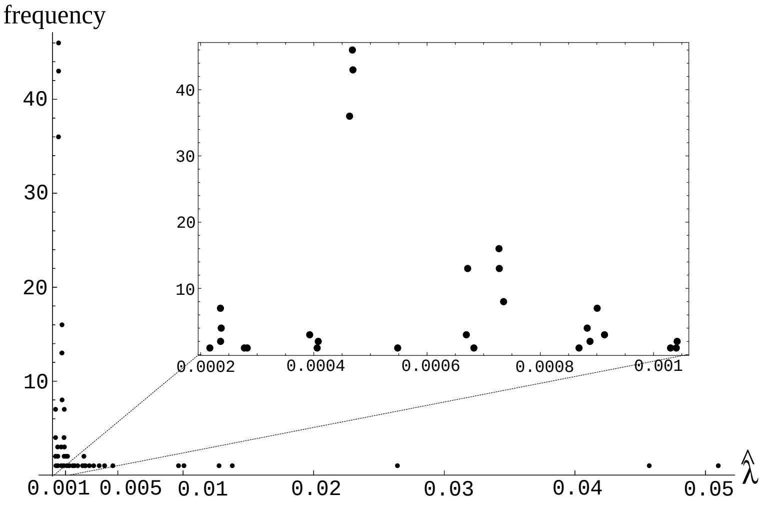


Figure 5.2: Inparanoid dataset: Values estimated for the duplication rate λ. On the x-axis the specific values for $\widehat{\lambda}$ are shown, while on the y-axis the frequencies are given. Gene families were excluded when the gene copy numbers for all species were equal and when the estimated duplication rate was 1.0E-15. 254 gene families remained. The interval for $\widehat{\lambda}$ between 0.0001 and 0.0011 is magnified.

Estimated ancestral number. The distribution of the inferred ancestral gene copy numbers $\widehat{\alpha}$ is illustrated in figure 5.1. Only the results for the 560 gene families without equal copy numbers for all species are shown. The most frequent ancestral gene numbers are one (209) and two (206), followed by three (48). Including the equal-copy gene families, $\widehat{\alpha} = 1$ became the most frequent value with an occurrence of 967. Then the next frequent was $\widehat{\alpha} = 2$ estimated for 225 gene families.

Estimated duplication and deletion rates. Here again gene families excluding equal-copy families are considered. For the remaining 560 gene families it was remarkable that in most cases one of the rates was estimated to be close to zero. For 306 gene families the duplication rate λ became 1.0E-15 and for 237 gene families the deletion rate μ became 1.0E-15. In figure 5.2 the distribution of the estimated duplication rates with $\widehat{\lambda}$ >1.0E-15 is shown. Most of the rates adopted values between 0.0002 and 0.01, but also higher values up to 0.05 could be found for the duplication rate.

The same plot was done for the deletion rate (see figure 5.3). The estimated values were also found between 0.0001 and 0.05. Here, the interval including most of the values was bigger than for the duplication rate and ranged from 0.0001 to 0.002. Overall, the estimated deletion rates are higher than the estimated duplication rates. That supports the previous observation and points at a small decrease in gene copy number from the MRCA to the recent species.

In only 17 cases both rates were larger than 1.0E-15, which is 3% of the total number of gene families. For these gene families a pairwise plot $\widehat{\lambda}$ against $\widehat{\mu}$ was made (figure 5.4). All points in the plot were found to be very close to the bisecting line. The Pearson correlation coefficient was computed with $\rho = 0.998859$ and indicates a very strong dependency between the two rates. Thus, if both rates were larger than 1.0E-15 they were almost equal.

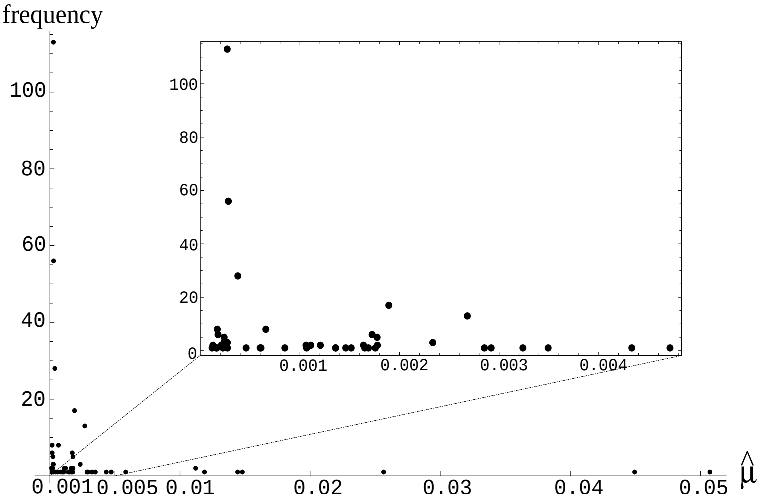


Figure 5.3: Inparanoid dataset: Values estimated for the deletion rate μ. On the x-axis the specific values for $\widehat{\mu}$ are shown, while on the y-axis the frequencies are given. Gene families were excluded when the gene copy numbers for all species were equal and when the estimated deletion rate was 1.0E-15. 323 gene families remained. The interval for $\widehat{\mu}$ between 0.0001 and 0.005 is magnified.

5.2.3 Discussion

For 1341 of 1626 gene families (82%) valid parameter values could be estimated. In 781 of these gene families the gene copy numbers for all species were equal and the ancestral gene number α was estimated to be equal to these number of genes. Consequently the estimated duplication and deletion rates are zero (1.0E-15). It is imaginable that also in these gene families changes in the gene copy number occurred, although today the numbers are the same for each species. Such changes can not be detected with method which uses only the number of genes.

For the remaining 560 gene families the rates were analyzed in more detail. For 306 gene families the duplication rate λ was estimated to be zero. Thus, the evolution of the genes in these gene families could be described with a pure death process. In contrast, for 237 gene families the deletion rate μ was estimated to be zero. Thus, a pure birth process could be applied to explain the change in gene copy number over time. For the remaining

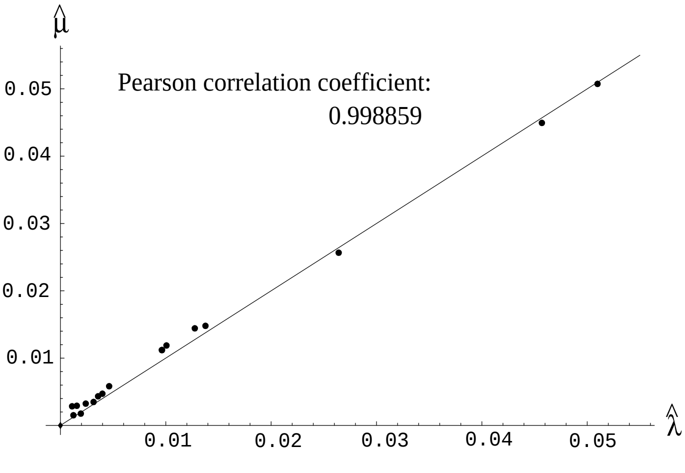


Figure 5.4: Inparanoid dataset: On the x-axis the estimated duplication rates $\widehat{\lambda}$ and on the y-axis the corresponding deletion rates $\widehat{\mu}$ are given. Here only gene families are included if non of the estimated rates was equal to zero (1.0E-15). 17 gene families remained.

17 gene families both rates are non zero, but happen to be nearly equal. In these cases, the evolution of gene copy numbers could be interpreted with a birth and death process with equal rates, $\lambda = \mu$.

The tendency to estimate one of the rates to be zero is an artifact of the method and was already observed in the simulation studies. So it is hard to distinguish if the results are due to the method or if it reflects the true evolution of the gene families. The fact that in all cases where both rates λ and μ were larger than zero, the inferred values are nearly equal is remarkable since this phenomenon was not observed in the simulation studies. This could indicate, that for biological data the most promising model is either a pure birth process, or a pure death process, or a BD process with equal rates.

In terms of finding global rates for duplication and deletion for this dataset the averaged estimated rates should be used. That would suggest a duplication rate of about 0.00026 gene^{-1} myr^{-1} and a deletion rate of 0.0003 gene^{-1} myr^{-1}. Here it should be considered that only those gene families were taken into account for which at least one gene was present in each of the species. That means, that gene families which were present in the MRCA of mouse, rat, human, and fruit fly but got extinct in one or more of the lineages are not considered at all. The current version of the program does not consider this evolutionary scenario. Including these families in the study would presumably result in an increase of the averaged estimated deletion rate $\widehat{\mu}$. This in turn would strengthen the observation, that the deletion rate was higher than the duplication rate.

Comparisons to previous studies (table 1.2) shows that our estimated global duplication and deletion rates are about one magnitude smaller. Nevertheless it is very hard to evaluate the rates, because rates of previous studies were specific for single species, whereas our rates are specific for individual gene families.

5.3 Ensembl database

The method was also applied to another biological dataset. There were two requirements on this dataset: First, it should cover more species. Second, the MRCA of all species considered should have lived more recent than in the previous study. There, the MRCA of mouse, rat, human, and fly was timed to 990 myr.

All requested properties are fulfilled by a dataset from an already published study on the evolution of mammalian gene families from Demuth *et al.* (2006). The five species included in this study are dog (*Canis familiaris*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), human (*Homo sapiens*), and chimp (*Pan troglodytes*). The MRCA of these species is dated back to 93 mya. The data was taken originally from the Ensembl database (release 41, October 2006; `http://www.ensembl.org/`).

5.3.1 Data assembly

Gene families were assembled by Demuth *et al.* (2006) using the MCL (Markov cluster) algorithm (Enright *et al.* (2002)). This method clusters proteins into families by simultaneous sequence analysis of all genes from all species. This approach differs from the typically used pairwise analysis. The MCL algorithm led to a dataset of 15,389 gene families. Demuth *et al.* (2006) excluded individual gene families and families which they assumed to be annotation artifacts, and arrived at a dataset containing 9990 gene families. The number of genes and gene families for the species can be found in the paper from Demuth *et al.* (2006) in table 1.

For the following analysis we excluded all gene families where at least one of the species was not represented. So the final dataset included 8569 gene families in total. 4477 gene families were found with single-copy genes. Further 3813 one-digit, 275 double-digit, and four three-digit gene families could be found. In the one-digit families 721 patterns of gene copy numbers 'D-M-R-H-C' could be determined, in the two-digit families 274 patterns and four patterns in the three-digit gene families. In total 1000 different patterns were detected.

5.3.2 ML method for individual gene families

Our software package (appendix A) was again used to estimate the parameters specific for each of the 1000 patterns. Therefore the tree tMRHCD including the five considered species dog, mouse, rat, human, and chimp from figure 1.3 was used. This species tree with its defined branch length was also taken from the study of Demuth *et al.* (2006).

The computations were aborted for 19 patterns including the ones of all three-digit gene families due to the expected running time. For very high copy numbers the precision in the computations had to be set very high. This in turn causes an unreasonably long running time, which can be up to years. For 279 of the remaining 981 patterns (28%) the method failed because the upper boundary for α was reached. The results of the estimated parameters for the remaining 702 patterns are summarized in table 5.5. The 702 patterns where the method succeeded, represent 8178 gene families. In contrast, to the 279 patterns where the method failed only 372 gene families could be assigned. On the level of gene families this results in a failure rate of only 4%.

In 5622 of the 8178 gene families the number of gene copies for all species were equal. That means for each pattern 'D-M-R-H-C' it is D=M=R=H=C=x. The following table shows how many gene families were found for different x:

x	1	2	3	4	5	6	7	8	9	12	14	27
no. gf.	4477	734	257	89	25	18	11	5	3	1	1	1

For all these gene families, the ancestral gene number α was estimated to be x, the duplication rate as well as the deletion rate was estimated to be zero (1.0E-15). The

	gf.		D	M	R	H	C	no.	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	$logL$
patterns	702	m	5.51	7.41	7.33	5.88	6.36	m	8.79	0.005694	0.006511	-3.001606
		σ	5.15	8.85	9.65	4.84	5.7	σ	10.89	0.018246	0.019549	-
		$Q_{.5}$	4	5	5	5	5	$Q_{.5}$	6	1.0E-15	0.001153	-
		$\sum$	3866	5199	5149	4130	4468	CI_u	27	0.022008	0.024989	-
no equals	2556	m	2.96	3.63	3.55	3.08	3.29	m	4.11	0.002798	0.002651	-1.999437
		σ	3.24	5.27	5.64	3.19	3.66	σ	6.75	0.009831	0.010809	-
		$Q_{.5}$	2	2	2	2	2	$Q_{.5}$	2	0.001126	1.0E-15	-
		$\sum$	7571	9279	9086	7876	8408	CI_u	13	0.007997	0.009908	-
all gf.	8178	m	1.84	2.05	2.03	1.88	1.95	m	2.2	0.000875	0.000829	-0.624916
		σ	2.1	3.22	3.4	2.1	2.36	σ	4.06	0.005646	0.006166	-
		$Q_{.5}$	1	1	1	1	1	$Q_{.5}$	1	1.0E-15	1.0E-15	-
		$\sum$	15073	16781	16588	15378	15910	CI_u	6	0.003289	0.003168	-

Table 5.5: Results for the Ensembl dataset. In the first five rows the results for the analysis of the 'patterns' are given. In the next five rows the results for all gene families without gene families where the gene copy numbers of all four species are equal (denoted with 'no equals') are summarized. The last five rows contain the results for all gene families (denoted with 'all gf.'). The column 'gf.' contains the number of gene families in the analyzes. Then for each species, namely dog(D), mouse(M), rat(R), human(H) and chimp(C), the mean m, standard deviation σ, median $Q_{.5}$ and sum $\sum$ of the gene copy numbers are given. Further the same characteristics and additional the upper boundary CI_u of the 95% one-sided confidence interval are given for the estimated ancestral gene numbers $\widehat{\alpha}$, the duplication rates $\widehat{\lambda}$, the deletion rates $\widehat{\mu}$, and the log-likelihood values $logL = log(L_{tMRHF}(\widehat{\alpha}, \widehat{\lambda}, \widehat{\mu}))$.

results for all gene families without these 5622 equal-copy families are summarized in table 5.5, row six to ten, denoted with 'no equals'. In addition, also the results for all gene families together are given in that table.

Mean of the estimates. The mean of the estimated ancestral gene numbers for the patterns is equal to 8.79. Further, the mean of the estimated duplication rates is 0.0057 gene^{-1} myr^{-1}, whereas the mean of the estimated deletion rate is 0.0065 gene^{-1} myr^{-1} for the patterns. As for the Inparanoid dataset, the averaged deletion rate for the patterns here seems to be bigger than the averaged duplication rate. Also the median for the deletion rate $Q_{.5}(\mu)$ was found to be larger than the median for the duplication rate $Q_{.5}(\lambda)$.

The estimates for the parameters for all gene families and for the gene families exclusive equal-copy families were compared. As already seen for the Inparanoid dataset, the estimates do not change with respect to their relative size to each other, but in their absolute values. The estimated global rate for duplications turned out to be 0.000875 gene^{-1} myr^{-1} whereas the global rate for deletions was estimated to be 0.000829 gene^{-1} myr^{-1}. Thus, when gene families are considered instead of pattern the rates were nearly the same. Actually, in this case the duplication rate was a little bit higher than the deletion rate.

Estimated ancestral number. The ancestral number of gene copies α is also strongly influenced by the gene families with equal gene copy numbers for all species. In figure 5.5

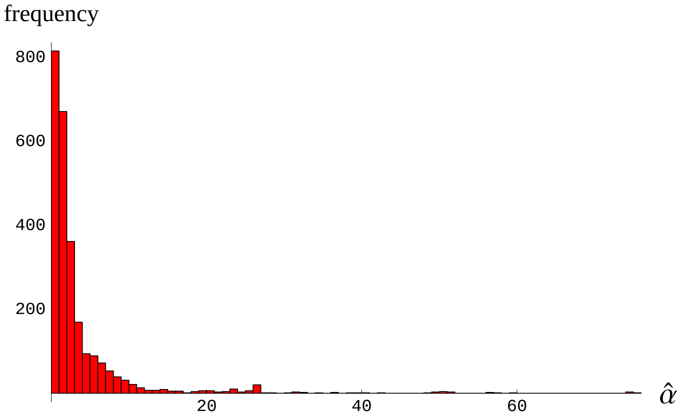


Figure 5.5: Ensembl dataset: Values estimated for the ancestral number of gene copies α. On the x-axis the specific values for $\widehat{\alpha}$ can be found, while on the y-axis the frequencies are given. Gene families were excluded when the gene copy numbers for all species were equal. The remaining 2556 gene families are included.

the distribution of α for all gene families without the ones with equal copy numbers is shown. Already here it becomes apparent that $\alpha = 1$ is most frequent, followed by $\alpha = 2$ and $\alpha = 3$. The averaged estimated ancestral gene number is found to be 2 when all gene families without equal-copy families are considered. The fact that 4477 gene families have exactly one gene copy for each species is the reason that for all gene families together the estimated global α turned out to be 1.

Estimated duplication and deletion rates. Furthermore the distributions for the estimated rates were examined. In this analysis again all equal-copy families were excluded, since for all these gene families both rates were estimated to be zero (1.0E-15). Thus, for the following analyzes 2556 gene families were considered.
First, the distribution of the estimated duplication rate $\widehat{\lambda}$ was analyzed. Here only values greater than 1.0E-15 were taken into account. In total, for 1447 gene families (57%) such a duplication rate was estimated. The distribution is shown in figure 5.6. Most duplication rates range between 0.0001 and 0.01 gene^{-1} myr^{-1}, with a maximum at around 0.003.

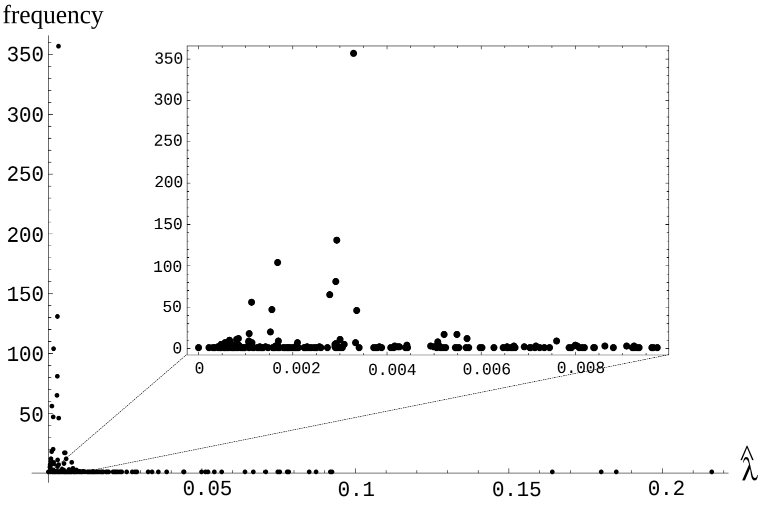


Figure 5.6: Ensembl dataset: Values estimated for the duplication rate λ. On the x-axis the specific values for $\widehat{\lambda}$ are shown, while on the y-axis the frequencies are given. Gene families were excluded when the gene copy numbers for all species were equal and when the estimated duplication rate was 1.0E-15. 1447 gene families remained. The interval for $\widehat{\lambda}$ between 0 and 0.01 is magnified.

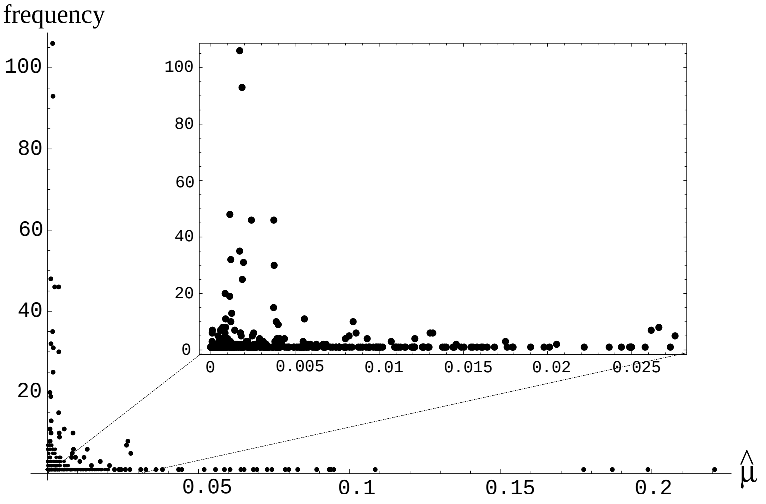


Figure 5.7: Ensembl dataset: Values estimated for the deletion rate μ. On the x-axis the specific values for $\widehat{\mu}$ are shown, while on the y-axis the frequencies are given. Gene families were excluded when the gene copy numbers for all species were equal and when the estimated deletion rate was 1.0E-15. 1204 gene families remained. The interval for $\widehat{\mu}$ between 0.0001 and 0.03 is magnified.

Also the distribution of the estimated deletion rate $\widehat{\mu}$ ($\widehat{\mu}$ >1.0E-15) was analyzed. In this case only 1204 (47%) gene families were found that meet this requirement. The range where most values are situated is more wide-stretched as for the duplication rate (0.0001-0.03), with a maximum at around 0.002.

Comparing both distributions, it seems that for the Ensembl data the duplication rates are slightly higher than the deletion rates. Nevertheless the distributions look very similar and the differences between the rates are not outstanding.

95 of the 2556 gene families (4%) remained with both rates greater than 1.0E-15. The estimated values for the duplication rate λ and the deletion rate μ of these gene families we show in a pairwise plot (figure 5.8). As for the Inparanoid dataset, it is striking that all datapoints are near the bisecting line. Accordingly, the two rates are highly correlated (Pearson correlation coefficient 0.97575). Thus, if duplication rate and deletion rate are larger than 1.0E-15, they appear to have nearly the same values.

Averaged duplication-deltion rate. In the study of Demuth *et al.* (2006) on the same dataset, a BD model assuming equal birth and death rates was used to estimate a joint birth-death rate. To compare our estimates with the results from the study of Demuth *et al.* (2006) the average of the duplication rate and the deletion rate was calculated for each of the 2556 gene families. The distribution of these averaged rate is shown in figure 5.9. The mean m, the standard deviation σ, the median $Q_{.5}$, and the boundaries of the 95% confidence interval CI_{left} and CI_{right} were computed for the 2556 gene families under consideration, which means exclusively equal-gene families ('no equals'), and additionally for all 8550 gene families together (all gf.):

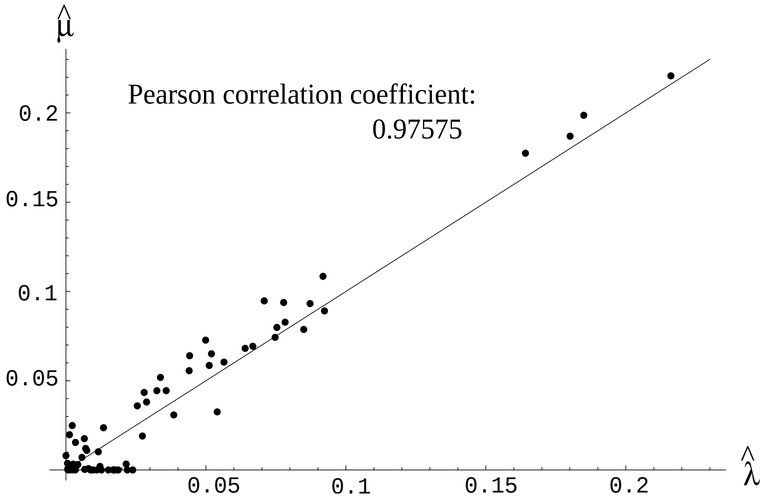


Figure 5.8: Ensembl dataset: On the x-axis the estimated duplication rates $\widehat{\lambda}$ and on the y-axis the corresponding deletion rates $\widehat{\mu}$ are given. Here only gene families are shown if $\lambda >$ 1.0E-15 & $\mu >$ 1.0E-15. 95 gene families remained.

	m	σ	Q_5	CI_{left}	CI_{right}
no equals	0.002725	0.009954	0.001456	0.000281	0.011937
all gf.	0.000852	0.005706	1.0E-15	1.0E-15	0.004697

Since for the 5622 gene families with equal gene copy numbers for all species both rates were estimated to be zero (1.0E-15), also the average rate is zero (1.0E-15). The adding of these gene families to the analysis leads to a huge decrease (3.4 times) in the mean of the averaged duplication-deletion rates.

Looking at the distribution of figure 5.9 more closely, it turns out that most averaged duplication-deletion rates lie between 0.0001 and 0.0025. The most frequent value for the rate is 0.0016 gene^{-1} myr^{-1} with a occurrence of 364 (415 values with values between 0.00156 and 0.00168). This value is very near to the median $Q_{.5} = 0.0015$ of the averaged duplication-deletion rates. Of course, taking also gene families with equal copy numbers into account the most frequent averaged rate would be 1.0E-15 with an occurrence of 5622.

5.3.3 Discussion

For 96% of all gene families valid estimates for the ancestral gene copy number α, the duplication rate λ, and the deletion rate μ could be inferred. For 69% of these gene families the copy numbers of all species were equal. As already discussed for the Inparanoid dataset, these gene families do not provide any information for the estimation of the parameters.

Estimated rates. The remaining 2556 gene families were analyzed further. For 1109 (43%) gene families the estimated duplication rate was zero (1.0E-15). For 1352 (53%) gene families the deletion rate was estimated to be zero (1.0E-15). Finally, in 95 (4%) gene families both rates were larger than 1.0E-15, but nearly equal for most of the 95 gene families.

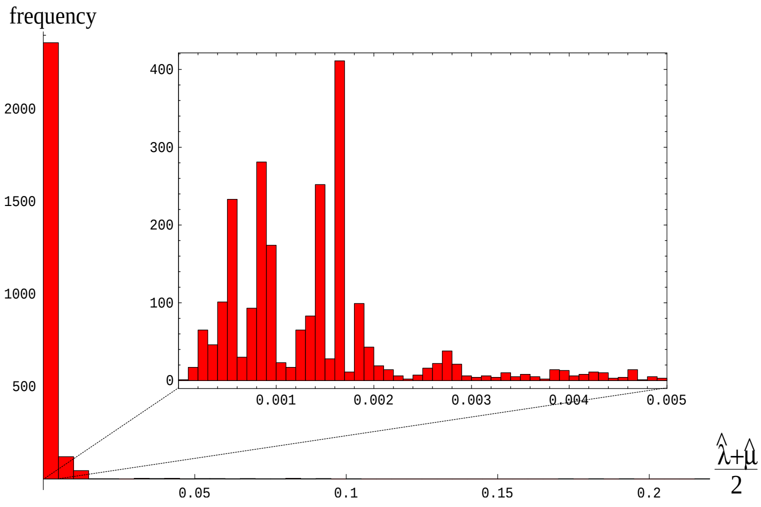


Figure 5.9: Ensembl dataset: Average of the estimated duplication rates $\widehat{\lambda}$ and the deletion rates $\widehat{\mu}$. On the x-axis the specific values for $(\widehat{\lambda}+\widehat{\mu})/2$ can be found, while on the y-axis the frequencies are given. Gene families were excluded when the gene copy numbers for all species were equal. 2556 gene families remained. The section where $(\widehat{\lambda}+\widehat{\mu})/2$ is between 0 and 0.005 is scaled up.

Again, parallels between the Inparanoid and the Ensembl dataset are seen. Also for the Ensembl dataset, the evolution of the individual gene families could be either explained with a pure birth process, or a pure death process, or a birth and death process with equal rates for duplications and deletions.

Comparison to Inparanoid dataset. To compare the results to previous estimates global rates have to be considered. Global rates for the Ensembl dataset are defined as the means of the individual rates from all gene families. That would suggest a duplication rate λ of 0.00088 gene^{-1} myr^{-1}. The 95% confidence interval for λ is then 0 .. 0.0052. The deletion rate for all gene families together was estimated to be 0.00083 gene^{-1} myr^{-1} with a 95% confidence interval of 0 .. 0.0057.

In comparison to the Inparanoid, dataset both rates are about 3.5 times higher. Furthermore, the duplication rate λ slightly exceeds the deletion rate μ. However, the difference between the two rates is small (0.058) compared to the Inparanoid data (0.14). As before, 1457 gene families where at least one of the species had no gene copy were excluded from the analysis. We speculate that the deletion rate μ for these gene families would be higher than the duplication rate λ, since the gene families became extinct in some of the lineages. This would, in turn, lead to a higher global deletion rate while the global duplication rate would stay the same. It is hard to say if both rates would be equal or if the deletion rate would be higher than the duplication rate afterwards.

Comparison to Cotton & Page. Our global rates were compared to previously reported estimates (see table 1.2). Compared to estimates from Lynch & Connery (2000, 2001, 2003) our estimates are about one magnitude lower. But recent estimates for the duplication rate from Cotton & Page (2005), which are 0.00097 gene^{-1} myr^{-1} and 0.00115 gene^{-1} myr^{-1} respectively, are very close to our estimate of 0.00088 gene^{-1} myr^{-1}. The estimated deletion rate from Cotton & Page (2005) depends on the age of the duplication events in the data. For data with duplication events which date from ∼4700 mya to present day, a deletion rate of 0.00048 gene^{-1} myr^{-1} was estimated, whereas for data with duplications from the last 200 myr only, a deletion of 0.0074 gene^{-1} myr^{-1} was inferred. Since our data incorporate the evolution of gene families for the last 93 myr only, our estimate was compared to the latter estimate of Cotton & Page (2005). In contrast to the duplication rate, our estimated deletion rate, 0.00083 gene^{-1} myr^{-1}, is again about one magnitude lower than the one proposed from Cotton & Page (2005) for this time interval.

Comparison to Demuth *et al.*. For the comparison of our estimates to the estimate from Demuth *et al.* (2006), the average of the estimated duplication and deletion rate $(\widehat{\lambda} + \widehat{\mu})/2$ was computed. Demuth *et al.* (2006) used a similar dataset composed of our dataset and the 1457 gene families where at least one species had no gene copy, which were not considered in our analysis. In contrast to our model, the model Demuth *et al.* (2006) used only allows for equal duplication and deletion rates. Their global estimate for this rate is 0.0016 gene^{-1} myr^{-1}. Our estimate for the average duplication-deletion rate for all gene families together is 0.00085 gene^{-1} myr^{-1}. That is half of the estimate of Demuth *et al.* (2006). But if the distribution of our averaged duplication-deletion rate is considered (figure 5.9), the most frequent value for the rate besides zero is 0.0016 gene^{-1} myr^{-1}. If we could include the 1457 gene families where at least one species had no gene copy, the results might change. Assuming that the deletion rates for these gene families would be high, we can speculate that our estimate for the average duplication-deletion rate would increase towards the estimate of Demuth *et al.* (2006).
Referring to our analysis of the Ensembl dataset, we cannot decide whether the assumption of equal birth and death rates from to Demuth *et al.* (2006) is justified. The occurrence of gene families with one rate equals zero in 89% on the pattern level and in 96% on the gene family level, excluding all equal-copy families, compared to 87% in simulation studies, makes it impossible to decide, if this is only due to artifacts of the method or also due to the data.

5.4 Conclusion

The application of the maximum likelihood method on biological data showed impressively how much the results of the estimation depend on the quality of the data. The method uses only very little information about a gene family, namely the number of gene copies for each species. Therefore at least these numbers should be correct.

The method failed if the upper boundary for the ancestral gene copy number α in the estimation procedure was reached. The value of this boundary and larger values for α are supposed to be implausible for the data. There can be several reasons why this upper boundary has been reached: It is an artifact of the method. In simulation studies we found that for the tree tMRHF an average failure rate of 20% and for the tree tMRHCD an average failure rate of 10% can be expected even for simulated data that fit the model perfectly.

These failure rates must be compared to the failure rates of the patterns of the biological datasets, since in the simulations gene families differed in their patterns too. The failure rate for the patterns of the Inparanoid dataset for the tree tMRHF with 60% is 3 times higher than expected. For the Ensembl dataset for the tree tMRHCD the failure rate for the patterns with 28% is 2.8 times higher than expected. So it seems that these high failure rates are not only caused by artifacts of the method.

Another explanation for this can be the quality of the data. Even small changes in the gene copy number can have big influences on the estimation of the parameters. That was also found in simulation studies. It is known that there are problems with redundancies, sequencing and annotation errors, etc. in current databases for nucleotide and protein data. Also the classification into gene families at all is not completely clear. Different method with slightly different criteria can be used for this classification which can lead to different gene families.

The third reason for the high failure rates can be the fact, that the analyzed data cannot be explained with the used BD model. As briefly summarized in the first chapter, there are many mechanisms which might have influenced the evolution of gene families. The BD model used here describes only the very simple mechanism of one gene duplication or one deletion at a time. Complex processes, like chromosome or whole genome duplications, are not taken into account by our model. This topic will be discussed further in the next chapter.

The results for both biological datasets suggest that the evolution of the gene families can be explained with either a pure birth process, or a pure death process, or a BD process with equal birth and death rates. On one hand this is again due to the method, but can also be caused by the data. Overall it should be noted that the few information which is given by a gene family with only four or five species might not be sufficient to overcome the problem of estimating one rate to be equal to zero. Studies on trees with more extant species and high quality data could shown if this problem disappears using more species, more data.

The estimated global rates are smaller than previous published rates. This could be due to the general underestimation of the rates as was seen in all simulation studies. Another reason for the underestimation of the rates can be the choice of the data. Gene families where at least one species has no gene copy at all were excluded in the study. These gene families are very problematic. For one species without any copy number it seems very likely that the gene must be lost on the branch to that species. But what happens if more or even all except of one species have no gene copies? How can be decide if the gene was present in internal ancestors? And even more questionable in these cases, is it possible to decide if the gene family was already present the MRCA or if the family was originated in recent time. That will need further investigation.

CHAPTER

SIX

An outlook to more complex models and future work

Our simulation studies revealed various artifacts of our method. One example is the estimation of the duplication or the deletion rate to be equal to 1.0E-15. Furthermore, the quality of our estimates was dependent on the proportions of the true rates, but in total, we found the inferred parameters to be underestimated. We observed on average a failure rate of 15% for simulated data, whereas the method was classified as failed, if no valid estimates for a specific gene family could be found. When applying our method to real data, we found an even higher failure rate, as well as many gene families with at least one estimated rate being 1.0E-15.

In the last chapters several suggestions for reasons for these artifacts were discussed. Here, some of them will be analyzed further based on individual simulation studies. Furthermore we will give an outlook on more complex models and future work.

6.1 Larger trees

All studies on simulated and real data were based on a four species or a five species tree. In chapter 4 it was already mentioned that the usage of trees with more leaves might overcome the artifacts of the method, since more data is included into the estimation of the parameters. To test this hypothesis one simulation study was performed.

6.1.1 Simulation study

The species tree for this study should have many contemporary species and a relatively short overall time from the MRCA to these species. For this purpose we choose again a pure mammalian tree including 11 species which is denoted by tMAMMALS (figure 6.1). The MRCA of these species is dated to only 92 mya (Hedges (2002)). For this time interval no big changes in the gene copy numbers, due to e.g. large-scale duplications, are expected. For that reason this tree might be also a good choice as a basis for real data studies.

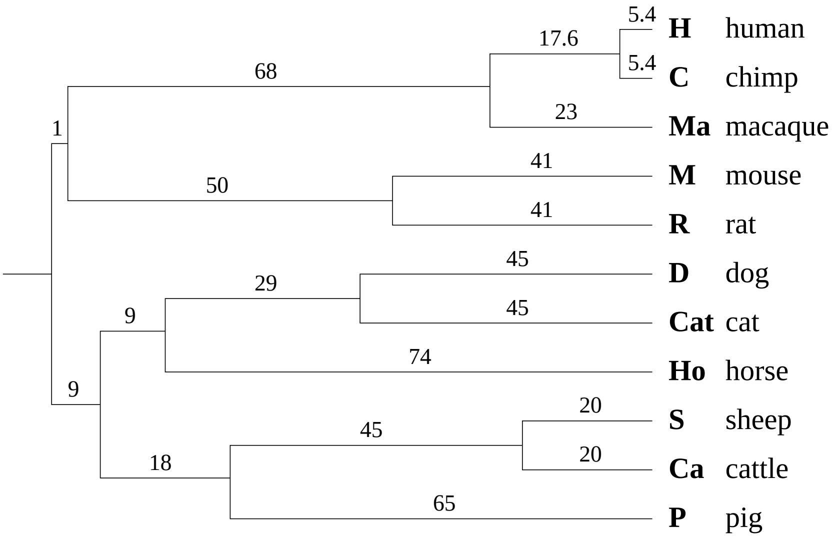


Figure 6.1: Species tree tMAMMALS for human, chimp, macaque, mouse, rat, dog, cat, horse, sheep, cattle, and pig. The edge lengths refer to time in myr. Time estimates are from Hedges (2002).

Along is tree, we simulated the evolution of 100 gene families using the following parameter values:

$$\alpha = 10, \quad \lambda = 0.002, \quad \mu = 0.008 \tag{6.1}$$

We then used our maximum likelihood method to estimate the ancestral gene copy number α, the duplication rate λ, and the deletion rate μ. The simulated gene copy numbers for the individual gene families as well as the particular results can be found in appendix C.2. In table 6.1 the statistics for the results are shown.

For 94 of the 100 gene families valid estimates could be found. That corresponds to a failure rate of 6%. To compare the results with regard to the failure rate and the quality of the estimates, the tree tMRHCD was taken. This tree includes five species and has with 93 mya almost the same time to the MRCA as the species in tree tMAMMALS. For the tree tMRCHD a simulation study with the same parameters was done (section 4.3.3). There the failure rate was 11% which is twice as high as the one for the tree tMAMMALS.

The mean of the estimated ancestral gene number is 9.3 which is close to the value of 10 we used for the simulation. The mean of the estimated duplication rates is 0.00296 gene^{-1} myr^{-1} while the mean of the estimated deletion rates is 0.00547 gene^{-1} myr^{-1}. Comparing these estimates to the values used for the simulation, it is easy to see that the duplication rate is overestimated while the deletion rate is underestimated. The ratio μ/λ is 4 for the 'true' values. In contrast, this ratio is only 1.85 for the estimates.

One thing stands out. That is the occurrence of 26 of 94 gene families (28%) where both estimated rates are larger than zero (1.0E-15). In the simulation study using tree tMRHCD this was found only in 15% of the analyzed families and for biological data in only 3-4% of the gene families. In contrast to the biological data, for these 26 gene families no strong dependency between the rates could be detected. The Pearson correlation coefficient was computed to be 0.6187.

tMAMMALS						tMRHCD				
gf.		$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	$logL$	gf.		$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
94	m	9.3	0.0029584563	0.0054652161	-8.4858620471	89	m	10.2	0.001541	0.005699
	σ	7.1	0.0038602705	0.0054760376	1.2512741518		σ	8.01	0.003046	0.006392
	$Q_{.5}$	8.0	1.0E-15	0.0044113565	-8.6685306046		-	-	-	-
	CI_l	2	1.0E-15	1.0E-15	-		CI_l	3	1.0E-15	1.0E-15
	CI_u	26	0.0107615299	0.00185442579	-		CI_u	27	0.008168	0.019656

Table 6.1: Results for tree tMAMMALS. The column 'gf.' contains the number of gene families with valid estimates. The mean m, standard deviation σ, median $Q_{.5}$, and additional the lower CI_l and the upper CI_u boundary of the 95% confidence interval are given for the estimated ancestral gene numbers $\widehat{\alpha}$, the duplication rates $\widehat{\lambda}$, the deletion rates $\widehat{\mu}$, and the corresponding log-likelihood values $logL = log(L_{tMRHF}(\widehat{\alpha}, \widehat{\lambda}, \widehat{\mu}))$. The left part entiteled 'tMAMMALS' includes the results for the current simulation study, while the right part includes the results from a previous study for the tree 'tMRHCD' for comparison (see also table 4.9).

6.1.2 Discussion and conclusion

Although the failure rate for the tMAMMALS data is only half that of the tree tMRHCD data, the quality of the estimates did not improve. Quite the contrary could be observed for the chosen parameter set. The estimates for the ancestral gene number are for both trees acceptable. However, the quality of the estimates for the rates differs. While for the tMRHCD data both rates are underestimated, the ratio of the rates are nearly the same as for the 'true' values. That could not be found for the tMAMMALS data.

This one simulation study is not sufficient to make reliable statements but gives the impression that larger trees do not necessarily increase the quality of the estimation. More simulation studies should be done to verify this observation. Unfortunately the running time for studies on large trees is very long, which makes extensive simulation studies difficult.

6.2 Whole genome duplications

In subsection 1.2.2 whole genome duplications were introduced as one important mechanism for the evolution of gene families. Examples like the Hox gene cluster imply that several rounds of WGD took place in the past. To verify the influence of WGD on our estimation of the ancestral gene copy number α, the duplication rate λ, and the deletion rate μ, only one simulation study was performed.

6.2.1 Simulation study

Again the species tree tMRHF (figure 6.2 left) was used. As described in recent studies, there are two WGDs assumed to have occurred at the origin of vertebrates (Holland *et al.* (1994)). Therefore, we simulated gene family evolution with a BD process with fixed rates for gene duplication and deletion, including two rounds of WGD (figure 6.2 to the right). On should note, that these WGDs itself, and even more the points in time when

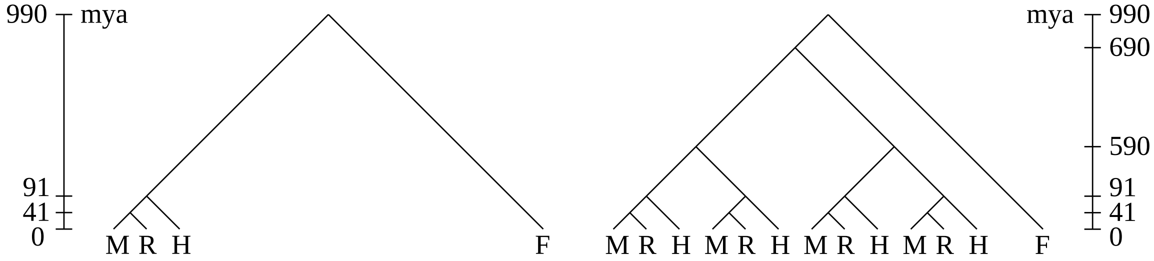


Figure 6.2: Schematic illustration of the species tree tMRHF for mouse, rat, human, and fruit fly on the left side. On the right side is a schematic illustration of the same tree, but with inclusion of two WGD, referring to the 2R hypothesis. One WGD 590 mya and the other one 690 mya. Times of the WGD are averaged from Panopoulou & Poustka (2005), speciation times are from Hedges (2002).

they might have occurred, are disputed. In Panopoulou & Poustka (2005) a time interval of ~530-738 mya for the two WGD was found to be strongly supported by several data sources. Furthermore, it was suggested that both WGDs might have happened within a relative short time interval of 90-106 myr. Referring to these time information the time points for the WGDs were averaged. Thus, the times for the first and the second WGD were assumed to be 690 mya and 590 mya, respectively.

Based on the species tree tMRHF the two WGD were included in the tree. A schematic illustration of this tree, referred to as tMRHF2R in the following, can be found in figure 6.2 to the right. On this tree simulations as described in chapter 4 were carried out. In doing so, to each leaf of tree tMRHF2R a specific number of gene copies was assigned. The parameters used for the simulation assuming a higher deletion rate associated with the occurrence of WGDs, were defined as:

$$\alpha = 5, \quad \lambda = 0.0001, \quad \mu = 0.002 \tag{6.2}$$

Further a second dataset for tree tMRHF was derived from the tMRHF2R data. Thereto the gene copy numbers from each tMRHF2R gene family were added up for mouse, rat, and human, respectively. In the end to each species exactly one number of gene copies was assigned. This second dataset refers to biological datasets, where also simply the total number of gene copies would be counted for each species. In table 6.2 an extract of the resulting simulated data is shown. Gene copy numbers for gene families of the first dataset can be found in the right hand part of the table entitled tMRHF2R, while the corresponding copy numbers of the gene families of the second dataset are in the left hand part entitled tMRHF. In total 100 gene families were included in the analysis.

For both trees tMRHF and tMRHF2R in conjunction with their corresponding datasets our maximum likelihood method was used to estimate the ancestral gene copy number α, the duplication rate λ, and the deletion rate μ. In doing so, for the tree tMRHF the parameters are estimated under the assumption that the gene families only evolved according to the BD model, whereas for the tree tMRHF2R additionally the occurrences of the two WGDs are included in the analysis. The estimated parameters for an extract

tree	tMRHF				tMRHF2R												
					subtree 1			subtree 2			subtree 3			subtree 4			
no.	M	R	H	F	M	R	H	M	R	H	M	R	H	M	R	H	F
1	5	5	7	2	1	1	0	1	1	2	2	2	3	1	1	2	2
2	7	6	8	2	2	3	3	2	1	2	1	1	1	2	1	2	2
3	12	12	12	3	3	3	3	3	3	2	3	3	4	3	3	3	3
4	8	10	9	3	2	2	1	2	4	4	2	2	2	2	2	2	3
5	6	6	8	2	2	2	1	1	1	3	1	0	1	2	3	3	2
6	9	8	6	2	3	3	2	1	1	1	2	1	2	3	3	1	2
7	4	8	7	2	1	2	2	1	3	2	2	2	2	0	1	1	2
8	7	6	4	3	1	1	1	2	1	1	1	1	0	3	3	2	3
9	10	10	11	1	2	3	3	3	3	3	3	2	4	2	2	1	1
10	4	5	6	3	1	1	1	1	1	1	1	1	2	1	2	2	3

Table 6.2: Extract of the simulated gene copy numbers for the trees tMRHF and tMRHF2R from figure 6.2 assuming two rounds of WGD. The gene families are consecutively numbered, denoted with 'no.'. M is the usual abbreviation for mouse, R for rat, H for human, and F for fruit fly.

of the gene families can be found in table 6.3. Every row corresponds to one gene family and includes the results for both estimation approaches in parallel. First, the results for the estimation using the tree tMRHF are given, followed by the results for the estimation using the tree tMRHF2R. Furthermore the mean m, the standard deviation σ, and the median $Q_{.5}$ for the estimated parameters of the entire dataset are given.

Highlighted in gray are estimations where the method failed because the ancestral gene number reached the upper boundary ($ub(\alpha) = 50$). We found that for some gene families in both estimation approaches the method failed, but that there are also gene families where either the tMRHF-estimation or the tMRHF2R-estimation failed. In total, the failure rate for the tree tMRHF was 30%, in contrast to the tree tMRHF2R where the failure rate was only 17%. In simulation studies for the tree tMRHF with gene family data simulated exclusively under the BD model without WGDs the failure rate was only about 20%. That results in an increase of the failure rate about a half when using data simulated with WGDs. For the tMRHF2R-estimation the obtained failure rate of 17% is hard to assess, because no comparable analyses were made. Without any further simulation studies on the tree tMRHF2R, to get an averaged failure rate, it is impossible to judge whether the 17% is a high or a low failure rate for this tree. Therefore it is not possible to compare the two estimation strategies here in terms of failure rate.

The mean of the estimated ancestral number $\widehat{\alpha}$ is for both estimation strategies almost the same, but about 1.5 times higher than the value used in the simulation. If the median is considered the 'true' value of α is found exactly for the tMRHF2R-estimation, while the median for the tMRHF-estimation is smaller than the 'true' value.

The averaged rates for gene duplication and deletion for the tMRHF-estimation are far

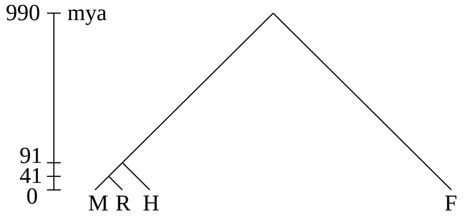
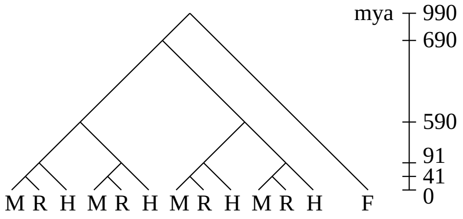
	without WGD				with WGD			
					 mya 990 690 590 91 41 0 M R H M R H M R H M R H F			
no.	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	2	0.000893	1.0E-15	-3.187494	50	1.0E-15	0.003396	-7.632127
2	50	0.001027	0.003380	-3.784293	24	1.0E-15	0.002563	-6.553977
3	50	0.003485	0.005918	-4.506535	50	1.0E-15	0.003459	-7.689670
4	1	0.001575	1.0E-15	-3.413099	15	1.0E-15	0.001949	-7.153791
5	50	1.0E-15	0.002492	-2.947468	4	1.0E-15	0.000849	-5.184240
6	2	0.000779	1.0E-15	-2.688531	4	1.0E-15	0.000837	-5.271556
7	10	0.001756	0.002507	-4.058039	20	1.0E-15	0.002407	-6.768968
8	50	0.000020	0.002307	-3.292370	5	1.0E-15	0.000894	-5.418727
9	50	0.003338	0.005876	-4.323313	50	1.0E-15	0.003537	-7.063629
10	2	0.000683	1.0E-15	-2.536195	6	1.0E-15	0.001354	-5.209892
11	27	0.000273	0.002037	-3.400871	2	1.0E-15	0.000232	-3.621913
12	15	1.0E-15	0.001360	-2.858040	2	1.0E-15	0.000388	-3.627276
13	1	0.001376	1.0E-15	-3.299379	5	1.0E-15	0.001400	-5.615434
14	50	0.000155	0.002203	-3.462888	9	1.0E-15	0.001244	-6.663925
15	2	0.001101	1.0E-15	-3.236978	4	1.0E-15	0.000538	-5.626682
16	2	0.000606	1.0E-15	-2.116053	2	1.0E-15	0.000259	-2.049299
17	1	0.001494	1.0E-15	-3.257039	50	1.0E-15	0.003223	-8.686453
18	10	1.0E-15	0.000849	-2.620518	3	1.0E-15	0.000453	-3.990432
19	50	0.000794	0.003123	-3.594939	28	1.0E-15	0.002714	-6.564674
20	23	1.0E-15	0.001432	-2.771172	7	0.000267	0.001293	-6.066984
21	2	0.001082	1.0E-15	-3.518809	2	0.000537	0.000150	-6.828282
22	50	1.0E-15	0.002388	-3.503018	50	1.0E-15	0.003290	-7.038840
23	28	0.000366	0.001975	-3.486375	16	1.0E-15	0.001983	-7.257413
24	50	0.000157	0.002846	-2.919208	50	0.000322	0.003934	-6.926059
25	1	0.001645	1.0E-15	-3.548313	3	1.0E-15	0.000453	-4.167397
26	4	1.0E-15	0.000126	-0.736850	1	0.000374	1.0E-15	-1.729943
27	19	1.0E-15	0.001013	-3.045921	2	0.000490	1.0E-15	-5.422874
28	2	0.001082	1.0E-15	-3.518809	7	1.0E-15	0.001078	-6.270861
29	20	1.0E-15	0.001514	-3.136344	2	1.0E-15	0.000244	-3.185800
30	4	0.000456	1.0E-15	-2.667301	2	0.000470	0.000172	-4.747534
...			...				...	
m	7.42	0.000721	0.000592	-2.964852	7.61	0.000064	0.001050	-5.299133
σ	7.87	0.000590	0.000834	0.768922	6.91	0.000164	0.000802	1.468500
$Q_{.5}$	3.5	0.000711	0.000113	-3.212236	5	1.0E-15	0.000873	-5.621058

Table 6.3: Extract of the estimated parameter for the 2R dataset. In the first columns denoted with 'without WGD' the results using the normal estimation procedure are shown. In the next columns 'with WGD' two WGD 690 mya and 590 mya were taken into account in the estimation procedure. m denotes the mean, σ the standard deviation, and $Q_{.5}$ the median for the entire sample, respectively.

away from the values used in the simulation, not even the right ratio of the rates could be found. In contrast, for the tMRHF2R-estimation the averaged estimated rates have almost the same ratio than the 'true' rates. Nevertheless, both rates are again underestimated and received values which are about half of the values of the 'true' rates.

Taken as a whole, the results got better using the strategy which includes WGD, when the evolution of the gene families analyzed involved WGDs. Using this approach, it might be possible to get better results for the Inparanoid dataset including mouse, rat, human, and fly. If two rounds of WGD have happened at the origin of vertebrates, the evolution of gene families of these species could be influenced by these WGDs and it might be possible to find evidence for these events in the data. However, the application of this approach to real data is complicated. First, the gene trees for all families would be needed. Furthermore, these gene trees must be analyzed with respect to their topology. If no topology similar to the one of tree tMRHF2R would be found, it is unclear how to divide the data to get the structure of the tMRHF2R data. For some of the biological gene families from the Inparanoid dataset the gene trees were checked, but unfortunately a tree topology similar the tMRHF2R was not found.

6.2.2 2R hypothesis in reality

Two rounds of WGD at the origin of vertebrates would suggest that many gene families would have four times as many members in vertebrates than in fruit fly (4:1 rule) (Friedman & Hughes (2001)). Only a few years ago, estimates for the total number of genes were in the range of ∼70,000 for human and ∼20,000 for invertebrates (Fields *et al.* (1994)). Furthermore gene families were found, where a 4:1 relation between human and fly could be detected. These observations supported the hypothesis of two rounds of WGD, but under the assumption that not many gene deletions had occurred afterwards.

According to this, one way to find evidence for the 2R hypothesis is the analysis of the distribution of the gene family ratios between vertebrates and invertebrates. These ratios were e.g. extensive examined in a study of Friedman & Hughes (2001). Based on this study the ratios between mouse and fly, rat and fly, and human and fly in our gene families from the Inparanoid database were analyzed. The resulting distributions as well as the distribution from Friedmann & Hughes for human and fly are summarized in table 6.4.
In both analysis and for all different vertebrates the most frequent ratio is one (1:1). That means in most gene families the same number of gene copies was found for the vertebrate and fly. The next frequent ratio is 2 (2:1) which is twice as many gene copies in vertebrates than in fruit fly. The 4:1 relation between vertebrates and fly was only found in very few gene families (2.5%, 3.5%, 2.8%, 4.9%).

These results contradict the 4:1 rule. So either there were not two rounds of WGD or the WGDs must have been followed by extensive gene loss. The latter explanation is very likely since higher rates for gene deletion compared to gene duplication have been reported (Lynch & Conery (2003), Cotton & Page (2005)). Furthermore, our global gene

ratio		Inparanoid dataset							
		mouse : fly		rat : fly		human : fly		human : fly*	
		freq	%	freq	%	freq	%	freq	%
< 0.2	< 1:5	10	0.6	8	0.5	13	0.8	-	-
0.2	1:5	2	0.1	2	0.1	2	0.1	7	0.3
0.25	1:4	6	0.4	7	0.4	7	0.4	12	0.4
0.33	1:3	27	1.7	23	1.4	25	1.5	16	0.6
0.5	1:2	98	6.0	95	5.8	95	5.8	94	3.4
1	1:1	965	59.4	967	59.5	966	59.4	1180	42.7
2	2:1	328	20.2	281	17.3	311	19.1	489	17.7
3	3:1	105	6.5	110	6.8	118	7.3	265	9.6
4	4:1	41	2.5	57	3.5	46	2.8	136	4.9
5	5:1	18	1.1	24	1.5	12	0.7	78	2.8
> 5	> 5:1	26	1.6	52	3.2	31	1.9	-	-
total		1626		1626		1626		2761	

Table 6.4: Distribution of ratios of the gene copy numbers of vertebrates and fruit fly. Gene families are from the Inparanoid dataset (section 5.2). * Family size ratios for human and fly from Friedman & Hughes (2001). In 'total' the number of gene families included in the particular analyzes is given.

deletion rate for the Inparanoid dataset was also estimated to be slightly higher than our global duplication rate. Such extensive gene loss would make it impossible to find a 4:1 relation between vertebrates and invertebrates and makes comparisons of gene copy numbers between them an uninformative measure.

Friedman & Hughes (2001) analyzed the gene families which showed the 4:1 relation in more detail. They tested if the topology of the gene trees of these gene families was consistent with the 2R hypothesis. That is also called (AB)(CD) topology measure in literature. Only for 24% of the gene families the topology was found to be equal to the one predicted by the 2R hypothesis (table 2 from Friedman & Hughes (2001)). That corresponds to our observation for the Inparanoid dataset. But, since the accuracy of the (AB)(CD) topology measure was often contested (Panopoulou & Poustka (2005), Dehal & Boore (2005)), it cannot deliver true evidence whether or not two round of WGD have happened.

6.2.3 Discussion and conclusion

It remains unclear what events have shaped the evolution of gene families in particular, especially with regard to large-scale duplications. Whether two separate rounds of WGD, only one single WGD, or no WGD at all had happened, stay open. But the simple way we included WGDs in the parameter estimation is not applicable for biological data. However, the study showed that large-scale duplications would influence the estimation of the rates and are likely to increase to failure rate of the method.

Hence, it would be desirable to extend the model for the possibility to include large-scale duplications. In this context, it will be also necessary to think about different mechanisms for gene loss and their influence on the data.

6.3 Heterogeneous rates

In our BD model as well as in the models used in publications, there is another assumption which might be problematic. Both rates, the duplication rate λ and the deletion rate μ, are assumed to be equal over time and equal for all branches in the species tree. It is questionable if the rates did not change over time, especially if the time interval considered is very long.

For the biological data studies on the tree tMRHF a dataset from the Inparanoid database was used. A more detailed analysis of the gene families of this Inparanoid dataset, led to a couple of gene families where the assumption of equal rates over the tree seems violated. A selection of such gene families is given in table 6.5.

In the first gene family YGL123W the gene copy numbers (gcn) of mouse, rat, and human are on average 14 times higher than the gcn of fly. But there are also gene families where the opposite can be found. For example in gene family YPR186C the gcn of fly is 10 times higher than the ones of mouse, rat, and human. Further gene families were found where mouse and rat together have much more gcn than the rest (YOL127W, YGR217W), or especially one of the vertebrates has more gcn than the remaining species (mouse: YOR113W, rat: YFR032C-A, human: YJL056C).

For such gene families it might be necessary to allow for changes in the rates. That could be changes along the entire tree according to a given distribution for one or both of the rates, as it is known from sequence evolution models. In this context, the Γ distribution is most commonly used for modeling rate heterogeneity. See Swofford *et al.* (1996) for an

ID	M	R	H	F	failed	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	*logL*
YGL123W	**12**	**16**	**13**	1	N	16	0.0035759066	0.0043326448	-5.0017160881
YOL127W	**16**	**17**	10	1	N	26	0.0046252698	0.0058266855	-5.1913668303
YGR217W	**14**	**14**	1	5	Y	100	0.0084613969	0.0109811531	-6.5907788274
YOR113W	**40**	4	5	1	N	1	0.0509805566	0.0507347527	-11.7948486961
YOR120W	11	**21**	1	4	Y	50	0.0083971881	0.0100811530	-8.0666116282
YLL045C	12	**38**	7	1	N	1	0.0264119730	0.0256537190	-10.5584561914
YFR032C-A	15	**53**	6	1	N	1	0.0456930789	0.0449455123	-11.6947151433
YJL056C	299	271	**419**	2	n.s.	n.s.	n.s.	n.s.	n.s.
YIL172C	1	1	1	**10**	N	1	0.0016849379	1.0E-15	-2.3623953445
YPR186C	2	2	3	**19**	N	1	0.0024071059	1.0E-15	-3.6475623860

Table 6.5: Selected gene families from the Inparanoid dataset. Given are the identifier(ID), the gene copy number for mouse(M), rat(R), human(H), and fly(F). Further the estimated ancestral gene numbers $\widehat{\alpha}$, the duplication rates $\widehat{\lambda}$, the deletion rates $\widehat{\mu}$, and the corresponding log-likelihood values $logL = log(L_{tMRHF}(\widehat{\alpha}, \widehat{\lambda}, \widehat{\mu}))$ are given. 'n.s.' means 'not specified'. For gene family YJL056C no estimates could be inferred due to the enormous running time.

overview about rate heterogeneity for substitution models and the work of Ziheng Yang for more details (e.g. Yang (1993), Yang *et al.* (1994), Yang (1994)).

It is also imaginable to add only one secondary duplication rate or one secondary deletion rate to branches where a higher or lower rate can be expected with respect to the global rate. For the gene families from table 6.5 this would imply e.g. a potentially higher duplication rate on the branch from the MRCA to fly for gene family YPR186C. In contrast, for the gene family YGL123W a higher deletion rate on the branch to fly could be expected or a higher duplication rate on the remaining branches leading to mouse, rat, and human, respectively. For all branches leading to species where the gcn is bold marked in table 6.5 a secondary, higher duplication rate λ_2 could be added to the model to explain the observed pattern of gcn.

But, since only few data is used yet to infer the three parameters α, λ, and μ for the current model, it will be problematic to estimate even more parameters. Surely it will be hard perhaps even impossible to infer meaningful estimates at all in the current framework. Nevertheless these observations might give an explanation why the method sometimes failed or obtained inaccurate estimates.

6.4 Usage of the gene trees

Right at the beginning of this thesis it was noticed that using only the information about the number of gene copies for each species neglects potentially valuable information. That is, e.g., information about the gene tree for a certain gene family. In the following, a little example will show how different the gene family evolution could be, although the gene copy numbers of the recent species in the end are the same.

In the study on real data a majority of gene families were found where the number of gene copies was the same for all species. It was further established that these gene families provide no information for the estimation of the duplication rate λ nor the deletion rate μ. As an example a gene family including the species mouse, rat, human, and fly was considered, all with exactly two copies of the gene.

In figure 6.3 (a) the evolution of the gene family according to the estimated parameters from our method is shown. For this gene family the ancestral gene copy number α was estimated to be two. That corresponds to duplication of the gene before the first speciation event, so that the MRCA had already two copies. Both rates, the duplication rate λ and the deletion rate μ, were estimated to be zero. That means that no duplication or deletion event occurred since the time of the MRCA. Thus, in the gene tree two perfect species trees one for each gene copy are found.

In fact there are infinitely many evolutionary scenarios possible, which could result in the same numbers of gene copies for the species. Three of them are also displayed in figure 6.3 (b)-(d). It might be the case that only gene duplications beside speciation events happened as shown in trees (b) and (c). It is also possible that additionally gene deletions

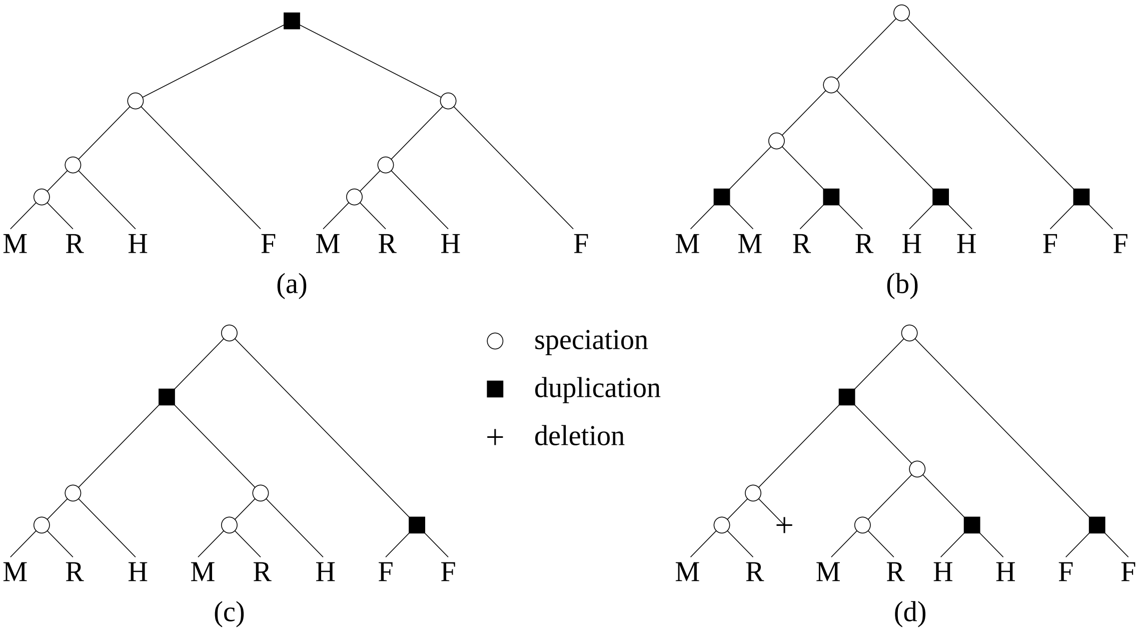


Figure 6.3: Schematic illustration of different ways in evolution of a gene family based on the species tree tMRHF. Different gene trees would be the consequence. For all gene trees the pattern of gene copies would be the same: 'M-R-H-F' = '2-2-2-2'. Further explanation of the individual gene trees are in text.

happened as for the tree (d). Using the information of these different gene trees should result in different estimates for α, λ, and μ since the evolutionary history of the gene family would have been quite different depending on the gene tree. But when using our current method, the estimates for the gene families with this pattern of gene copy numbers would be equal, independent of the underlying gene tree.

As hopefully has become clear, the usage of only numbers of gene copies has the big disadvantage that essential available information is not taken into account. From my point of view the next step should be the development of a model which makes use of the gene tree information as well. This could probably improve the quality and accuracy of the estimates.

6.5 Conclusion

The evolutionary history of a gene family can have many different facets, including duplications and deletions of single genes, large-scale duplications, like chromosome duplications or WGD, etc. Depending on the data, some processes might be more likely to have happened than others. The BD model we used here can be seen as a beginning to model gene duplications and deletions. It will be a challenge to develop more complex models to give consideration to different evolutionary processes. Such models will for sure improve the estimation of duplication and deletion rates.

Summary

Gene and genome duplications, and in this context also gene deletions, have played a major role in the evolution of genomes. Therefore models describing these processes are needed, to understand and reconstruct evolutionary events. We introduced a birth and death model describing duplications and deletions respectively. This model in conjunction with a phylogenetic tree describing the evolutionary relationships of the considered species, can be used to model the change in gene family size over time.

Two estimations strategies were applied to infer the parameters of our BD model, namely the ancestral gene number α, the duplication rate λ, and the deletion rate μ. The first strategy was the method of moments. We showed that the MOM can be applied, but with several restrictions. So it was not possible to took the whole information the species tree provided into account. The overall time of the tree could be included in the estimation, while the tree topology could not. Furthermore the parameter space of the variables could not be restricted. Although, we evaluated the MOM in several simulation studies. Thereby a high failure rate was detected due to invalid values for the parameters, e.g. negative or complex numbers. Also the quality of the estimated parameters α, λ, and μ was poor. Thus, we rejected the MOM for the parameter estimation.

The second estimation strategy considered was the maximum likelihood method. We showed how the likelihood function of a phylogenetic tree assuming a BD model can be computed. By maximizing this likelihood function for specific gene copy numbers for the contemporary species, the model parameters α, λ, and μ could be inferred. The ML method was also validated in several simulation studies. Unfortunately, there were also gene families were the method failed, but in contrast to the MOM this rate was much smaller. The estimates for the model parameters depend on the chosen parameter set in the simulation. In general, for datasets simulated with $\lambda > \mu$ the estimates were rather poor, whereas for dataset simulated with $\lambda < \mu$ the estimates were quite well. But, we found both rates λ and μ to be underestimated. This is due to the fact, that it is not possible to detect multiple duplications and deletions. Another artifact of the method is a high probability to estimate one rate being zero. We could show that this probability decreases with increasing number of contemporary species in the tree.

The ML method was also applied to biological gene family data of the Inparanoid database and the Ensembl database. The Inparanoid dataset included three mammalian species and fruit fly. The average of the estimated duplication rates for this dataset was 0.00026 gene^{-1} myr^{-1}, while the average of the estimated deletion rates was slightly higher with 0.0003 gene^{-1} myr^{-1}. The Ensembl dataset was a pure mammalian dataset, including five species. For this dataset an average duplication rate of 0.00088 gene^{-1} myr^{-1} and an average deletion rate of 0.00083 gene^{-1} myr^{-1} was estimated. These estimates are about one magnitude lower than previous reported rates. Furthermore, the failure rate in these analyzes was about twice as high as expected, which is assumed to be caused by model violations.

There are several problems with this model. On the one hand, it is not possible to include large-scale duplications or deletions and on the other hand dependencies between genes (e.g. for tandemly duplicated regions) are ignored. We have shown, that the occurrence of large-scale duplications in the evolution of gene families have a negative influence on the quality and the success of our method. The simple approach we used to include WGD in the analyzes turned out to be not suitable for the application to biological data. Another big problem is the traceability of gene deletions. If the DNA sequence is really lost, there is no hint for gene deletions at all and it is impossible to make statements about the amount of gene loss. If genes are deleted in form of pseudogenes, at least the existence of gene deletions is certain but it remains unclear how to involve this information into a mathematical model.

A software package was developed, to realize the simulation of gene family evolution along a phylogenetic tree and the estimation of duplication rate, deletion rate and ancestral gene number for the described applications. We have shown, that for the implementation high accuracy variable are essential.

BIBLIOGRAPHY

Acton, F. S. (1970) *Numerical Methods That Work*. Mathematical Association of America, Washington.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A. Galle, R. F. *et al.* (2000) The genome sequence of drosophila melanogaster. *Science*, **287**, 2185–2195.

Asenjo, A. B., Rim, J. & Oprian, D. D. (1994) Molecular determinants of human red/green color discrimination. *Neuron*, **12**, 1131–1138.

Bailey, N. T. J. (1964) *The elements of Stochastic Processes with applications to the natural sciences.* John Wiley & Sons, Inc., New York.

Bailey, W. J., Kim, J., Wagner, G. P. & Ruddle, F. H. (1997) Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Mol. Biol. Evol.*, **14**, 843–853.

Benton, M. J. & Ayala, F. J. (2003) Dating the Tree of Life. *Science*, **300**, 1698–1700.

Bertrand, S., Brunet, F., Escriva, H., Parmentier, G., Laudet, V. & Robinson-Rechavi, M. (2004) Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine system. *Mol. Biol. Evol.*, **21**, 1923–1937.

Blattner, F. R., Plunkett 3rd, G., Bloch, C. A., Perna, N. T., Burland, V. *et al.* (1997) The complete genome sequence of escherichia coli K-12. *Science*, **277**, 1453–1474.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S. & Van de Peer, Y. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.*, **7**, R43.

Bohn, W. F. & Flik, T. (2005) Zeichen- und Zahlendarstellungen. In Rechenberg, P. & Pomberger, G. (eds.), *Informatik-Handbuch*, Hanser, 4. bearb., aktualisierte und erweiterte Aufl., München/Wien.

Borenstein, E., Shlomi, T., Ruppin, E. & Sharan, R. (2006) Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. *Nucl. Acid. Res.*, **35**, e7.

Brent, R. P. (1973) *Algorithms for Minimization without Derivatives.* Prentice Hall, Englewood Cliffs, New Jersey, USA.

Chen, K., Durand, D. & Farach-Colton, M. (2000) Notung: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.*, **7**, 429–447.

Cotton, J. A. & Page, R. D. M. (2002) Going nuclear: gene family evolution and vertebrate phylogeny reconciled. *Proc. R. Soc. Lond. B*, **7**, 1555–1561.

Cotton, J. A. & Page, R. D. M. (2005) Rates and patterns of gene duplication and loss in the human genome. *Proc. R. Soc. Lond. B*, **272**, 277–283.

Cowlishaw, M. F. (2003) Decimal floating-point: Algorism for computers. In *Proceedings of the 16th IEEE Symposium on Computer Arithmetics (ARITH'03)*, pp. 104– 111.

Csűrös, M. & Miklós, I. (2006) A probabilistic model for gene content evolution with duplication, loss, and horizontal trannsfer. In *Proceedings of the 10th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2006)*, Lecture Notes in Computer Science, pp. 206–220, Springer, Berlin.

Cuyt, A., Verdonk, B., Becuwe, S. & Kuterna, P. (2001) A remarkable example of catastrophic cancellation unraveled. *Computing*, **66**, 309–320.

Dehal, P. & Boore, J. L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, **3**, e314.

Demuth, J., De Bie, T., Stajich, J., Cristianini, N. & Hahn, M. (2006) The evolution of mammalian gene families. *PloS ONE*, **1**.

Duret, L., Perrière, G. & Gouy, M. (1999) Hovergen: database and software for comparative analysis of homologous vertebrate genes. In Letovsky, S. (ed.), *Bioinformatics Databases and Systems.*, Kluwer Academic Publishers, Boston, MA, pp. 1329.

Edwards, A. W. F. (1992) *Likelihood.* The John Hopkins University Press, Baltimore.

Enright, A. J., Van Dongen, S. & Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Eulenstein, O., Mirkin, B. & Vingron, M. (1997) Comparison of annotation duplication, tree mapping, and copying as methods to compare gene trees with species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences*, **37**, 71–93.

Ewens, W. J. & Grant, G. R. (2001) *Statistical Methods in Bioinformatics: An Introduction.* Springer Verlag, New York, USA.

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2004) *Statistik - Der Weg zur Datenanalyse.* Springer Verlag, Berlin.

Feller, W. (1950) *An Introduction to Probability Theory and its Applications.* John Wiley & Sons, Inc., New York.

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

Fields, C., Adams, M. D., White, O. & Venter, J. C. (1994) How many genes in the human genome? *Nat. Genet.*, **7**, 345–346.

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. Royal Soc. London A*, **222**, 309–368.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F. & Kerlavage, A. R. e. a. (1995) Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, **269**, 496–498,507–512.

Forster, O. (2004) *ARIBAS interpreter for Arithmetic, Version 1.50.* Munich, available from http://www.mathematik.uni-muenchen.de/ forster/sw/aribas.html.

Friedman, R. & Hughes, A. L. (2001) Pattern and timing of gene duplication in animal genomes. *Genome Res.*, **11**, 1842–1847.

Friedmann, R. & Hughes, A. L. (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.*, **20**, 154–161.

Garcia-Fernàndez, J. (2005) Hox, ParaHox, ProtoHox: facts and guesses. *Heredity*, **94**, 145–152.

Garcia-Fernàndez, J. & Holland, P. W. H. (1994) Archetypal organization of the amphioxus Hox gene cluster. *Nature*, **370**, 563–566.

Genome Project Database (2007) `http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj`.

Glazko, G. V. & Nei, M. (2003) Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.*, **20**, 424–434.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B. & Feldmann, H. e. a. (1996) Life with 6000 genes. *Science*, **274**, 546, 563–7.

Goldberg, D. (1991) What every computer scientist should know about floating-point arithmetic. *ACM Comput. Surveys*, **23**, 5–48.

Goldman, N. (1990) Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of DNA substitution and to parsimony analyses. *Syst. Zool.*, **39**, 345–361.

Graur, D. & Li, W.-H. (2000) *Fundamentals of Molecular Evolution.* 2nd edition, Sinauer Associates, Sunderland, Massachusetts.

Grimmett, G. R. & Stirzaker, D. R. (2001) *Probability and Random Processes.* Oxford University Press Inc., New York.

Gu, X. & Huang, W. (2002) Testing the parsimony test of genome duplications: A counterexample. *Genome Res.*, **12**, 1–2.

Gu, X., Wang, Y. & Gu, J. (2002a) Age-distribution of human gene families showing equal roles of large and small-scale duplications in vertebrate evolution. *Nature Genetics*, **31**, 205–209.

Gu, Z., Cavalcanti, A., Chen, F.-C., Bouman, P. & Li, W.-H. (2002b) Extent of gene duplication in the genomes of drosophila, nematode, and yeast. *Mol. Biol. Evol.*, **19**, 256–262.

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.*, **15**, 1153–1160.

Haldane, J. B. S. (1932) *The causes of evolution.* Longmans and Green, London.

Hardison, R. C. (2006) Globin genes: Evolution. In *Encyclopedia of Life Sciences*, John Wiley & Sons, Ltd: Chichester.

Harisson, P. M., Echols, N. & Gerstein, M. B. (2001) Digging for dead genes: An analysis of the characteristics of the pseudogene population in the *C. elegans* genome. *Nucleic Acids Res.*, **29**, 818–830.

Harisson, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M., Bertone, P., Echols, N., Johnson, T., & Gerstein, M. B. (2002) Molecular fossils in the human genome: Identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.*, **12**, 272–280.

Hedges, S. B. (2002) The origin and evolution of model organisms. *Nat. Rev. Genet.*, **3**, 838–849.

Hennessy, J. L. & Patterson, D. A. (1996) *Computer Architecture - A quantitative approach.* Morgan Kaufmann.

Hoegg, S. & Meyer, A. (2005) Hox clusters as models for vertebrate genome evolution. *Trends Genet.*, **21**, 421–424.

Holland, P. W. H. (1997) Vertebrate evolution: something fishy about Hox genes. *Curr. Biol.*, **7**, R570–R572.

Holland, P. W. H., Garcia-Fernàndez, J., Williams, N. A. & Sidow, A. (1994) Gene duplications and the origin of vertebrate development. *Dev. Suppl.*, **43**, 125–133.

Horner, W. G. (1819) A new method of solving numerical equations of all orders, by continuous approximation. *Phil. Trans. R. Soc. Lond. B.*

Hughes, A. L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.*, **48**, 565–576.

Hughes, A. L. (2005) Gene duplication and the origin of novel proteins. *Proc. Natl. Acad. Sci. USA*, **25**, 8791–8792.

Human Genome Project (2007) `http://www.ornl.gov/sci/techresources/Human\_Genome/home.shtml.`

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lahres, H. (1964) *Einführung in die diskreten Markoff-Processe und ihre Anwendungen.* Vieweg Verlag, Braunschweig.

Li, W.-H. (1997) *Molecular Evolution.* Sinauer Associates, Sunderland, Massachusetts.

Long, M. & Thornton, K. (2001) Gene duplication and evolution. *Science*, **293**, 1551a.

Lynch, M. & Conery, J. S. (2000) The evolutionary fate and consequences of duplicated genes. *Science*, **290**, 1151–1155.

Lynch, M. & Conery, J. S. (2001) Gene duplication and evolution: response. *Science*, **293**, 1551a.

Lynch, M. & Conery, J. S. (2003) The evolutionary demography of duplicate genes. *J. Struc. Func. Genom.*, **3**, 35–44.

Lynch, M. & Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459–473.

Lynch, M., O'Hely, M., Walsh, B. & A., F. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics*, **159**, 1789–1804.

Málaga-Trillo, E. & Meyer, A. (2001) Genome duplications and accelerated evolution of Hox genes and cluster architecture in Teleost Fishes. *Amer. Zool.*, **41**, 676–686.

Meyer, A. & Málaga-Trillo, E. (1999) Vertebrates genomics: more fishy tales about Hox genes. *Curr. Biol.*, **9**, R210–R213.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520 –562.

Nagl, W. (1990) Polyploidy in differentiation and evolution. *Int. J. Cell Cloning*, **8**, 216–223.

Nei, M., Rogozin, I. B. & Piontkivska, H. (2000) Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc. Natl. Acad. Sci. USA*, **97**, 10866–10871.

Novozhilov, A. S., Karev, G. P. & Koonin, E. V. (2006) Biological applications of the theory of birth-and-death processes. *Briefings in Bioinformatics*, **7**, 70–85.

Nowak, M. A., Boerlijst, M. C., Cooke, J. & Maynard Smith, J. (1997) Evolution of genetic redundancy. *Nature*, **388**, 167–171.

O'Brien, K. P., Remm, M. & Sonnhammer, E. L. L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucl. Acids Res.*, **33**, D476–D480.

Ohno, S. (1970) *Evolution by gene duplication.* (eds. George Allen and Unwin London), Springer-Verlag, London.

Page, R. D. M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, **43**, 58–77.

Page, R. D. M. (2000) Extracting species trees from complex gene trees: Reconciled trees and vertebrate phylogeny. *Mol. Phylogenet. Evol.*, **14**, 89–106.

Page, R. D. M. & Charleston, M. A. (1997) From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.*, **7**, 231–240.

Panopoulou, G. & Poustka, A. J. (2005) Timing and mechanism of ancient vertebrate genome duplications the adventure of a hypothesis. *Trends Genet.*, **21**, 559–567.

Pearson, K. (1894) Contributions to the mathematical theory of evolution. *Phil. Trans. Royal Soc. London A*, **185**, 71–110.

Piatigorsky, J. & Wistow, G. J. (1991) The recruitment of crystallins: new functions precede gene duplication. *Science*, **252**, 1078–1079.

Piontkivska, H., Rooney, A. P. & Nei, M. (2002) Purifying selection and birth-and-death evolution in the histone h4 gene family. *Mol. Biol. Evol.*, **19**, 689–697.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1992) *Numerical Recipes in C: The Art of Scientific Computing.* 2nd edition, Cambridge University Press, Cambridge.

Purvis, A. (1995) A composite estimate of primate phylogeny. *Phil. Trans. R. Soc. Lond. B*, **348**, 405–421.

Ralston, A. & Rabinowitz, P. (1978) *A First Course in Numerical Analysis, 2nd ed.* McGraw-Hill, New York.

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D. & Liberles, D. (2006) Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J. Exp. Zool. (Mol. Dev. Evol.)*, **306**.

Skrabanek, L. & Wolfe, K. H. (1998) Eukaryote genome duplication - where's the evidence? *Curr. Opin. Genet. Dev.*, **8**, 694–700.

Strachan, T. & Read, A. (2003) Our place in the tree of life. In Strachan, T. & Read, A. (eds.), *Human Molecular Genetics*, 3rd edition, Garland Science, Taylor & Francis Group, London.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996) Phylogeny inference. In Hillis, D. M., Moritz, C. & Mable, B. K. (eds.), *Molecular Systematics*, 2nd edition, pp. 407–514, Sinauer Associates, Sunderland, Massachusetts.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, **408**, 796–815.

The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode c. elegans: A platform for investigating biology. *Science*, **282**, 2012–8.

The PARI-Group (2000) *PARI/GP, Version 2.1.0.* Bordeaux, available from http://www.parigp-home.de/.

Waddell, P. J., Cao, Y., Hasegawa, M. & Mindell, D. P. (1999) Assessing the cretaceous superordinal divergence times within birds and placental mammals by using whole mitochondrial protein sequences and an extended statistical framework. *Syst. Biol.*, **48**, 119–137.

Wagner, A. (2002) Selection and gene duplication: a view from the genome. *Genome Biol.*, **3**, 1012.1–1012.3.

Wang, Y. & Gu, X. (2000) Evolutionary patterns of gene families generated in the early stage of vertebrates. *J. Mol. Evol.*, **51**, 88–96.

Wolfe, K. H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.*, **2**, 333–341.

Wolfram Research, Inc. (2005) *Mathematica Edition: Version 5.2.* Wolfram Research, Inc., Champaign, Illinois.

Wray, G. A. (2001) Dating branches on the Tree of Life using DNA. *Genome Biol.*, **3**, 0001.1–0001.7.

Yang, Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, **10**, 1396–1401.

Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximative methods. *J. Mol. Evol.*, **39**, 306–314.

Yang, Z., Goldman, N. & Friday, A. E. (1994) Comparison of models for nucleotide substitution used in maximum–likelihood phylogenetic estimation. *Mol. Biol. Evol.*, **11**, 316–324.

Yokoyama, S. & Yokoyama, R. (1989) Molecular evolution of human visual pigment genes. *Mol. Biol. Evol.*, **6**, 186–197.

Zhang, J. (2003) Evolution by gene duplication: an update. *TREE*, **18**, 292–298.

Zhang, L., Gaut, B. S. & Vision, T. J. (2001) Gene duplication and evolution. *Science*, **293**, 1551a.

APPENDIX

A

Software

DupliDeli is a computer program for the estimation of gene family specific duplication and deletion rates, as well as the ancestral gene number. The program uses a stochastic birth and death process to model changes in gene family size along a user defined species tree. Based on the number of gene family members and the phylogenetic tree of the species considered, DupliDeli returns the maximum likelihood estimates for the ancestral gene number, the duplication rate, and the deletion rate specific for a gene family.
Furthermore, DupliDeli provides the opportunity to simulate gene family data for a given species tree, a fix number of ancestral genes, and specified rates for gene duplication and deletion. Supported features are:

- Simulation of gene family data for given $\mathcal{T}$, α, λ, and μ
- Computation of the likelihood of a tree for given $\mathcal{T}$, α, λ, μ, and gene family data
- Estimation of λ and μ for given $\mathcal{T}$, fixed α, and gene family data
- Estimation of α, λ, and μ for given $\mathcal{T}$ and gene family data
- Estimation of α, λ, and μ for given $\mathcal{T}$ and a set of gene family data

The species tree $\mathcal{T}$ have to be provided in Newick format with branch length, the ancestral gene number α must be a natural number, whereas the duplication rate λ and the deletion rate μ are real numbers. The gene family data contains the number of family members for each species included in the tree $\mathcal{T}$. For the estimation initial ranges of values for α, λ, and μ can be specified. For all applications the precision can be changed, if required.

DupliDeli is written in the object-oriented, platform independent programming language Java and is available from `http://www.cibiv.at/~andrea/duplideli/`.

APPENDIX

B

Deriving the probability distribution for the BD process

The equations 2.4 can be solved successively but this can lead to considerable difficulties. It is also possible to use the methods of generating functions. The probability generating function $P(t,x)$ for the probability distribution $p_i(t)$ is defined as

$$P(t,x) = \sum_{i=0}^{\infty} p_i(t) \cdot x^i \tag{B.1}$$

The initial conditions for $p_i(t)$ can be converted for $P(t,x)$ to an equivalent form:

$$\begin{aligned} &p_0(0) = 1 \;\Leftrightarrow\; P(0,x) \equiv 1, \quad p_1(0) = 1 \;\Leftrightarrow\; P(0,x) \equiv x \\ &general : p_i(0) = 1 \;\Leftrightarrow\; P(0,x) \equiv x^i;\; i = 0, 1, \ldots \end{aligned} \tag{B.2}$$

and the partial derivative of the generating function are described by

$$\frac{\partial P(t,x)}{\partial t} = \sum_{i=0}^{\infty} p'_i(t)\, x^i \qquad \frac{\partial P(t,x)}{\partial x} = \sum_{i=0}^{\infty} i \cdot p_i(t)\, x^{i-1} \tag{B.3}$$

Multiply the first equation of 2.4 with x^i, summation over all values of i and using eq. B.3 leads to a partial differential equation for $P(t,x)$

$$\begin{aligned}
p'_i(t) &= -(\lambda+\mu)\, i\, p_i(t) + \lambda(i-1)\, p_{i-1}(t) + \mu(i+1)\, p_{i+1}(t) \quad \mid \cdot x^i \\
p'_i(t)\, x^i &= -(\lambda+\mu)\, i\, p_i(t)\, x^i + \lambda(i-1)\, p_{i-1}(t)\, x^i + \mu(i+1)\, p_{i+1}(t)\, x^i \\
\textstyle\sum_{i=0}^{\infty} p'_i(t) x^i &= -(\lambda+\mu) \textstyle\sum_{i=0}^{\infty} i p_i(t) x^i + \lambda \textstyle\sum_{i=0}^{\infty} (i-1) p_{i-1}(t) x^i + \mu \textstyle\sum_{i=0}^{\infty} (i+1) p_{i+1}(t) x^i \\
\underbrace{\sum_{i=0}^{\infty} p'_i(t) x^i}_{\frac{\partial P(t,x)}{\partial t}} &= -(\lambda+\mu) x \underbrace{\sum_{i=0}^{\infty} i p_i(t) x^{i-1}}_{\frac{\partial P(t,x)}{\partial x}} + \lambda x^2 \underbrace{\sum_{i=0}^{\infty} (i-1) p_{i-1}(t) x^{i-2}}_{\frac{\partial P(t,x)}{\partial x}} + \mu \underbrace{\sum_{i=0}^{\infty} (i+1) p_{i+1}(t) x^i}_{\frac{\partial P(t,x)}{\partial x}}
\end{aligned}$$

$$\begin{aligned}\tfrac{\partial P(t,x)}{\partial t} &= (-(\lambda+\mu)x+\lambda x^2+\mu)\tfrac{\partial P(t,x)}{\partial x}\\ \tfrac{\partial P(t,x)}{\partial t} &= (x-1)(\lambda x-\mu)\tfrac{\partial P(t,x)}{\partial x}\end{aligned}$$

This is a linear homogeneous partial differential equation

$$\frac{\partial P(t,x)}{\partial t} - (x-1)(\lambda x-\mu)\,\frac{\partial P(t,x)}{\partial x} = 0 \tag{B.4}$$

which can be solved by integration. The integration of the partial differential equation is equivalent to the integration of the corresponding so called characteristic system:

$$\begin{aligned}\tfrac{dt}{1} &= \tfrac{dx}{-(x-1)(\lambda x-\mu)}\\ dt &= \tfrac{-dx}{(x-1)(\lambda x-\mu)}\end{aligned} \tag{B.5}$$

For $\lambda \neq \mu$ the integration of both sides of eq. B.5 with the arbitrary constant c leads to

$$\begin{aligned}&t + C = -\tfrac{1}{\lambda-\mu}\, log\left(\tfrac{x-1}{\lambda x-\mu}\right)\\ &t + \tfrac{1}{\lambda-\mu}\, log\left(\tfrac{x-1}{\lambda x-\mu}\right) = -C\\ &(\lambda-\mu)t + log\left(\tfrac{x-1}{\lambda x-\mu}\right) = c \text{ with } c = (\lambda-\mu)(-C)\end{aligned} \tag{B.6}$$

and results in the general solution for $P(t,x)$ of the form:

$$P(t,x) = \Psi\left((\lambda-\mu)t + log\left(\frac{x-1}{\lambda x-\mu}\right)\right) \tag{B.7}$$

where Ψ is a continuously differentiable function. The assignment of Ψ results from the information of the initial conditions:

$$\begin{aligned}&p_k(0) = 1 \;\Leftrightarrow\; P(0,x) \equiv x^k\\ &P(0,x) = \Psi\left((\lambda-\mu)0 + log\left(\tfrac{x-1}{\lambda x-\mu}\right)\right) = \Psi\left(log\left(\tfrac{x-1}{\lambda x-\mu}\right)\right) = x^k\end{aligned} \tag{B.8}$$

The outcome of this is the following equation for Ψ

$$\Psi(y) = \frac{\mu e^y - 1}{\lambda e^y - 1} \text{ with } y = (\lambda-\mu)t + log\left(\frac{x-1}{\lambda x-\mu}\right)^k \tag{B.9}$$

By insert y in Ψ and after some transformations the explicit solution for the generating function become

$$P(t,x) = \left(\frac{(\mu e^{(\lambda-\mu)t} - \lambda)x \;+\; \mu(1-e^{(\lambda-\mu)t})}{\lambda(e^{(\lambda-\mu)t}-1)x \;+\; (\mu - \lambda e^{(\lambda-\mu)t})}\right)^k \tag{B.10}$$

To get the solution for $p_i(t)$ the generating function $P(t,x)$ has to be expand in a power series to x, also called taylor expansion according to the following equation:

$$P(t,x) = \sum_{i=0}^{\infty} a_i x^i = \sum_{i=0}^{\infty} \underbrace{\frac{1}{i\,!}\,\frac{\partial^{(i)}P}{\partial x^i}(t,0)}_{a_i \Leftrightarrow p_i(t)}\; x^i \tag{B.11}$$

The coefficients of the taylor expansion a_i arise from the partial derivatives of $P(t,x)$ at the position $x=0$, as shown below

$$\begin{aligned}
a_0 &= P(t,0) &&= \left(\tfrac{\mu(E-1)}{E\lambda-\mu}\right)^k \\
a_1 &= \tfrac{\partial^{(1)}P}{\partial x}(t,0) &&= k\left(\tfrac{\mu(E-1)}{E\lambda-\mu}\right)^{k-1}\left(\tfrac{\mu(E-1)\lambda(E-1)}{(E\lambda-\mu)^2} + \tfrac{\lambda-E\mu}{E\lambda-\mu}\right) \\
a_2 &= \tfrac{1}{2}\tfrac{\partial^{(2)}P}{\partial x^2}(t,0) &&= k\left(\tfrac{\mu(E-1)}{E\lambda-\mu}\right)^{k-1}\left(\tfrac{(\lambda(E-1))^2\,\mu(E-1)}{(E\lambda-\mu)^3} - \tfrac{\lambda(E-1)(E\mu-\lambda)}{(E\lambda-\mu)^2}\right) \\
& &&\quad +\tfrac{1}{2}k(k-1)\left(\tfrac{\mu(E-1)}{E\lambda-\mu}\right)^{k-2}\left(\tfrac{\mu(E-1)\lambda(E-1)}{(E\lambda-\mu)^2} + \tfrac{\lambda-E\mu}{E\lambda-\mu}\right)^2 \\
a_3 &= \tfrac{1}{6}\tfrac{\partial^{(3)}P}{\partial x^3}(t,0) &&= \dots
\end{aligned} \tag{B.12}$$

with $e^{(\lambda-\mu)t}$ denoted by E. The probability distribution $p_i(t)$ for the most general case $\lambda \neq \mu$ and $k > 1$ for $i > 0$ is then obtained by several manipulations

$$\begin{aligned}
a_0 &= p_0(t) &&= \Bigg(\underbrace{\frac{\mu(E-1)}{E\lambda-\mu}}_{A}\Bigg)^k \\
a_1 &= p_1(t) &&= k\Bigg(\underbrace{\frac{\mu(E-1)}{E\lambda-\mu}}_{A}\Bigg)^{k-1}\Bigg(\underbrace{\frac{\mu(E-1)}{E\lambda-\mu}}_{A}\;\underbrace{\frac{\lambda(E-1)}{E\lambda-\mu}}_{B} + \underbrace{\frac{E\mu+\lambda}{E\lambda-\mu}}_{1-A-B}\Bigg) \\
a_2 &= p_2(t) &&= \dots \\
&\dots \\
a_i &= p_i(t) &&= \sum_{j=0}^{\min(k,i)} \tbinom{k}{j}\tbinom{k+i-j-1}{k-1}\, A^{k-j}\, B^{i-j}\,(1-A-B)^j
\end{aligned} \tag{B.13}$$

APPENDIX

C

Estimated parameters for single gene families from simulated data

C.1 Method of moments

C.1.1 parameters: tMRHF, $\alpha = 5$, $\lambda = 0.0008$, $\mu = 0.0002$

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
1	7	7	11	9	7.71	0.00020482	0.000106299
2	8	8	8	5	19.4	-0.000313072	0.000682828
3	12	11	12	8	64.7	-0.000634639	0.00117816
4	13	14	14	7	0. - 5.36 ⅈ	0.000568068 + 0.000592436 ⅈ	-0.000245305 - 0.000994227 ⅈ
5	11	11	11	15	8.91	0.000258598	-0.0000417593
6	8	8	8	15	4.39	0.000714014	-0.0000919507
7	16	15	16	6	0. - 3.22 ⅈ	0.000904507 + 0.000515524 ⅈ	-0.000525053 - 0.00107114 ⅈ
8	6	6	6	8	5.96	0.000099871	0.0000116643
9	8	8	9	9	8.51	0.0000142079	0.0000155203
10	7	6	6	13	3.83	0.000735051	-0.00000900194
11	6	6	6	6	6.	indeterminate	indeterminate
12	9	8	8	10	8.56	0.0000505405	0.00002794
13	8	9	10	5	0. - 9.57 ⅈ	0.000103712 + 0.000608524 ⅈ	0.000284415 - 0.000978139 ⅈ
14	3	4	3	17	0. - 2.12 ⅈ	0.0014937 - 0.000450581 ⅈ	0.000325343 - 0.00203724 ⅈ
15	8	9	8	12	7.18	0.000256772	0.00000113158
16	6	6	6	7	6.17	0.0000218512	0.00000824821
17	8	8	9	12	7.18	0.000256772	0.00000113158
18	4	4	4	10	2.8	0.000774483	0.000094289
19	7	8	7	10	7.1	0.000149424	0.0000288949
20	6	7	8	12	5.56	0.0004582	0.0000599338
21	5	5	5	10	3.57	0.000565268	0.
22	7	7	9	12	6.79	0.000340699	0.0000838469
23	5	5	7	5	4.99	0.000115126	0.0000159642
24	8	7	10	8	7.72	0.000103921	0.0000366873
25	7	8	6	8	7.66	0.0000213375	0.000077117
26	2	2	3	6	2.61	0.000483692	0.000264065
27	14	14	15	7	0. - 4.8 ⅈ	0.00065003 + 0.000577049 ⅈ	-0.000316416 - 0.00100961 ⅈ
28	8	7	7	8	7.51	0.0000160059	0.0000176922
29	9	9	12	9	8.16	0.000170048	-0.0000103712
30	9	9	8	7	8.64	0.0000195334	0.0000666188
31	8	8	8	8	8.	indeterminate	indeterminate

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
32	5	4	5	4	4.52	0.0000257746	0.0000304728
33	8	8	8	7	7.85	0.00000566002	0.0000189389
34	9	10	8	7	8.79	0.0000585854	0.0000924638
35	13	14	13	9	0. - 18.9 𝕚	-0.0000637361 + 0.000655242 𝕚	0.000374637 - 0.000931421 𝕚
36	7	7	8	9	7.57	0.000056343	0.000032196
37	13	13	13	6	0. - 4.34 𝕚	0.000647674 + 0.000577462 𝕚	-0.000314303 - 0.0010092 𝕚
38	13	16	13	12	11.6	0.000155708	0.
39	9	9	9	12	8.16	0.000170048	-0.0000103712
40	11	12	9	9	10.2	0.0000854891	0.0000803885
41	12	12	10	9	11.7	0.0000384448	0.000127185
42	6	7	8	10	7.31	0.000167927	0.000108938
43	8	8	8	9	8.16	0.0000167104	0.00000612633
44	11	11	10	11	10.9	0.00000417184	0.0000135279
45	19	20	19	15	55.8	-0.000394959	0.000733916
46	8	9	7	4	0. - 6.88 𝕚	0.000204987 + 0.000596263 𝕚	0.000187264 - 0.0009904 𝕚
47	10	8	9	8	8.56	0.0000505405	0.00002794
48	11	11	12	12	11.5	0.000010625	0.0000113415
49	13	11	9	8	10.3	0.000180839	0.000182927
50	12	11	13	8	0. - 31.5 𝕚	-0.000302171 + 0.00062096 𝕚	0.000760732 - 0.000965703 𝕚
51	10	9	11	15	7.74	0.000380871	0.00000324644
52	10	10	10	9	9.85	0.00000457311	0.0000149519
53	4	4	4	10	2.8	0.000774483	0.000094289
54	9	9	10	10	9.51	0.0000127725	0.0000138229
55	9	9	7	10	10.1	0.00000153301	0.000145557
56	7	7	6	6	6.51	0.0000183235	0.0000205698
57	15	15	16	11	0. - 32.9 𝕚	-0.00025549 + 0.000660363 𝕚	0.000589331 - 0.0009263 𝕚
58	5	5	4	8	4.36	0.000300539	0.0000666972
59	5	4	6	7	5.98	0.0000771659	0.000162201
60	11	10	12	13	11.7	0.0000462805	0.0000645066
61	7	7	7	8	7.17	0.0000189382	0.00000703143
62	8	8	8	11	7.22	0.00018514	-0.00000840191
63	4	4	7	6	5.63	0.000132554	0.000203688
64	10	11	10	9	10.	0.000023399	0.0000272012
65	5	5	6	9	4.66	0.000334719	0.0000391773
66	8	8	8	12	6.36	0.000315896	-0.0000341781
67	13	14	11	14	15.2	-0.000016723	0.000142713
68	7	6	8	6	6.58	0.0000636408	0.0000379364
69	8	8	5	15	5.6	0.000831656	0.000352322
70	12	13	9	18	11.2	0.000453938	0.000302148
71	11	11	11	14	10.	0.000146296	-0.0000123206
72	7	7	7	6	6.85	0.00000642079	0.0000218524
73	10	9	7	8	8.79	0.0000585854	0.0000924638
74	11	11	11	7	0. - 14.6 𝕚	-0.0000407555 + 0.000658103 𝕚	0.000340569 - 0.00092856 𝕚
75	7	7	8	11	6.32	0.000278212	0.00000892056
76	10	8	7	10	9.9	0.0000412064	0.000165867
77	6	6	6	12	3.76	0.000663902	-0.0000324449
78	9	10	9	8	9.04	0.0000257746	0.0000304728
79	8	8	8	6	8.55	-0.0000121242	0.000119878
80	10	10	10	9	9.85	0.00000457311	0.0000149519
81	8	8	9	10	8.56	0.0000505405	0.00002794
82	12	12	12	13	12.2	0.0000113634	0.00000404135
83	9	8	8	4	0. - 5.53 𝕚	0.000305412 + 0.000611308 𝕚	0.0000327634 - 0.000975355 𝕚
84	6	7	7	13	4.14	0.000674792	-0.0000207048
85	14	11	11	13	12.1	0.0000741381	0.0000643343

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
86	12	13	12	5	0. - 3.56 ⅈ	0.000727144 + 0.000551442 ⅈ	-0.00036637 - 0.00103522 ⅈ
87	10	13	11	13	12.7	0.0000367633	0.000113915
88	17	16	17	10	0. - 7.95 ⅈ	0.000467 + 0.000619599 ⅈ	-0.000174766 - 0.000967064 ⅈ
89	7	9	5	8	12.5	-0.0000759031	0.000470503
90	9	8	8	6	9.19	-0.00000169028	0.000169989
91	12	12	15	6	0. - 4.69 ⅈ	0.000647516 + 0.000532686 ⅈ	-0.000236577 - 0.00105398 ⅈ
92	14	14	14	15	14.2	0.0000097961	0.00000345302
93	19	20	18	4	0. - 1.92 ⅈ	0.00127654 + 0.000395938 ⅈ	-0.000818072 - 0.00119072 ⅈ
94	7	7	7	6	6.85	0.00000642079	0.0000218524
95	7	8	8	8	7.85	0.00000566002	0.0000189389
96	7	6	7	10	6.06	0.000243368	0.0000285861
97	14	13	15	5	0. - 2.99 ⅈ	0.000887799 + 0.000508183 ⅈ	-0.000493392 - 0.00107848 ⅈ
98	9	9	9	7	9.5	-0.00000920177	0.000103388
99	9	9	9	7	9.5	-0.00000920177	0.000103388
100	14	14	16	7	0. - 4.66 ⅈ	0.000687672 + 0.000557321 ⅈ	-0.000328385 - 0.00102934 ⅈ

C.1.2 parameters: tMRHF, $\alpha = 10$, $\lambda = 0.0002$, $\mu = 0.0008$

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
1	6	5	5	4	5.08	0.0000431989	0.0000585823
2	5	5	5	7	4.99	0.000115126	0.0000159642
3	4	4	4	5	4.17	0.0000315589	0.0000125872
4	2	3	3	9	3.01	0.000939235	0.000590857
5	8	8	6	2	0. - 2.45 ⅈ	0.000665435 + 0.000521669 ⅈ	-0.000239494 - 0.00106499 ⅈ
6	6	6	7	8	6.58	0.0000636408	0.0000379364
7	5	4	5	6	5.08	0.0000431989	0.0000585823
8	7	6	8	6	6.58	0.0000636408	0.0000379364
9	7	7	7	5	7.61	-0.000016723	0.000142713
10	6	6	7	5	6.06	0.0000369929	0.000047626
11	4	4	4	6	4.02	0.000135916	0.0000232172
12	4	4	5	3	4.1	0.0000517685	0.0000760156
13	8	6	8	4	0. - 13.6 ⅈ	-0.000113172 + 0.000581587 ⅈ	0.000630651 - 0.00100508 ⅈ
14	7	7	5	9	8.06	0.0000837359	0.000225606
15	2	3	3	6	2.88	0.000392234	0.000196117
16	3	2	4	3	3.13	0.0000640468	0.000107992
17	2	2	2	2	2.	indeterminate	indeterminate
18	4	5	5	10	3.48	0.00062372	0.000072385
19	1	2	3	7	0. - 4.25 ⅈ	0.000612025 + 0.0000987828 ⅈ	0.000882434 - 0.00148788 ⅈ
20	8	8	7	6	7.66	0.0000213375	0.000077117
21	6	4	4	7	5.63	0.000132554	0.000203688
22	4	5	5	4	4.52	0.0000257746	0.0000304728
23	9	9	9	5	0. - 8.33 ⅈ	0.000132093 + 0.000642384 ⅈ	0.000173177 - 0.000944279 ⅈ
24	6	6	7	4	7.56	-0.0000189591	0.000257404
25	6	6	5	7	6.06	0.0000369929	0.000047626
26	5	5	4	6	5.08	0.0000431989	0.0000585823
27	5	5	6	8	5.24	0.000186543	0.0000492032
28	5	5	5	8	4.51	0.000253163	0.00000887628
29	3	4	5	7	4.86	0.000223822	0.000246627
30	4	5	5	7	4.86	0.000149064	0.0000706584
31	8	8	5	5	8.12	0.0000825278	0.000307505
32	9	9	10	7	10.1	0.00000153301	0.000145557
33	6	6	6	3	0. - 10. ⅈ	-0.000139499 + 0.00063637 ⅈ	0.000513921 - 0.000950293 ⅈ

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
34	5	7	7	3	0. - 7.3 ⅈ	0.0000934258 + 0.000574694 ⅈ	0.000378938 - 0.00101197 ⅈ
35	7	7	7	3	0. - 4.54 ⅈ	0.000306437 + 0.000615099 ⅈ	0.0000238032 - 0.000971564 ⅈ
36	6	8	8	5	8.22	0.0000393713	0.000238918
37	2	2	2	4	2.17	0.000212984	0.0000689547
38	7	7	8	5	8.33	-0.00000751744	0.000204524
39	5	5	6	4	5.08	0.0000431989	0.0000585823
40	8	7	6	3	0. - 4.87 ⅈ	0.000307769 + 0.000579537 ⅈ	0.0000960227 - 0.00100713 ⅈ
41	2	2	2	6	2.12	0.000597614	0.00024754
42	6	6	6	3	0. - 10. ⅈ	-0.000139499 + 0.00063637 ⅈ	0.000513921 - 0.000950293 ⅈ
43	3	3	3	8	2.51	0.000689071	0.000157424
44	2	2	1	5	2.54	0.000450324	0.000465707
45	3	3	3	9	2.5	0.000853786	0.000258506
46	2	2	2	8	2.94	0.000979572	0.000804874
47	2	2	3	3	2.54	0.0000431989	0.0000585823
48	5	5	5	5	5.	indeterminate	indeterminate
49	9	9	10	5	0. - 7.22 ⅈ	0.000229757 + 0.000624576 ⅈ	0.0000957401 - 0.000962087 ⅈ
50	6	6	7	6	6.17	0.0000218512	0.00000824821
51	4	4	4	6	4.02	0.000135916	0.0000232172
52	8	9	6	2	0. - 2.36 ⅈ	0.000722891 + 0.000492585 ⅈ	-0.000261153 - 0.00109408 ⅈ
53	9	8	8	7	8.05	0.000028682	0.0000346355
54	7	7	6	3	0. - 5.43 ⅈ	0.000207889 + 0.000611833 ⅈ	0.000149411 - 0.00097483 ⅈ
55	6	6	3	2	0. - 4.26 ⅈ	0.000296918 + 0.000495633 ⅈ	0.000298778 - 0.00109103 ⅈ
56	5	5	5	6	5.17	0.0000258231	0.0000099696
57	5	5	4	8	4.36	0.000300539	0.0000666972
58	6	6	5	3	13.9	-0.000273494	0.00075692
59	4	4	4	4	4.	indeterminate	indeterminate
60	7	7	5	5	6.27	0.0000640468	0.000107992
61	5	4	4	6	4.62	0.0000857734	0.0000584575
62	2	2	2	4	2.17	0.000212984	0.0000689547
63	7	6	6	7	6.51	0.0000183235	0.0000205698
64	6	6	7	6	6.17	0.0000218512	0.00000824821
65	5	5	5	7	4.99	0.000115126	0.0000159642
66	5	5	5	4	4.86	0.00000875997	0.0000315642
67	7	7	6	5	6.69	0.0000233564	0.0000915474
68	5	3	6	5	7.02	-0.0000450434	0.000350186
69	8	8	9	6	9.19	-0.00000169028	0.000169989
70	6	5	6	6	5.86	0.00000741374	0.0000258254
71	3	3	2	9	3.01	0.000939235	0.000590857
72	7	7	6	7	6.85	0.00000642079	0.0000218524
73	4	5	5	8	4.36	0.000300539	0.0000666972
74	6	6	6	4	6.71	-0.0000246156	0.000176526
75	7	7	7	4	0. - 36.9 ⅈ	-0.000626057 + 0.000624979 ⅈ	0.00116731 - 0.000961684 ⅈ
76	6	6	6	6	6.	indeterminate	indeterminate
77	4	4	4	8	3.28	0.00045625	0.0000312163
78	2	2	1	3	2.22	0.0000804961	0.000185364
79	3	3	4	3	3.18	0.0000405681	0.0000170307
80	7	7	8	4	0. - 12.2 ⅈ	-0.000119215 + 0.000625004 ⅈ	0.000519129 - 0.000961659 ⅈ
81	3	2	3	4	3.13	0.0000640468	0.000107992
82	5	5	5	4	4.86	0.00000875997	0.0000315642
83	5	5	5	5	5.	indeterminate	indeterminate
84	3	5	6	4	5.13	0.0000832412	0.000216216
85	3	4	4	4	3.86	0.0000106738	0.0000405826
86	3	3	1	5	0. - 5.2 ⅈ	0.0000903014 + 0.000476752 ⅈ	0.000645156 - 0.00110991 ⅈ
87	3	3	3	5	3.08	0.000165929	0.0000370019

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
88	5	4	4	7	4.34	0.000212984	0.0000689547
89	5	5	4	4	4.52	0.0000257746	0.0000304728
90	7	8	7	5	8.33	-0.00000751744	0.000204524
91	2	3	3	7	2.76	0.000579448	0.000268976
92	6	5	7	8	6.89	0.000070312	0.000129705
93	5	4	3	5	4.78	0.0000268318	0.000146431
94	4	3	4	3	3.53	0.0000323193	0.0000401094
95	3	3	3	8	2.51	0.000689071	0.000157424
96	8	8	7	6	7.66	0.0000213375	0.000077117
97	5	5	5	6	5.17	0.0000258231	0.0000099696
98	0	0	0	1	1.	0.	0.0014003
99	5	5	6	1	0. - 1.99 𝕚	0.000588017 + 0.00054194 𝕚	-0.00017911 - 0.00104472 𝕚
100	6	6	6	5	5.86	0.00000741374	0.0000258254

C.1.3 parameters: tMRHF, $\alpha = 20$, $\lambda = 0.0004$, $\mu = 0.0006$

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
1	15	15	16	18	14.8	0.0000833284	0.00000787579
2	11	11	12	12	11.5	0.000010625	0.0000113415
3	16	18	18	24	11.2	0.000446739	-0.000082774
4	14	15	16	9	0. - 8.48 𝕚	0.000390862 + 0.000615945 𝕚	-0.0000783138 - 0.000970718 𝕚
5	16	15	16	18	15.6	0.000057394	0.0000148785
6	16	16	16	17	16.2	0.00000860872	0.00000301404
7	15	14	15	18	13.5	0.000139142	-0.00000255887
8	16	18	14	14	14.2	0.000130832	0.0000404837
9	21	19	19	17	19.3	0.0000450769	0.0000621507
10	13	13	14	16	12.9	0.0000936699	0.0000101422
11	10	10	11	8	11.	0.00000340276	0.000127323
12	12	12	13	11	12.	0.0000197526	0.0000223899
13	20	20	22	9	0. - 3.82 𝕚	0.00096595 + 0.000506113 𝕚	-0.000585843 - 0.00108055 𝕚
14	18	17	15	10	0. - 8.38 𝕚	0.000463633 + 0.000594516 𝕚	-0.000124343 - 0.000992147 𝕚
15	22	23	23	25	22.5	0.0000412376	0.00000954854
16	16	16	17	17	16.5	0.00000747962	0.00000782757
17	16	14	15	10	0. - 14.8 𝕚	0.000118587 + 0.000639543 𝕚	0.000195452 - 0.00094712 𝕚
18	16	16	18	17	16.5	0.0000276875	0.000013481
19	10	9	10	19	4.29	0.000917033	-0.000122221
20	15	16	16	11	0. - 20.5 𝕚	-0.0000315946 + 0.00066246 𝕚	0.000319381 - 0.000924202 𝕚
21	13	13	13	16	12.	0.000128429	-0.0000129342
22	19	21	21	16	38.9	-0.000202723	0.000508862
23	15	17	14	26	6.37	0.000883718	-0.000165188
24	18	18	19	21	17.8	0.0000714974	0.0000056775
25	16	19	16	18	17.	0.0000554877	0.0000427025
26	20	18	23	22	24.5	0.0000129345	0.000182014
27	20	22	23	13	0. - 7.72 𝕚	0.000630525 + 0.00058395 𝕚	-0.000304885 - 0.00100271 𝕚
28	12	11	11	17	7.53	0.000452175	-0.0000797778
29	18	18	18	21	16.8	0.0000984829	-0.0000124755
30	10	10	11	17	6.	0.000597805	-0.000101686
31	15	13	15	15	15.4	-0.00000294363	0.0000567492
32	20	21	20	17	24.1	-0.0000431463	0.000172585
33	13	15	15	17	15.4	0.0000544265	0.0000821074
34	16	17	17	16	16.5	0.00000747962	0.00000782757
35	15	14	16	14	14.5	0.0000312185	0.000015497

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
36	20	19	21	18	19.6	0.0000293432	0.0000356079
37	18	21	21	11	0. - 6.13 𝕚	0.000713215 + 0.000558063 𝕚	-0.000360474 - 0.0010286 𝕚
38	21	21	23	11	0. - 5.01 𝕚	0.000849133 + 0.000541464 𝕚	-0.000497622 - 0.0010452 𝕚
39	18	17	18	25	9.92	0.000527102	-0.000155499
40	16	15	19	12	21.7	0.0000706681	0.000408541
41	15	15	16	20	12.	0.000271455	-0.0000505484
42	12	13	12	19	7.23	0.000550235	-0.00011749
43	13	14	15	15	14.6	0.0000126447	0.0000366701
44	15	16	18	15	14.8	0.0000833284	0.00000787579
45	19	19	18	20	19.	0.0000127725	0.0000138229
46	10	9	11	9	9.55	0.0000458183	0.0000246635
47	22	22	22	23	22.2	0.0000063131	0.00000218159
48	6	6	6	15	3.03	0.0010463	0.0000353624
49	15	16	18	23	11.	0.000454028	-0.000040902
50	16	14	15	13	14.7	0.000038091	0.0000494784
51	13	14	14	22	6.63	0.000703089	-0.000171197
52	18	18	16	16	17.1	0.0000271514	0.0000324215
53	21	22	23	14	0. - 9.25 𝕚	0.00053534 + 0.000613194 𝕚	-0.000243226 - 0.000973469 𝕚
54	11	9	10	9	9.55	0.0000458183	0.0000246635
55	18	17	19	15	19.4	0.00000798423	0.000127859
56	19	19	19	20	19.2	0.00000728433	0.00000253119
57	23	25	23	15	0. - 9.05 𝕚	0.000586525 + 0.000604212 𝕚	-0.00028756 - 0.000982451 𝕚
58	16	18	17	13	29.5	-0.000161457	0.000456908
59	11	10	12	8	13.4	-0.0000110024	0.000256359
60	16	16	17	14	16.8	0.00000586723	0.000072844
61	17	16	15	10	0. - 9.7 𝕚	0.000354812 + 0.000623752 𝕚	-0.0000508637 - 0.000962911 𝕚
62	19	22	18	13	0. - 13.9 𝕚	0.000327142 + 0.000583879 𝕚	0.0000628757 - 0.00100278 𝕚
63	12	13	13	17	10.2	0.000265804	-0.0000330082
64	24	23	23	18	0. - 34.2 𝕚	-0.0000926747 + 0.000678435 𝕚	0.000351782 - 0.000908228 𝕚
65	16	16	16	18	15.8	0.000043027	0.00000195965
66	21	20	17	18	19.5	0.0000538752	0.0000808129
67	24	23	24	18	0. - 26.5 𝕚	0.0000352941 + 0.000676074 𝕚	0.000210607 - 0.000910589 𝕚
68	18	18	20	25	12.6	0.000400011	-0.000081191
69	20	22	22	19	21.5	0.0000247897	0.0000587455
70	19	18	18	16	18.8	0.00000576614	0.0000637699
71	19	21	19	18	18.6	0.0000491422	0.0000120316
72	17	17	16	18	17.	0.0000142079	0.0000155203
73	16	17	19	18	17.6	0.0000323193	0.0000401094
74	25	29	25	16	0. - 8.19 𝕚	0.000716668 + 0.00055448 𝕚	-0.000358716 - 0.00103218 𝕚
75	12	12	12	20	5.89	0.000709585	-0.000165262
76	10	11	11	15	8.52	0.000296662	-0.000027863
77	12	10	13	12	12.9	0.00000517267	0.000101877
78	19	19	19	22	17.8	0.0000941045	-0.0000122628
79	25	24	27	17	0. - 11.5 𝕚	0.000506316 + 0.000612176 𝕚	-0.000207414 - 0.000974487 𝕚
80	13	15	11	17	17.8	0.0000811993	0.000324797
81	17	16	17	13	25.7	-0.000137481	0.000355286
82	11	11	12	12	11.5	0.000010625	0.0000113415
83	13	14	12	22	6.05	0.000782597	-0.00015064
84	19	18	18	13	0. - 17.9 𝕚	0.000106605 + 0.0006626 𝕚	0.000158557 - 0.000924063 𝕚
85	16	16	15	17	16.	0.0000150535	0.0000165353
86	11	11	14	18	9.31	0.000442446	0.000067285
87	13	13	14	31	3.47	0.00148674	-0.000162781
88	14	14	15	17	13.9	0.000088196	0.0000089012
89	19	17	15	19	21.4	-0.0000139778	0.000189207

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
90	10	10	10	15	7.03	0.000402423	-0.0000728213
91	19	19	20	10	0. - 5.21 𝕚	0.000763284 + 0.000564822 𝕚	-0.000431427 - 0.00102184 𝕚
92	22	22	20	14	0. - 10.6 𝕚	0.000452011 + 0.00062224 𝕚	-0.000162072 - 0.000964423 𝕚
93	17	18	17	11	0. - 9.71 𝕚	0.000384454 + 0.000633429 𝕚	-0.000104186 - 0.000953234 𝕚
94	15	16	13	13	14.1	0.0000653785	0.0000535391
95	14	15	15	16	15.	0.0000160059	0.0000176922
96	9	10	12	16	8.24	0.000436428	0.0000784071
97	15	15	12	15	18.4	-0.0000601773	0.000195576
98	15	18	14	22	12.7	0.00039732	0.0000873539
99	18	18	19	21	17.8	0.0000714974	0.0000056775
100	15	15	15	15	15.	indeterminate	indeterminate

C.1.4 parameters: tMRHCD, $\alpha = 5$, $\lambda = 0.008$, $\mu = 0.002$

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
1	6	5	4	5	10	3.47	0.00590331	0.0000207442
2	16	12	8	8	7	5.85	0.00746512	0.00148943
3	8	8	10	9	9	8.71	0.000395149	0.000285635
4	4	4	8	8	10	0. - 5.37 𝕚	0.00433651 + 0.00517485 𝕚	0.00179521 - 0.0117154 𝕚
5	14	14	6	7	8	0. - 10.1 𝕚	0.00518185 + 0.00315144 𝕚	0.00548832 - 0.0137388 𝕚
6	10	12	7	7	14	0. - 88.4 𝕚	-0.00220737 + 0.0031026 𝕚	0.0212244 - 0.0137877 𝕚
7	9	9	8	9	8	8.66	0.000114013	0.000187083
8	12	14	7	8	11	0. - 18.8 𝕚	0.000656618 + 0.00517739 𝕚	0.00702354 - 0.0117129 𝕚
9	7	8	5	5	13	4.03	0.00778589	0.000970943
10	4	3	10	11	7	0. - 3.09 𝕚	0.00782668 + 0.00415547 𝕚	-0.000956292 - 0.0127348 𝕚
11	3	2	7	8	10	0. - 2.23 𝕚	0.00856248 + 0.00421104 𝕚	-0.00209054 - 0.0126792 𝕚
12	7	7	7	7	8	7.11	0.000185597	0.0000518693
13	10	8	7	7	9	7.97	0.00103066	0.000727712
14	14	14	6	6	9	0. - 5.97 𝕚	0.00667499 + 0.00385817 𝕚	0.0013372 - 0.0130321 𝕚
15	4	9	8	8	6	0. - 8.65 𝕚	0.00106624 + 0.00636697 𝕚	0.00333689 - 0.0105233 𝕚
16	12	10	10	10	9	9.58	0.000827009	0.000153652
17	11	9	9	10	8	9.29	0.000656438	0.000526021
18	4	5	13	10	10	0. - 3.59 𝕚	0.00776326 + 0.00441816 𝕚	-0.00136991 - 0.0124721 𝕚
19	15	13	9	9	13	74.9	-0.00417799	0.015691
20	12	12	6	7	10	0. - 9.05 𝕚	0.00293808 + 0.00568021 𝕚	0.0025355 - 0.0112101 𝕚
21	14	12	12	12	9	17.1	-0.0005972	0.00338905
22	10	10	7	8	5	0. - 17.6 𝕚	-0.00121838 + 0.00601554 𝕚	0.00723659 - 0.0108747 𝕚
23	13	14	5	5	7	0. - 4.62 𝕚	0.00822613 + 0.0027737 𝕚	0.00128829 - 0.0141166 𝕚
24	8	12	4	4	12	0. - 3.1 𝕚	0.00858889 + 0.00393686 𝕚	-0.00162143 - 0.0129534 𝕚
25	10	11	6	6	7	8.03	0.00294232	0.0029828
26	10	13	8	8	7	6.41	0.00415275	0.000273054
27	10	8	12	13	9	11.2	0.00145533	0.00223118
28	19	20	14	14	9	0. - 7.39 𝕚	0.00655314 + 0.00519077 𝕚	-0.00120514 - 0.0116995 𝕚
29	6	6	6	7	14	2.83	0.00933546	-0.00156951
30	14	12	11	10	10	9.95	0.00171935	0.00025264
31	6	5	11	10	4	0. - 8.46 𝕚	0.00416815 + 0.00362831 𝕚	0.00589846 - 0.013262 𝕚
32	8	7	12	9	8	6.49	0.00318433	-0.0000903521
33	7	6	15	14	4	0. - 3.35 𝕚	0.00989827 + 0.00257247 𝕚	-0.000969577 - 0.0143178 𝕚
34	11	14	6	5	10	0. - 5.48 𝕚	0.0063402 + 0.00435172 𝕚	0.000767574 - 0.0125386 𝕚
35	10	11	12	12	8	16.7	-0.00102262	0.00384827
36	3	5	8	7	10	0. - 4.82 𝕚	0.00472661 + 0.00513291 𝕚	0.00135415 - 0.0117574 𝕚
37	9	9	8	9	9	8.9	0.0000362896	0.000160344

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
38	11	13	9	9	6	0. - 23.1 ⅈ	-0.000737801 + 0.00531379 ⅈ	0.00868445 - 0.0115765 ⅈ
39	15	15	11	12	7	0. - 6.91 ⅈ	0.00510209 + 0.00596148 ⅈ	-0.000828274 - 0.0109288 ⅈ
40	4	6	11	11	10	0. - 3.86 ⅈ	0.0066074 + 0.00544445 ⅈ	-0.0017499 - 0.0114458 ⅈ
41	8	6	10	10	10	0. - 44.4 ⅈ	-0.00579439 + 0.00656421 ⅈ	0.011601 - 0.0103261 ⅈ
42	6	6	10	11	6	7.38	0.00362111	0.0030304
43	6	5	7	6	6	6.04	0.000323544	0.00039571
44	5	4	8	8	10	0. - 8.06 ⅈ	0.00240336 + 0.00540087 ⅈ	0.00392522 - 0.0114894 ⅈ
45	8	9	6	7	8	8.16	0.000381603	0.00114249
46	13	12	6	8	13	0. - 5.83 ⅈ	0.00537599 + 0.00579352 ⅈ	-0.0008467 - 0.0110968 ⅈ
47	7	5	4	4	7	5.87	0.00146329	0.00235452
48	5	4	12	12	6	0. - 4.56 ⅈ	0.00747828 + 0.00313333 ⅈ	0.00169649 - 0.0137569 ⅈ
49	4	4	7	6	7	8.02	0.000171636	0.00403379
50	4	4	7	7	11	6.59	0.00541363	0.00539471
51	6	6	5	4	8	5.74	0.00168039	0.00156503
52	17	13	14	13	8	0. - 8.97 ⅈ	0.00412841 + 0.00608213 ⅈ	0.000142832 - 0.0108082 ⅈ
53	7	8	12	12	11	0. - 18.2 ⅈ	-0.000570289 + 0.00618764 ⅈ	0.00588837 - 0.0107026 ⅈ
54	15	15	6	6	10	0. - 4.48 ⅈ	0.0082113 + 0.00381081 ⅈ	-0.000838304 - 0.0130795 ⅈ
55	5	7	5	5	10	4.19	0.00486971	0.000310508
56	4	3	14	13	12	0. - 1.85 ⅈ	0.0117481 + 0.00364004 ⅈ	-0.00549986 - 0.0132502 ⅈ
57	8	8	15	16	5	0. - 4.54 ⅈ	0.00873017 + 0.00309645 ⅈ	-0.000193872 - 0.0137938 ⅈ
58	5	7	7	6	18	2.19	0.0139708	-0.000755285
59	14	14	11	10	12	13.3	0.000721592	0.00163152
60	12	9	10	10	12	10.7	0.000681959	0.000785852
61	14	12	5	5	9	0. - 4.3 ⅈ	0.00770843 + 0.00387911 ⅈ	-0.000240866 - 0.0130112 ⅈ
62	13	13	13	12	12	12.7	0.0000791209	0.00012614
63	9	11	9	10	5	0. - 6.95 ⅈ	0.00297773 + 0.00661939 ⅈ	0.000446384 - 0.0102709 ⅈ
64	13	10	9	8	13	12.3	0.00149947	0.00313754
65	8	8	7	6	7	7.42	0.000263413	0.000585495
66	8	11	8	7	11	9.11	0.0016182	0.00174718
67	4	6	6	7	6	7.27	-0.000220533	0.00220891
68	8	9	9	11	7	8.48	0.00125343	0.00085779
69	13	12	3	3	9	0. - 1.89 ⅈ	0.011157 + 0.00342243 ⅈ	-0.00434363 - 0.0134679 ⅈ
70	6	6	8	8	7	7.14	0.000513148	0.000728052
71	11	13	9	10	16	8.81	0.00398678	0.000846715
72	9	7	5	5	6	5.41	0.00262835	0.000828973
73	9	8	9	8	5	0. - 18.5 ⅈ	-0.00271447 + 0.00691976 ⅈ	0.00657246 - 0.00997052 ⅈ
74	10	9	10	9	6	0. - 61.1 ⅈ	-0.00762036 + 0.00677518 ⅈ	0.0132157 - 0.0101151 ⅈ
75	19	17	5	5	9	0. - 2.51 ⅈ	0.0125491 + 0.00159815 ⅈ	-0.00335072 - 0.0152921 ⅈ
76	12	11	6	6	12	0. - 5.66 ⅈ	0.00523706 + 0.00556652 ⅈ	-0.000224621 - 0.0113238 ⅈ
77	8	5	13	12	14	0. - 4.05 ⅈ	0.00747606 + 0.00533543 ⅈ	-0.00266133 - 0.0115548 ⅈ
78	11	9	8	8	7	7.55	0.00177809	0.000375723
79	10	9	9	9	10	9.36	0.000159092	0.000114881
80	9	10	8	8	6	10.2	0.000128556	0.00243579
81	11	9	12	11	11	11.9	-0.0000153039	0.00101775
82	11	11	8	8	7	9.11	0.0016182	0.00174718
83	4	6	13	13	7	0. - 4.11 ⅈ	0.00806994 + 0.00343349 ⅈ	0.000127924 - 0.0134568 ⅈ
84	8	7	6	6	4	8.95	-0.00015187	0.00379933
85	7	7	8	7	16	2.93	0.0100466	-0.00202224
86	9	10	8	8	5	0. - 15.5 ⅈ	-0.00135979 + 0.0066194 ⅈ	0.00572965 - 0.0102709 ⅈ
87	7	8	11	11	7	8.9	0.00200262	0.00212668
88	10	16	11	9	4	0. - 4.24 ⅈ	0.00806682 + 0.00408853 ⅈ	-0.00116765 - 0.0128018 ⅈ
89	5	5	7	8	7	7.21	0.000643061	0.00192316
90	5	6	5	5	3	6.53	-0.000407012	0.00290582
91	4	4	10	8	11	0. - 3.77 ⅈ	0.00674263 + 0.00470145 ⅈ	-0.000512116 - 0.0121888 ⅈ

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
92	14	12	3	3	7	0. - 2. ⅈ	0.0114817 + 0.0024296 ⅈ	-0.00314591 - 0.0144607 ⅈ
93	3	3	8	8	4	0. - 6.82 ⅈ	0.00368944 + 0.00366495 ⅈ	0.00660603 - 0.0132253 ⅈ
94	7	7	12	10	18	5.58	0.00937099	0.00226318
95	6	8	9	9	7	8.94	0.000269014	0.00173667
96	11	13	9	9	12	11.7	0.000912407	0.00173513
97	10	9	10	9	6	0. - 61.1 ⅈ	-0.00762036 + 0.00677518 ⅈ	0.0132157 - 0.0101151 ⅈ
98	10	10	12	11	6	0. - 8.68 ⅈ	0.0023096 + 0.00673164 ⅈ	0.00100574 - 0.0101586 ⅈ
99	7	7	7	7	11	5.2	0.00361313	-0.000752038
100	4	5	9	12	10	0. - 4.07 ⅈ	0.00670085 + 0.00476047 ⅈ	-0.000571235 - 0.0121298 ⅈ

C.1.5 parameters: tMRHCD, $\alpha = 10$, $\lambda = 0.002$, $\mu = 0.008$

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
1	6	4	4	5	7	5.16	0.00144398	0.00135689
2	5	6	5	4	3	5.42	0.000439723	0.00219516
3	5	5	4	3	3	4.26	0.000766361	0.00145364
4	5	4	8	7	6	7.35	0.000889949	0.00306987
5	5	6	3	2	4	8.	0.	0.0074532
6	4	5	7	7	4	5.87	0.00146329	0.00235452
7	2	3	3	2	6	2.3	0.00484539	0.0012739
8	6	7	8	7	5	7.21	0.000410763	0.00135944
9	4	7	5	6	5	5.41	0.00102422	0.00104906
10	6	8	3	5	3	7.86	0.00238081	0.00724297
11	5	6	5	6	4	5.45	0.00033873	0.000846826
12	3	4	5	6	5	5.42	0.000439723	0.00219516
13	4	4	6	5	6	5.2	0.000662704	0.00109232
14	8	9	7	7	8	7.72	0.000441025	0.000326872
15	5	4	7	7	8	11.7	-0.000891904	0.0059568
16	3	4	5	5	3	4.26	0.000766361	0.00145364
17	3	3	4	4	4	3.67	0.000252357	0.00047183
18	4	5	6	6	5	5.45	0.00033873	0.000846826
19	7	8	6	6	7	6.73	0.000498869	0.000381836
20	7	6	7	7	8	7.03	0.000281513	0.000334439
21	3	3	8	9	5	0. - 4.64 ⅈ	0.00517536 + 0.00406089 ⅈ	0.0031543 - 0.0128294 ⅈ
22	4	4	2	2	4	4.57	0.	0.00383521
23	4	5	4	3	6	4.48	0.00118219	0.00138318
24	5	4	6	6	9	4.82	0.00341767	0.00107304
25	6	6	6	6	4	6.76	-0.000337011	0.00168506
26	5	4	3	3	7	3.81	0.00332111	0.00176689
27	7	8	6	6	4	8.95	-0.00015187	0.00379933
28	8	7	3	3	4	14.6	0.00194468	0.0134893
29	6	4	5	5	6	5.45	0.00033873	0.000846826
30	8	6	5	6	4	5.74	0.00168039	0.00156503
31	1	0	8	8	4	0. - 0.924 ⅈ	0.0119179 + 0.00277118 ⅈ	-0.00436609 - 0.0141191 ⅈ
32	6	5	8	8	4	9.32	0.000509394	0.0048979
33	6	7	4	5	6	6.29	0.000436065	0.00167845
34	6	5	6	6	3	17.	-0.00397625	0.00877872
35	9	8	5	5	5	5.93	0.00286471	0.00205229
36	5	6	6	7	8	6.37	0.000900143	0.000842489
37	7	7	3	6	4	0. - 9.52 ⅈ	-0.000150531 + 0.00596071 ⅈ	0.00595063 - 0.0109296 ⅈ
38	4	7	8	8	4	0. - 9.05 ⅈ	0.000850989 + 0.00584424 ⅈ	0.00491632 - 0.011046 ⅈ
39	5	4	5	5	8	4.1	0.00307384	0.000107932

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
40	9	7	8	8	4	0. - 7.4 ⅈ	0.00161991 + 0.00669203 ⅈ	0.00191247 - 0.0101983 ⅈ
41	4	5	7	7	5	6.	0.00105737	0.00180483
42	6	6	5	5	5	5.37	0.000269918	0.000206573
43	7	7	3	2	6	0. - 3.01 ⅈ	0.00537607 + 0.00540824 ⅈ	-0.0000883889 - 0.011482 ⅈ
44	4	4	5	6	7	5.16	0.00144398	0.00135689
45	5	6	4	4	4	4.33	0.00105339	0.000397473
46	4	5	4	4	6	4.33	0.00105339	0.000397473
47	6	5	7	8	8	8.07	0.000251281	0.00208817
48	7	8	6	6	6	6.29	0.000771504	0.000245905
49	5	6	3	3	4	4.29	0.00164836	0.00186936
50	2	4	7	7	6	0. - 3.44 ⅈ	0.00454277 + 0.00584722 ⅈ	0.000098896 - 0.0110431 ⅈ
51	8	8	5	5	4	6.94	0.00191135	0.00348197
52	4	4	5	4	5	4.37	0.000326753	0.000257881
53	6	6	5	4	4	5.2	0.000662704	0.00109232
54	5	5	5	4	6	5.05	0.000380081	0.00048431
55	6	8	7	7	9	7.33	0.000801617	0.000702626
56	5	5	5	5	8	4.26	0.00267149	-0.000267149
57	6	3	6	4	8	13.7	-0.000348523	0.00964416
58	9	5	7	6	6	5.71	0.00216833	0.000618544
59	6	4	4	4	4	3.88	0.0014093	0.0000585444
60	4	5	6	5	6	5.45	0.00033873	0.000846826
61	7	9	7	7	2	0. - 2.77 ⅈ	0.0065205 + 0.00591912 ⅈ	-0.00249866 - 0.0109712 ⅈ
62	2	1	4	4	8	0. - 3.77 ⅈ	0.0064116 + 0.00206429 ⅈ	0.00632064 - 0.014826 ⅈ
63	7	6	6	6	4	7.27	-0.000220533	0.00220891
64	8	7	3	3	6	0. - 4.78 ⅈ	0.00372047 + 0.00527647 ⅈ	0.00240919 - 0.0116138 ⅈ
65	5	6	7	7	8	7.21	0.000410763	0.00135944
66	2	1	4	5	6	0. - 2.37 ⅈ	0.00529877 + 0.00501627 ⅈ	0.000822497 - 0.011874 ⅈ
67	5	5	8	8	6	6.68	0.00134658	0.00181224
68	3	2	7	7	5	0. - 3.58 ⅈ	0.00459182 + 0.00517257 ⅈ	0.00145163 - 0.0117177 ⅈ
69	6	6	6	6	4	6.76	-0.000337011	0.00168506
70	8	8	7	6	6	7.14	0.000513148	0.000728052
71	5	4	6	8	3	5.8	0.00264226	0.00381974
72	4	4	8	9	7	0. - 7.78 ⅈ	0.00211 + 0.00544952 ⅈ	0.00420946 - 0.0114408 ⅈ
73	6	5	9	7	5	5.41	0.00262835	0.000828973
74	1	2	1	1	5	1.64	0.00689381	0.00478606
75	9	7	6	5	7	6.63	0.00151279	0.00123438
76	3	6	5	5	4	5.42	0.000439723	0.00219516
77	8	7	6	5	9	8.06	0.000891382	0.00240161
78	6	5	7	7	3	0. - 9.41 ⅈ	-0.000480835 + 0.00644118 ⅈ	0.00509794 - 0.0104491 ⅈ
79	8	4	6	6	5	5.74	0.00168039	0.00156503
80	8	6	6	6	6	5.82	0.00102484	-0.00000000000000000711126
81	6	7	6	8	6	6.29	0.000771504	0.000245905
82	6	6	7	7	4	8.66	-0.000688422	0.00325765
83	5	5	3	4	5	5.11	0.0000334895	0.00165113
84	6	6	2	3	6	0. - 3.78 ⅈ	0.0035095 + 0.00585267 ⅈ	0.00140559 - 0.0110376 ⅈ
85	7	6	8	8	4	0. - 22.3 ⅈ	-0.00372801 + 0.00641379 ⅈ	0.00937659 - 0.0104765 ⅈ
86	7	6	5	5	3	9.14	-0.000649158	0.00540965
87	8	11	5	5	4	5.02	0.00617567	0.0032263
88	5	7	10	9	8	34.5	-0.00407668	0.0119182
89	8	6	8	7	10	7.54	0.00137191	0.0010142
90	5	6	7	7	11	4.69	0.00479404	0.000180955
91	5	7	8	10	7	8.14	0.00149661	0.00252604
92	9	11	6	5	4	14.7	0.00340572	0.0113986
93	6	6	10	9	5	7.42	0.00268886	0.00301095

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
94	6	5	2	1	11	0. - 1.84 ⅈ	0.0105717 + 0.00162412 ⅈ	-0.000168014 - 0.0152662 ⅈ
95	12	9	6	6	8	6.56	0.00410523	0.00169791
96	3	3	11	10	4	0. - 2.63 ⅈ	0.00933573 + 0.00247687 ⅈ	0.000120138 - 0.0144134 ⅈ
97	10	10	4	4	8	0. - 3.96 ⅈ	0.00609763 + 0.00505335 ⅈ	-0.000334567 - 0.0118369 ⅈ
98	7	6	10	10	6	8.17	0.00211979	0.0026209
99	6	8	11	12	8	14.	0.00117219	0.0059222
100	11	12	8	5	5	0. - 6.25 ⅈ	0.00514819 + 0.00446973 ⅈ	0.00223574 - 0.0124206 ⅈ

C.1.6 parameters: tMRHCD, $\alpha = 20$, $\lambda = 0.001$, $\mu = 0.009$

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
1	10	10	4	4	14	0. - 2.97 ⅈ	0.00922619 + 0.00361625 ⅈ	-0.0019384 - 0.013274 ⅈ
2	12	11	8	6	14	0. - 7.51 ⅈ	0.00440658 + 0.00544068 ⅈ	0.0011214 - 0.0114496 ⅈ
3	10	12	6	6	11	0. - 7.84 ⅈ	0.00348756 + 0.00559255 ⅈ	0.002007 - 0.0112977 ⅈ
4	9	10	8	7	7	7.97	0.00103066	0.000727712
5	8	10	6	6	10	11.1	0.000763369	0.0042791
6	8	10	8	9	11	8.94	0.00093908	0.000627532
7	10	9	5	5	5	7.06	0.00379548	0.00419385
8	7	8	6	4	9	0. - 16.9 ⅈ	-0.00179383 + 0.00601811 ⅈ	0.00801245 - 0.0108722 ⅈ
9	10	11	12	13	9	11.6	0.000722503	0.00128402
10	9	7	10	11	6	16.1	-0.000474666	0.00626312
11	9	8	11	11	8	9.41	0.00104606	0.0010601
12	13	10	7	8	10	8.93	0.00267702	0.00190326
13	9	8	9	9	12	7.65	0.00205232	-0.000156307
14	10	12	7	9	12	27.6	-0.00237798	0.00854155
15	12	11	14	14	10	13.3	0.000721592	0.00163152
16	14	15	9	10	13	0. - 44.6 ⅈ	-0.00317671 + 0.00577231 ⅈ	0.0107632 - 0.011118 ⅈ
17	10	11	10	12	8	11.8	0.000194033	0.00180367
18	9	8	7	7	4	0. - 9.82 ⅈ	0.000261232 + 0.00654959 ⅈ	0.00389645 - 0.0103407 ⅈ
19	11	8	12	11	12	24.4	-0.0028139	0.00595982
20	13	12	7	8	11	0. - 15.7 ⅈ	0.000562372 + 0.00588046 ⅈ	0.00520896 - 0.0110098 ⅈ
21	6	6	4	5	10	4.01	0.00522069	0.000536199
22	12	13	4	4	10	0. - 2.56 ⅈ	0.00962943 + 0.00413017 ⅈ	-0.00341287 - 0.0127601 ⅈ
23	4	7	11	9	11	0. - 4.42 ⅈ	0.00574062 + 0.00571785 ⅈ	-0.00116114 - 0.0111724 ⅈ
24	9	9	8	8	15	4.52	0.00669317	-0.00163626
25	12	12	10	10	11	11.1	0.000349575	0.00043557
26	10	8	9	11	12	10.7	0.000766361	0.00145364
27	7	6	9	9	12	8.21	0.00283834	0.00234143
28	9	10	16	15	11	13.4	0.00306957	0.00410415
29	9	8	13	13	12	0. - 30. ⅈ	-0.00239888 + 0.006165 ⅈ	0.00838527 - 0.0107253 ⅈ
30	8	8	11	12	13	19.7	-0.000460288	0.00638702
31	11	12	11	12	11	11.4	0.000131982	0.0000939907
32	6	6	9	9	12	9.13	0.00291237	0.0038133
33	11	9	10	10	8	10.1	0.000328174	0.00086608
34	11	13	9	10	16	8.81	0.00398678	0.000846715
35	12	14	13	14	7	0. - 6.36 ⅈ	0.00497129 + 0.00659573 ⅈ	-0.00186041 - 0.0102945 ⅈ
36	6	7	8	7	7	7.03	0.000281513	0.000334439
37	6	6	12	10	10	0. - 11. ⅈ	0.00196037 + 0.00545323 ⅈ	0.00440718 - 0.0114371 ⅈ
38	10	9	11	11	12	11.1	0.000305467	0.000772595
39	6	9	10	9	7	11.3	-0.0000862228	0.00340301
40	7	6	7	7	13	3.66	0.00704147	-0.0013695
41	11	10	10	11	8	11.8	-0.000194696	0.0015954

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
42	9	7	6	7	7	6.65	0.00111828	0.00025888
43	6	8	9	10	19	3.26	0.0117798	-0.000698277
44	12	14	13	13	9	0. - 75. 𝕚	-0.00713334 + 0.00682426 𝕚	0.0123995 - 0.010066 𝕚
45	9	14	10	8	8	5.9	0.00481688	-0.000639143
46	12	11	9	9	9	9.27	0.0012345	0.00042173
47	7	8	10	11	6	9.79	0.00155327	0.0031949
48	10	11	9	9	12	9.91	0.000862162	0.000550878
49	9	10	11	11	10	10.4	0.000195859	0.000400109
50	12	11	14	11	10	10.4	0.00140054	0.000212571
51	13	10	12	12	8	36.4	-0.00374575	0.0091072
52	9	9	6	5	13	11.7	0.00410589	0.00770098
53	11	13	12	11	15	10.9	0.00160839	0.000210501
54	11	11	7	7	8	8.9	0.00200262	0.00212668
55	7	9	12	11	11	0. - 60.6 𝕚	-0.00623715 + 0.00631209 𝕚	0.0131421 - 0.0105782 𝕚
56	13	13	9	10	11	12.4	0.000750294	0.00183331
57	13	11	10	10	9	9.43	0.00150713	0.000252607
58	10	10	11	11	5	0. - 5.32 𝕚	0.00460214 + 0.00664304 𝕚	-0.00151578 - 0.0102472 𝕚
59	9	8	9	8	12	7.19	0.00243965	-0.000213803
60	11	11	11	10	10	10.7	0.0000934211	0.000150685
61	10	10	14	14	14	26.6	-0.00172792	0.00649094
62	9	10	11	11	8	10.8	0.000264444	0.00130083
63	15	17	9	10	7	0. - 12.7 𝕚	0.00500739 + 0.00309688 𝕚	0.00594781 - 0.0137934 𝕚
64	7	8	8	8	13	4.9	0.00519863	-0.00110036
65	5	8	9	9	8	0. - 18.5 𝕚	-0.00271447 + 0.00691976 𝕚	0.00657246 - 0.00997052 𝕚
66	9	10	6	7	16	4.6	0.00859757	0.000677206
67	10	9	8	8	10	9.11	0.000416282	0.000545256
68	12	9	14	13	7	0. - 9.12 𝕚	0.00332004 + 0.00601259 𝕚	0.0013019 - 0.0108777 𝕚
69	11	13	8	7	8	7.74	0.00365699	0.00156584
70	6	6	8	9	16	3.41	0.0100246	-0.00039847
71	9	9	7	8	13	6.16	0.00413549	-0.000184958
72	8	10	7	6	9	8.88	0.000856894	0.00197323
73	11	11	10	9	11	10.9	0.0000724598	0.00060579
74	10	9	12	12	13	13.5	0.000142998	0.00212803
75	10	12	10	9	3	0. - 2.84 𝕚	0.00809866 + 0.00570859 𝕚	-0.00407019 - 0.0111817 𝕚
76	7	5	17	19	6	0. - 2.76 𝕚	0.0124938 + 0.000982915 𝕚	-0.00215988 - 0.0159074 𝕚
77	9	9	4	5	14	0. - 6.32 𝕚	0.00685529 + 0.00258761 𝕚	0.00405565 - 0.0143027 𝕚
78	11	12	12	12	12	11.9	0.0000276021	0.000118816
79	10	11	11	10	13	10.1	0.00101285	0.000111703
80	8	9	9	9	11	8.6	0.000905613	0.000178919
81	10	9	13	15	14	0. - 44.6 𝕚	-0.00317671 + 0.00577231 𝕚	0.0107632 - 0.011118 𝕚
82	4	8	9	9	10	0. - 4.95 𝕚	0.00426264 + 0.00652333 𝕚	-0.000890009 - 0.010367 𝕚
83	10	8	5	5	11	0. - 8.24 𝕚	0.00315695 + 0.00504366 𝕚	0.00374915 - 0.0118466 𝕚
84	13	10	10	9	9	8.09	0.00225848	-0.000235294
85	6	10	9	9	12	41.1	-0.00380608	0.0122777
86	8	9	11	11	12	12.7	0.000103831	0.00242818
87	12	12	9	10	8	11.5	0.000775349	0.00209196
88	10	12	10	11	8	11.8	0.000194033	0.00180367
89	14	11	9	11	12	11.6	0.00118208	0.00132454
90	10	7	11	10	11	31.	-0.0041951	0.00818573
91	10	9	10	10	11	10.	0.000202378	0.000228247
92	12	13	12	11	9	14.8	-0.000416917	0.00238869
93	9	10	14	15	7	0. - 16.5 𝕚	0.00234077 + 0.00439603 𝕚	0.00670639 - 0.0124943 𝕚
94	8	9	5	5	9	0. - 18. 𝕚	-0.00158546 + 0.00584041 𝕚	0.00826713 - 0.0110499 𝕚
95	9	8	14	13	8	10.	0.00357153	0.00316454

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$
96	11	10	11	11	7	0. - 27.4 i	-0.00362533 + 0.00707493 i	0.00722903 - 0.00981535 i
97	8	7	11	11	13	38.1	-0.00251345	0.0118766
98	7	7	6	6	7	6.66	0.000146187	0.000246649
99	10	9	10	10	5	0. - 6.55 i	0.00307505 + 0.006834 i	-0.000102588 - 0.0100563 i
100	13	12	10	9	8	11.2	0.00145533	0.00223118

C.2 Maximum likelihood estimation

C.2.1 parameters: tMRHF, $\alpha = 2$, $\lambda = 0.0009$, $\mu = 0.0005$

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	2	2	2	3	3	1.0E-15	0.0001774038	-0.7549599838
2	2	2	3	1	1	0.0007276592	1.0E-15	-1.9476741746
3	1	1	1	6	1	0.0012066049	1.0E-15	-1.8900829848
4	4	3	5	7	50	1.0E-15	0.0022274118	-3.1289965579
5	4	4	3	4	4	1.0E-15	0.0001179610	-1.7976165759
6	5	5	4	10	16	0.0002024385	0.0010125363	-2.9358748682
7	2	2	2	6	2	0.0006716931	1.0E-15	-1.7877007860
8	3	3	3	3	3	1.0E-15	1.0E-15	-5,503369E-12
9	2	2	2	3	3	1.0E-15	0.0001774038	-0.7549599838
10	1	2	1	1	1	0.0004690824	1.0E-15	-2.1503929938
11	5	2	2	2	50	0.0031849901	0.0060941658	-3.9278203062
12	2	2	1	3	12	1.0E-15	0.0016975881	-2.0041164732
13	6	6	7	2	2	0.0007536586	1.0E-15	-2.5808924032
14	4	4	4	4	4	1.0E-15	1.0E-15	-7,337857E-12
15	2	2	4	5	50	0.0001416897	0.0027068279	-3.0189007284
16	5	5	6	3	3	0.0003970357	1.0E-15	-2.1830030037
17	1	2	1	3	50	1.0E-15	0.0031772157	-2.5969234191
18	4	4	6	1	1	0.0012669266	1.0E-15	-2.9417103384
19	8	8	9	3	3	0.0006402053	0.0000005656	-2.8500648036
20	2	2	3	4	9	1.0E-15	0.0010454721	-2.2438932039
21	2	2	2	2	2	1.0E-15	1.0E-15	-3,668928E-12
22	1	1	1	2	2	1.0E-15	0.0002810651	-0.7276794184
23	1	2	2	5	49	1.0E-15	0.0027587135	-2.7938250556
24	3	3	2	5	10	1.0E-15	0.0009926825	-2.2525143224
25	1	1	1	5	1	0.0010862257	1.0E-15	-1.7212351231
26	4	3	3	3	2	0.0005595647	1.0E-15	-2.4168650737
27	3	3	5	4	50	1.0E-15	0.0025295506	-2.8659293032
28	6	6	5	4	8	1.0E-15	0.0005247411	-2.0888768832
29	2	2	2	6	2	0.0006716931	1.0E-15	-1.7877007860
30	3	3	2	2	4	1.0E-15	0.0006085728	-1.8620982391
31	11	10	12	4	9	0.0008253127	0.0009926825	-3.7925304886
32	2	1	2	12	15	0.0024132701	0.0032219133	-3.8303272017
33	7	8	6	1	1	0.0014439474	1.0E-15	-3.3940866400
34	1	1	1	3	1	0.0006716931	1.0E-15	-1.2433353952
35	2	2	2	4	2	0.0003926693	1.0E-15	-1.2272569378
36	2	2	2	2	2	1.0E-15	1.0E-15	-3,668928E-12
37	5	5	5	4	5	1.0E-15	0.0001000000	-0.7423348566
38	1	1	1	1	1	1.0E-15	1.0E-15	-1,834729E-12
39	2	1	3	1	50	0.0000206924	0.0034551839	-2.6802625863
40	5	3	4	8	50	0.0007333239	0.0029004470	-3.7085151112
41	4	3	4	2	8	1.0E-15	0.0010605303	-2.3774690837

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
42	5	5	7	1	1	0.0013479136	1.0E-15	-3.0364650267
43	1	1	1	7	1	0.0013605019	1.0E-15	-2.0326646224
44	2	1	2	2	2	1.0E-15	0.0002358791	-2.1441057012
45	1	1	2	2	50	1.0E-15	0.0033963508	-1.9603804472
46	4	4	4	2	4	1.0E-15	0.0002710496	-1.2397974492
47	4	5	6	3	50	1.0E-15	0.0024916529	-2.9474682951
48	3	3	3	1	3	1.0E-15	0.0003763051	-1.2491930173
49	5	6	4	1	1	0.0012913261	1.0E-15	-3.2213095365
50	2	2	2	2	2	1.0E-15	1.0E-15	-3,668928E-12
51	3	3	3	3	3	1.0E-15	1.0E-15	-5,503369E-12
52	4	4	4	2	4	1.0E-15	0.0002710496	-1.2397974492
53	3	3	2	1	10	1.0E-15	0.0017240851	-1.9595078862
54	1	1	1	2	2	1.0E-15	0.0002810651	-0.7276794184
55	5	3	3	9	50	0.0018935591	0.0039869743	-3.9365369057
56	2	2	5	7	50	0.0018704036	0.0041681232	-3.7195123164
57	2	2	2	3	3	1.0E-15	0.0001774038	-0.7549599838
58	2	3	1	1	50	0.0008612252	0.0043537499	-2.7323503861
59	2	2	1	1	19	1.0E-15	0.0026803316	-1.7401201765
60	3	2	2	1	1	0.0007352617	1.0E-15	-2.2849906562
61	1	1	1	3	1	0.0006716931	1.0E-15	-1.2433353952
62	4	4	3	1	12	1.0E-15	0.0016382851	-2.2369315831
63	1	1	1	1	1	1.0E-15	1.0E-15	-1,834729E-12
64	4	4	4	3	4	1.0E-15	0.0001258283	-0.7368496229
65	2	2	2	3	3	1.0E-15	0.0001774038	-0.7549599838
66	1	1	2	3	50	1.0E-15	0.0031468615	-2.1454255633
67	3	3	3	4	4	1.0E-15	0.0001290479	-0.7678848270
68	2	2	2	7	2	0.0007801272	1.0E-15	-1.9987967168
69	5	5	4	4	5	1.0E-15	0.0002008260	-1.9148994073
70	7	5	7	1	44	0.0015714263	0.0040329239	-3.7899934884
71	5	5	4	2	10	1.0E-15	0.0011086264	-2.2621198520
72	2	2	3	1	1	0.0007276592	1.0E-15	-1.9476741746
73	1	1	1	1	1	1.0E-15	1.0E-15	-1,834729E-12
74	3	3	3	2	3	1.0E-15	0.0001703122	-0.7275168265
75	1	1	1	1	1	1.0E-15	1.0E-15	-1,834729E-12
76	2	2	3	2	2	0.0002351470	1.0E-15	-1.8045952222
77	1	1	1	4	1	0.0008828739	1.0E-15	-1.5133744400
78	4	4	5	2	2	0.0005416387	1.0E-15	-2.1602977878
79	6	5	6	3	10	1.0E-15	0.0008487511	-2.6205183940
80	3	2	2	2	2	0.0002353635	1.0E-15	-2.1483056995
81	2	2	3	7	2	0.0009685857	0.0000337722	-2.8195115793
82	2	2	3	3	3	1.0E-15	0.0001644235	-2.0456787263
83	1	1	1	5	1	0.0010862257	1.0E-15	-1.7212351231
84	2	2	2	1	2	1.0E-15	0.0002695163	-0.7079396163
85	8	10	7	2	49	0.0015461986	0.0038038343	-4.0302577905
86	7	7	4	5	50	1.0E-15	0.0022151467	-3.2882322751
87	4	4	4	1	5	1.0E-15	0.0006571019	-1.7918257802
88	3	3	4	4	3	0.0002900013	1.0E-15	-2.0308452889
89	1	1	2	2	50	1.0E-15	0.0033963508	-1.9603804472
90	1	1	1	2	2	1.0E-15	0.0002810651	-0.7276794184
91	1	1	1	5	1	0.0010862257	1.0E-15	-1.7212351231
92	2	2	2	3	3	1.0E-15	0.0001774038	-0.7549599838
93	2	4	2	3	50	0.0013601805	0.0042364917	-3.3995226387
94	4	4	5	3	3	0.0002824400	1.0E-15	-1.9691404622
95	3	3	3	1	3	1.0E-15	0.0003763051	-1.2491930173

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
96	4	4	5	3	3	0.0002824400	1.0E-15	-1.9691404622
97	4	4	5	2	2	0.0005416387	1.0E-15	-2.1602977878
98	3	4	4	5	6	1.0E-15	0.0003767874	-2.3789170976
99	3	3	3	5	3	0.0002773487	1.0E-15	-1.2242233901
100	5	4	4	3	2	0.0006828028	1.0E-15	-2.5361950585

C.2.2 parameters: tMRHF, $\alpha = 5$, $\lambda = 0.0008$, $\mu = 0.0002$

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	7	7	8	13	5	0.0007201371	1.0E-15	-2.89312060081
2	12	13	13	12	13	1.0E-15	0.0000742551	-2.29093659024
3	10	10	10	5	12	1.0E-15	0.0004470490	-2.48833822238
4	6	6	6	10	6	0.0002746746	1.0E-15	-1.90429781147
5	12	12	12	10	12	1.0E-15	0.0000821716	-1.22861591620
6	8	10	11	7	100	1.0E-15	0.0024888347	-3.68676375645
7	11	9	11	5	100	1.0E-15	0.0025681275	-3.80559224931
8	7	7	9	17	4	0.0011660711	1.0E-15	-3.78208065481
9	13	13	12	16	18	1.0E-15	0.0002400061	-2.48753626135
10	8	8	7	10	12	1.0E-15	0.0003270484	-2.26379894014
11	8	7	7	4	26	1.0E-15	0.0015359320	-2.95347585676
12	14	14	15	8	7	0.0004659554	1.0E-15	-3.08713966094
13	8	9	9	18	12	0.0007616036	0.0006655512	-3.81231745692
14	6	6	7	3	3	0.0004858426	1.0E-15	-2.40638775552
15	13	13	12	11	14	1.0E-15	0.0001833375	-2.13667731370
16	9	10	9	7	7	0.0001872231	1.0E-15	-2.56593524565
17	7	7	7	4	7	1.0E-15	0.0002265207	-1.70928531870
18	6	5	5	7	32	1.0E-15	0.0016814164	-2.81140685033
19	7	7	7	11	7	0.0002397226	1.0E-15	-1.91763253495
20	6	6	7	11	5	0.0005753557	1.0E-15	-2.75012919640
21	10	10	9	11	11	1.0E-15	0.0000889627	-2.01606562963
22	7	7	7	3	8	1.0E-15	0.0004441894	-2.15641640616
23	9	9	10	4	3	0.0007826896	1.0E-15	-2.86895547266
24	16	16	15	7	5	0.0007917824	1.0E-15	-3.57512115949
25	9	10	8	8	57	1.0E-15	0.0019295514	-3.13114471346
26	11	11	12	11	11	0.0000429604	1.0E-15	-1.80079646321
27	12	12	13	15	16	0.0000080084	0.0001701754	-2.48465509070
28	7	7	7	12	7	0.0002917089	1.0E-15	-2.21139365470
29	8	8	7	11	14	1.0E-15	0.0004233255	-2.48824575301
30	7	7	7	12	7	0.0002917089	1.0E-15	-2.21139365470
31	2	3	3	9	22	0.0007796246	0.0021499433	-3.40133159961
32	9	8	9	4	18	1.0E-15	0.0010393336	-3.09378440915
33	12	12	14	13	27	1.0E-15	0.0007454814	-3.16620001374
34	8	7	7	10	26	1.0E-15	0.0011276832	-2.97140766085
35	9	8	9	8	9	1.0E-15	0.0001082062	-2.28340827239
36	10	10	10	6	11	1.0E-15	0.0003015925	-2.09994387238
37	14	15	15	7	5	0.0008008923	0.0000089583	-3.65601928063
38	14	13	11	12	100	1.0E-15	0.0021221619	-3.52790530521
39	7	6	8	13	68	1.0E-15	0.0019453877	-3.59597410093
40	6	6	6	8	6	0.0001466541	1.0E-15	-1.22330649094
41	8	8	8	6	8	1.0E-15	0.0001258917	-1.23066105329
42	9	9	9	7	9	1.0E-15	0.0001111127	-1.22993181222
43	8	7	8	8	8	1.0E-15	0.0000592593	-2.14567718088

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
44	10	10	10	9	10	1.0E-15	0.0000484894	-0.75301154713
45	6	8	7	6	100	1.0E-15	0.0027606276	-3.30381769094
46	11	11	10	13	14	1.0E-15	0.0001860662	-2.26041886951
47	7	8	8	7	8	1.0E-15	0.0001221664	-2.28031928053
48	9	9	9	10	9	0.0000512828	1.0E-15	-0.77447156273
49	8	8	8	9	8	0.0000575183	1.0E-15	-0.77583520131
50	5	5	5	6	6	1.0E-15	0.0000835779	-0.78037485850
51	5	6	6	9	14	1.0E-15	0.0006689993	-2.81711297938
52	9	9	8	9	9	1.0E-15	0.0000527372	-1.79890575738
53	9	10	9	9	9	0.0000525519	1.0E-15	-2.14667258637
54	7	7	7	7	7	1.0E-15	1.0E-15	-0.00000000001
55	18	17	19	8	4	0.0011949505	1.0E-15	-4.09542513573
56	11	12	11	12	11	0.0000842434	1.0E-15	-2.32643430226
57	9	9	10	8	8	0.0001137580	1.0E-15	-1.98682970944
58	3	3	3	8	3	0.0005831913	1.0E-15	-2.07209416147
59	13	15	15	9	70	1.0E-15	0.0017966611	-3.82215913594
60	11	11	12	8	7	0.0003384368	1.0E-15	-2.46427402898
61	4	4	4	5	5	1.0E-15	0.0001014414	-0.77542895462
62	7	7	6	14	10	0.0005914410	0.0005770365	-3.28060086325
63	6	6	10	15	100	0.0015784780	0.0037873188	-4.39289006354
64	2	2	2	8	2	0.0008869552	1.0E-15	-2.18162508202
65	12	11	13	14	24	1.0E-15	0.0006242089	-3.15716078740
66	6	6	6	10	6	0.0002746746	1.0E-15	-1.90429781147
67	8	8	10	10	32	1.0E-15	0.0012339375	-3.13101152949
68	12	11	11	9	9	0.0001482858	1.0E-15	-2.57604183946
69	5	6	5	5	29	1.0E-15	0.0017536520	-2.68363304617
70	11	11	13	9	7	0.0004386853	1.0E-15	-3.04812695478
71	10	11	9	9	49	1.0E-15	0.0016649092	-3.16391860867
72	4	4	5	14	2	0.0015224134	1.0E-15	-3.31936161200
73	8	8	7	11	14	1.0E-15	0.0004233255	-2.48824575301
74	11	10	9	14	62	1.0E-15	0.0016757933	-3.43939365110
75	18	16	15	7	52	0.0008557133	0.0023556727	-4.42508166171
76	6	6	7	8	6	0.0002194137	1.0E-15	-2.23000210564
77	5	5	5	6	6	1.0E-15	0.0000835779	-0.78037485850
78	13	13	14	12	12	0.0000768597	1.0E-15	-1.99144626377
79	14	14	14	4	3	0.0010215442	1.0E-15	-3.61397956921
80	12	12	12	4	3	0.0009005594	1.0E-15	-3.24666100285
81	8	8	9	11	7	0.0003464796	1.0E-15	-2.41306332231
82	13	11	12	7	100	1.0E-15	0.0023644493	-3.72254203014
83	16	15	15	12	11	0.0002309289	1.0E-15	-2.79219997080
84	9	9	8	5	13	1.0E-15	0.0006468277	-2.52683369304
85	3	4	4	5	6	1.0E-15	0.0003739636	-2.37886864268
86	9	8	9	13	19	1.0E-15	0.0005769146	-3.02346424620
87	8	7	7	7	20	1.0E-15	0.0010527208	-2.76721671676
88	6	6	6	11	6	0.0003330600	1.0E-15	-2.18893787276
89	11	13	12	6	100	0.0000853998	0.0024993090	-3.86448903698
90	15	15	13	11	24	1.0E-15	0.0006515395	-3.05234265462
91	12	11	11	16	32	1.0E-15	0.0008718363	-3.26147966260
92	11	12	11	12	11	0.0000842434	1.0E-15	-2.32643430226
93	6	6	5	9	13	1.0E-15	0.0005948571	-2.48850544341
94	7	7	7	7	7	1.0E-15	1.0E-15	-0.00000000001
95	7	7	8	10	6	0.0003950420	1.0E-15	-2.42133644255
96	7	9	8	7	100	1.0E-15	0.0026168738	-3.34636039352
97	9	9	9	9	9	1.0E-15	1.0E-15	-0.00000000001

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
98	10	13	10	11	100	1.0E-15	0.0028370044	-4.15263887900
99	4	4	4	7	4	0.0003043214	1.0E-15	-1.56993911347
100	12	12	13	8	7	0.0003829502	1.0E-15	-2.66857708097

C.2.3 parameters: tMRHF, $\alpha = 10$, $\lambda = 0.0002$, $\mu = 0.0008$

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	6	5	5	4	39	1.0E-15	0.0021380208	-2.6883441783
2	5	5	5	7	5	0.0001726203	1.0E-15	-1.2232788229
3	4	4	4	5	5	1.0E-15	0.0001000000	-0.7754736233
4	2	3	3	9	22	0.0007903328	0.0021627723	-3.4013578458
5	8	8	6	2	27	0.0002729528	0.0020370993	-3.4008712318
6	6	6	7	8	6	0.0002193658	1.0E-15	-2.2300021629
7	5	4	5	6	6	1.0E-15	0.0001665169	-2.3772613837
8	7	6	8	6	27	1.0E-15	0.0014187139	-2.9992181359
9	7	7	7	5	5	0.0001677750	1.0E-15	-1.3285934538
10	6	6	7	5	5	0.0001770326	1.0E-15	-1.9795129343
11	4	4	4	6	6	1.0E-15	0.0001763525	-1.2866134010
12	4	4	5	3	3	0.0002824400	1.0E-15	-1.9691404622
13	8	6	8	4	50	1.0E-15	0.0021780935	-3.5149395668
14	7	7	5	9	26	1.0E-15	0.0012436309	-2.9895043571
15	2	3	3	6	22	1.0E-15	0.0016720845	-2.8064003490
16	3	2	4	3	50	1.0E-15	0.0027951783	-2.7660692636
17	2	2	2	2	2	1.0E-15	1.0E-15	-3.668928E-12
18	4	5	5	10	26	0.0000908099	0.0013875368	-3.2340098653
19	1	2	3	7	50	0.0012576483	0.0037051393	-3.4943612447
20	8	8	7	6	9	1.0E-15	0.0003019938	-2.1137785291
21	6	4	4	7	50	0.0009960493	0.0031937625	-3.6575412490
22	4	5	5	4	5	1.0E-15	0.0002008060	-2.2633416027
23	9	9	9	5	10	1.0E-15	0.0003364817	-2.1127535980
24	6	6	7	4	4	0.0003100612	1.0E-15	-2.1987688400
25	6	6	5	7	7	1.0E-15	0.0001432400	-2.0254854080
26	5	5	4	6	6	1.0E-15	0.0001665423	-2.0301941797
27	5	5	6	8	11	1.0E-15	0.0005027771	-2.4750446826
28	5	5	5	8	5	0.0002512266	1.0E-15	-1.5796588177
29	3	4	5	7	50	1.0E-15	0.0022274118	-3.1289965579
30	4	5	5	7	10	1.0E-15	0.0005675565	-2.6025914369
31	8	8	5	5	50	1.0E-15	0.0021380208	-3.3342853290
32	9	9	10	7	7	0.0001863666	1.0E-15	-2.2252715302
33	6	6	6	3	6	1.0E-15	0.0002717588	-1.7259136734
34	5	7	7	3	50	1.0E-15	0.0023883315	-3.5030176209
35	7	7	7	3	8	1.0E-15	0.0004461222	-2.1564406756
36	6	8	8	5	50	1.0E-15	0.0020960988	-3.4173784965
37	2	2	2	4	2	0.0003926693	1.0E-15	-1.2272569378
38	7	7	8	5	5	0.0002532150	1.0E-15	-2.2101114665
39	5	5	6	4	4	0.0002188162	1.0E-15	-1.9752470243
40	8	7	6	3	50	0.0002023025	0.0025162882	-3.2826813427
41	2	2	2	6	2	0.0006716931	1.0E-15	-1.7877007860
42	6	6	6	3	6	1.0E-15	0.0002717588	-1.7259136734
43	3	3	3	8	3	0.0005785554	1.0E-15	-2.0721599793
44	2	2	1	5	27	1.0E-15	0.0021509279	-2.4787073143
45	3	3	3	9	3	0.0006716931	1.0E-15	-2.2828481542

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
46	2	2	2	8	2	0.0008828739	1.0E-15	-2.1816508399
47	2	2	3	3	3	1.0E-15	0.0001644235	-2.0456787263
48	5	5	5	5	5	1.0E-15	1.0E-15	-9.172297E-12
49	9	9	10	5	4	0.0005870354	1.0E-15	-2.6696053932
50	6	6	7	6	6	0.0000779666	1.0E-15	-1.8015183038
51	4	4	4	6	6	1.0E-15	0.0001763525	-1.2866134010
52	8	9	6	2	50	0.0010267363	0.0033801412	-3.7842925331
53	9	8	8	7	21	1.0E-15	0.0010233180	-2.8103841119
54	7	7	6	3	12	1.0E-15	0.0009131029	-2.5271562164
55	6	6	3	2	50	0.0001850806	0.0028681734	-3.4083442519
56	5	5	5	6	6	1.0E-15	0.0000831552	-0.7803803730
57	5	5	4	8	13	1.0E-15	0.0007422787	-2.4882098498
58	6	6	5	3	10	1.0E-15	0.0008438689	-2.2807275123
59	4	4	4	4	4	1.0E-15	1.0E-15	-7.337857E-12
60	7	7	5	5	23	1.0E-15	0.0014315764	-2.7711723446
61	5	4	4	6	50	1.0E-15	0.0023121560	-2.7623119162
62	2	2	2	4	2	0.0003926693	1.0E-15	-1.2272569378
63	7	6	6	7	23	1.0E-15	0.0012680780	-2.7688105860
64	6	6	7	6	6	0.0000779666	1.0E-15	-1.8015183038
65	5	5	5	7	5	0.0001726203	1.0E-15	-1.2232788229
66	5	5	5	4	5	1.0E-15	0.0001000000	-0.7423348566
67	7	7	6	5	9	1.0E-15	0.0004553504	-2.1053613479
68	5	3	6	5	50	1.0E-15	0.0023529692	-3.6335285056
69	8	8	9	6	6	0.0002163400	1.0E-15	-2.2186955957
70	6	5	6	6	6	1.0E-15	0.0000786956	-2.1455053899
71	3	3	2	9	11	0.0007966603	0.0014640939	-3.0899295168
72	7	7	6	7	7	1.0E-15	0.0000676550	-1.7986094372
73	4	5	5	8	14	1.0E-15	0.0008266984	-2.8162577491
74	6	6	6	4	6	1.0E-15	0.0001732823	-1.2332004225
75	7	7	7	4	7	1.0E-15	0.0002280625	-1.7093150470
76	6	6	6	6	6	1.0E-15	1.0E-15	-1.100674E-11
77	4	4	4	8	4	0.0003926693	1.0E-15	-1.8646883408
78	2	2	1	3	12	1.0E-15	0.0016975881	-2.0041164732
79	3	3	4	3	3	0.0001580951	1.0E-15	-1.8030679010
80	7	7	8	4	4	0.0003904802	1.0E-15	-2.4328751084
81	3	2	3	4	5	1.0E-15	0.0004722813	-2.3784323266
82	5	5	5	4	5	1.0E-15	0.0001000000	-0.7423348566
83	5	5	5	5	5	1.0E-15	1.0E-15	-9.172297E-12
84	3	5	6	4	50	1.0E-15	0.0024370731	-3.6101016052
85	3	4	4	4	4	1.0E-15	0.0001182673	-2.1451513425
86	3	3	1	5	50	1.0E-15	0.0026975362	-2.9570868832
87	3	3	3	5	3	0.0002773487	1.0E-15	-1.2242233901
88	5	4	4	7	44	1.0E-15	0.0020960988	-2.8918797148
89	5	5	4	4	5	1.0E-15	0.0002008260	-1.9148994073
90	7	8	7	5	25	1.0E-15	0.0014248619	-2.8409089021
91	2	3	3	7	39	0.0000141508	0.0021580109	-3.0316297460
92	6	5	7	8	40	1.0E-15	0.0017539905	-3.0711121071
93	5	4	3	5	50	0.0001075442	0.0025669684	-3.0312530036
94	4	3	4	3	5	1.0E-15	0.0004560745	-2.2488222691
95	3	3	3	8	3	0.0005785554	1.0E-15	-2.0721599793
96	8	8	7	6	9	1.0E-15	0.0003019938	-2.1137785291
97	5	5	5	6	6	1.0E-15	0.0000831552	-0.7803803730
98	0	0	0	1	50	1.0E-15	0.0045258864	-3.0099759373
99	5	5	6	1	1	0.0011292454	1.0E-15	-2.5274319038

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
100	6	6	6	5	6	1.0E-15	0.0000817781	-0.7459335591

C.2.4 parameters: tMRHF, $\alpha = 20$, $\lambda = 0.0004$, $\mu = 0.0006$

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	15	15	16	18	14	0.0001866532	1.0E-15	-2.3887963745
2	11	11	12	12	11	0.0000837266	1.0E-15	-1.9811014116
3	16	18	18	24	50	1.0E-15	0.0009020903	-4.0266456754
4	14	15	16	9	41	0.0000031302	0.0012274847	-3.6460328549
5	16	15	16	18	19	1.0E-15	0.0001348582	-2.5980006193
6	16	16	16	17	16	0.0000289281	1.0E-15	-0.7696684240
7	15	14	15	18	21	1.0E-15	0.0002603881	-2.8169543798
8	16	18	14	14	50	0.0004502970	0.0016793939	-4.1927313490
9	21	19	19	17	50	0.0000612441	0.0010816668	-3.7953524392
10	13	13	14	16	12	0.0002168023	1.0E-15	-2.3928487229
11	10	10	11	8	8	0.0001656220	1.0E-15	-2.2305679532
12	12	12	13	11	11	0.0000832176	1.0E-15	-1.9905864827
13	20	20	22	9	5	0.0010730415	1.0E-15	-4.1261090310
14	18	17	15	10	50	0.0001803589	0.0015135027	-3.9713736308
15	22	23	23	25	26	1.0E-15	0.0000965840	-2.5956842142
16	16	16	17	17	16	0.0000581060	1.0E-15	-1.9749656947
17	16	14	15	10	50	1.0E-15	0.0013934090	-3.7720956942
18	16	16	18	17	13	0.0002903344	1.0E-15	-2.9754989176
19	10	9	10	19	12	0.0007237145	0.0005570883	-3.8162673466
20	15	16	16	11	21	1.0E-15	0.0004551934	-3.0679362233
21	13	13	13	16	13	0.0001041363	1.0E-15	-1.6105801226
22	19	21	21	16	41	1.0E-15	0.0008189899	-3.7584371169
23	15	17	14	26	50	0.0009328766	0.0018376090	-4.5846886268
24	18	18	19	21	22	0.0000012003	0.0001172848	-2.4826273684
25	16	19	16	18	50	0.0005084900	0.0015791889	-4.3608794457
26	20	18	23	22	50	0.0004777530	0.0013382236	-4.5301582398
27	20	22	23	13	50	0.0004125966	0.0014661522	-4.3888789676
28	12	11	11	17	38	1.0E-15	0.0010000000	-3.4055533426
29	18	18	18	21	21	1.0E-15	0.0000720010	-1.7573862033
30	10	10	11	17	8	0.0005362419	1.0E-15	-3.0316214536
31	15	13	15	15	27	1.0E-15	0.0006291229	-3.5131626247
32	20	21	20	17	16	0.0001647869	1.0E-15	-2.7963231108
33	13	15	15	17	33	1.0E-15	0.0007610441	-3.6254860832
34	16	17	17	16	17	1.0E-15	0.0000563053	-2.2948860888
35	15	14	16	14	24	1.0E-15	0.0005090142	-3.1277179501
36	20	19	21	18	29	1.0E-15	0.0004261960	-3.2104296297
37	18	21	21	11	50	0.0008385327	0.0020127544	-4.6512497104
38	21	21	23	11	7	0.0008393335	1.0E-15	-4.0190802043
39	18	17	18	25	33	0.0000712712	0.0005148934	-3.5741637391
40	16	15	19	12	50	0.0004103838	0.0016474166	-4.1708288714
41	15	15	16	20	13	0.0003156188	1.0E-15	-2.7356050971
42	12	13	12	19	9	0.0005602378	1.0E-15	-3.3534754566
43	13	14	15	15	23	1.0E-15	0.0004687439	-3.1149471407
44	15	16	18	15	50	0.0000006236	0.0011746160	-3.6333432286
45	19	19	18	20	20	1.0E-15	0.0000475094	-2.0094993407
46	10	9	11	9	23	1.0E-15	0.0008889400	-3.0730981561
47	22	22	22	23	22	0.0000213173	1.0E-15	-0.7679476209

no.	M	R	H	F	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
48	6	6	6	15	5	0.0007084469	1.0E-15	-2.9078058030
49	15	16	18	23	50	0.0001536337	0.0010975227	-4.0200652977
50	16	14	15	13	50	1.0E-15	0.0012803731	-3.5649349114
51	13	14	14	22	21	0.0003603057	0.0005328486	-3.7128023996
52	18	18	16	16	24	1.0E-15	0.0003810879	-2.9033073586
53	21	22	23	14	24	0.0003353751	0.0006172857	-3.9251038828
54	11	9	10	9	50	1.0E-15	0.0016855764	-3.4437723995
55	18	17	19	15	30	1.0E-15	0.0005962600	-3.2826118760
56	19	19	19	20	19	0.0000246326	1.0E-15	-0.7686670096
57	23	25	23	15	9	0.0007705477	1.0E-15	-4.2196720182
58	16	18	17	13	50	1.0E-15	0.0012214435	-3.6974622787
59	11	10	12	8	30	1.0E-15	0.0011484927	-3.2222824876
60	16	16	17	14	14	0.0000978174	1.0E-15	-2.2482189644
61	17	16	15	10	7	0.0006412793	1.0E-15	-3.6067055034
62	19	22	18	13	50	0.0008865789	0.0020156449	-4.5962530214
63	12	13	13	17	22	1.0E-15	0.0004084790	-3.0219966259
64	24	23	23	18	16	0.0002599756	1.0E-15	-3.1919193354
65	16	16	16	18	18	1.0E-15	0.0000545795	-1.2988032396
66	21	20	17	18	50	0.0001259875	0.0011340844	-4.0412430846
67	24	23	24	18	30	1.0E-15	0.0003638456	-3.2554252055
68	18	18	20	25	13	0.0005305396	1.0E-15	-3.4943200202
69	20	22	22	19	33	1.0E-15	0.0005061510	-3.6093098485
70	19	18	18	16	27	1.0E-15	0.0004650068	-2.9884007116
71	19	21	19	18	50	0.0000424804	0.0010367919	-3.7695049621
72	17	17	16	18	18	1.0E-15	0.0000536973	-2.0102707523
73	16	17	19	18	50	1.0E-15	0.0010534014	-3.6350117738
74	25	29	25	16	50	0.0012452792	0.0021134393	-5.0458697098
75	12	12	12	20	10	0.0004437694	1.0E-15	-2.8767394161
76	10	11	11	15	20	1.0E-15	0.0004526449	-3.0227252933
77	12	10	13	12	50	1.0E-15	0.0014633277	-3.7184041920
78	19	19	19	22	19	0.0000720451	1.0E-15	-1.6181736656
79	25	24	27	17	28	0.0004348939	0.0007036610	-4.2168468552
80	13	15	11	17	50	0.0007301916	0.0019579492	-4.3404583052
81	17	16	17	13	20	1.0E-15	0.0003062119	-2.8703592029
82	11	11	12	12	11	0.0000843345	1.0E-15	-1.9810920706
83	13	14	12	22	7	0.0009211461	0.0000000780	-3.9963684479
84	19	18	18	13	32	1.0E-15	0.0007259771	-3.3294829078
85	16	16	15	17	17	1.0E-15	0.0000566681	-2.0107955813
86	11	11	14	18	50	0.0003122748	0.0015242752	-4.0054602143
87	13	13	14	31	7	0.0011202224	1.0E-15	-4.1594892593
88	14	14	15	17	13	0.0002015364	1.0E-15	-2.3907129483
89	19	17	15	19	50	0.0004638673	0.0015083169	-4.2775751586
90	10	10	10	15	9	0.0003152509	1.0E-15	-2.2311735338
91	19	19	20	10	7	0.0007186952	1.0E-15	-3.6210073867
92	22	22	20	14	50	1.0E-15	0.0010478763	-3.6826854231
93	17	18	17	11	35	1.0E-15	0.0009090880	-3.4834205085
94	15	16	13	13	50	0.0000411801	0.0013563595	-3.6089184235
95	14	15	15	16	16	1.0E-15	0.0000605672	-2.3572603637
96	9	10	12	16	50	0.0002113229	0.0015571171	-3.9097047911
97	15	15	12	15	37	1.0E-15	0.0009626989	-3.4749933703
98	15	18	14	22	50	0.0010717080	0.0020732112	-4.7137869199
99	18	18	19	21	22	0.0000012003	0.0001172848	-2.4826273684
100	15	15	15	15	15	1.0E-15	1.0E-15	-2.751689E-11

C.2.5 parameters: tMRHCD, $\alpha = 5$, $\lambda = 0.008$, $\mu = 0.002$

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	6	5	4	5	10	26	0.0006927156	0.0154341391	-4.2778485346
2	16	12	8	8	7	5	0.0074673324	1.0E-15	-4.9436918536
3	8	8	10	9	9	8	0.0012162670	1.0E-15	-2.6326698603
4	4	4	8	8	10	4	0.0061983537	1.0E-15	-3.5706539401
5	14	14	6	7	8	14	0.0042137533	0.0082646935	-4.8314983644
6	10	12	7	7	14	6	0.0063044255	1.0E-15	-4.3330410571
7	9	9	8	9	8	9	1.0E-15	0.0007651383	-2.4460950658
8	12	14	7	8	11	31	1.0E-15	0.0115640387	-4.5518746569
9	7	8	5	5	13	4	0.0079987383	1.0E-15	-4.0610233469
10	4	3	10	11	7	7	0.0071447382	0.0072440880	-4.6050753886
11	3	2	7	8	10	2	0.0135317367	1.0E-15	-4.7256283255
12	7	7	7	7	8	7	0.0004721311	1.0E-15	-0.9467070847
13	10	8	7	7	9	7	0.0022372812	1.0E-15	-3.3345931756
14	14	14	6	6	9	5	0.0070037866	1.0E-15	-4.2603456461
15	12	10	10	10	9	9	0.0014097219	1.0E-15	-3.2142927957
16	11	9	9	10	8	18	1.0E-15	0.0073085961	-3.9554802670
17	15	13	9	9	13	8	0.0044768552	1.0E-15	-4.0443941850
18	12	12	6	7	10	16	1.0E-15	0.0055251050	-4.1968511735
19	14	12	12	12	9	18	1.0E-15	0.0049625573	-3.8507184600
20	10	10	7	8	5	13	1.0E-15	0.0058218187	-3.9869943387
21	13	14	5	5	7	4	0.0080529126	1.0E-15	-4.3688126278
22	8	12	4	4	12	8	0.0104725844	0.0094662142	-5.1739388728
23	10	11	6	6	7	6	0.0029818233	1.0E-15	-3.3142425261
24	10	13	8	8	7	6	0.0045269352	1.0E-15	-4.2230966444
25	10	8	12	13	9	23	1.0E-15	0.0088513589	-4.2459728694
26	19	20	14	14	9	7	0.0073040659	1.0E-15	-4.9882094319
27	6	6	6	7	14	4	0.0085166910	1.0E-15	-4.4788654372
28	14	12	11	10	10	22	1.0E-15	0.0072893966	-4.1424244737
29	6	5	11	10	4	3	0.0085954686	1.0E-15	-4.3219746776
30	7	6	15	14	4	3	0.0106943739	1.0E-15	-4.9773366698
31	10	11	12	12	8	8	0.0025255605	1.0E-15	-3.3488422478
32	9	9	8	9	9	9	1.0E-15	0.0003757818	-2.1269455051
33	11	13	9	9	6	5	0.0062117017	1.0E-15	-4.1819781846
34	15	15	11	12	7	19	0.0020929045	0.0077125272	-4.8096858181
35	4	6	11	11	10	22	1.0E-15	0.0101495278	-4.5545717582
36	8	6	10	10	10	12	1.0E-15	0.0032645925	-3.6392117121
37	6	6	10	11	6	5	0.0043292879	1.0E-15	-3.6289689011
38	6	5	7	6	6	8	1.0E-15	0.0033921192	-3.2915390355
39	5	4	8	8	10	13	0.0003512093	0.0063132012	-3.8767974533
40	8	9	6	7	8	11	1.0E-15	0.0040035898	-3.4698814809
41	7	5	4	4	7	12	1.0E-15	0.0081047866	-3.5428402624
42	5	4	12	12	6	3	0.0096793665	1.0E-15	-4.3675057251
43	4	4	7	6	7	9	1.0E-15	0.0047978916	-3.4746834917
44	4	4	7	7	11	4	0.0061997824	1.0E-15	-3.5961901667
45	6	6	5	4	8	9	1.0E-15	0.0040947189	-3.4950306732
46	7	8	12	12	11	14	1.0E-15	0.0034853330	-3.5419879176
47	15	15	6	6	10	5	0.0077398257	1.0E-15	-4.5133287492
48	5	7	5	5	10	4	0.0064071522	1.0E-15	-3.9753070731
49	4	3	14	13	12	2	0.0169288659	1.0E-15	-5.2835485448
50	8	8	15	16	5	4	0.0088979119	1.0E-15	-4.9278080481
51	5	7	7	6	18	3	0.0139056998	0.0003954418	-5.4106774387

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
52	14	14	11	10	12	15	1.0E-15	0.0022473494	-3.4646568667
53	12	9	10	10	12	18	1.0E-15	0.0055552898	-4.0450595479
54	14	12	5	5	9	4	0.0088040796	1.0E-15	-4.5045236267
55	13	13	13	12	12	13	1.0E-15	0.0005265918	-2.4511887582
56	9	11	9	10	5	30	1.0E-15	0.0138855939	-4.4789160496
57	13	10	9	8	13	43	1.0E-15	0.0146608388	-4.6241819152
58	8	8	7	6	7	8	1.0E-15	0.0013259012	-2.6294576998
59	8	11	8	7	11	44	1.0E-15	0.0166851717	-4.4580402082
60	4	6	6	7	6	20	1.0E-15	0.0133373007	-3.8214624223
61	8	9	9	11	7	37	1.0E-15	0.0158078399	-4.3669739832
62	13	12	3	3	9	3	0.0114910730	0.0006929547	-4.7424572276
63	6	6	8	8	7	6	0.0015645043	1.0E-15	-1.8352075049
64	11	13	9	10	16	34	1.0E-15	0.0107834284	-4.6728052476
65	9	7	5	5	6	5	0.0030207569	1.0E-15	-3.3356251362
66	9	8	9	8	5	14	1.0E-15	0.0069631071	-3.8809577601
67	10	9	10	9	6	15	1.0E-15	0.0063345255	-3.8855802385
68	19	17	5	5	9	3	0.0135528258	1.0E-15	-5.2226150099
69	12	11	6	6	12	5	0.0072017458	1.0E-15	-4.1277210838
70	11	9	8	8	7	7	0.0021708411	1.0E-15	-3.3867825569
71	10	9	9	9	10	9	0.0007357141	1.0E-15	-2.0278429751
72	9	10	8	8	6	6	0.0028263808	1.0E-15	-3.0712351378
73	11	9	12	11	11	17	1.0E-15	0.0049855446	-3.9311322528
74	11	11	8	8	7	7	0.0021426917	1.0E-15	-2.8790390843
75	4	6	13	13	7	12	0.0059058583	0.0099534463	-4.9504319502
76	8	7	6	6	4	9	1.0E-15	0.0046115684	-3.1242818438
77	7	7	8	7	16	5	0.0076220763	1.0E-15	-4.6656458910
78	9	10	8	8	5	12	1.0E-15	0.0049613849	-3.4799043244
79	7	8	11	11	7	6	0.0037121784	1.0E-15	-3.3978173778
80	5	5	7	8	7	5	0.0029942581	1.0E-15	-3.0662400133
81	5	6	5	5	3	6	1.0E-15	0.0031949126	-2.7528252668
82	4	4	10	8	11	3	0.0108493660	1.0E-15	-4.8830939250
83	14	12	3	3	7	2	0.0144085897	1.0E-15	-4.7573989784
84	3	3	8	8	4	3	0.0051878074	1.0E-15	-3.0336996892
85	7	7	12	10	18	4	0.0119728627	1.0E-15	-5.3947255324
86	6	8	9	9	7	10	1.0E-15	0.0030745033	-3.3673084838
87	11	13	9	9	12	8	0.0035936244	1.0E-15	-3.6373800369
88	10	9	10	9	6	15	1.0E-15	0.0063345255	-3.8855802385
89	10	10	12	11	6	16	0.0003507067	0.0063164164	-4.1666342894
90	7	7	7	7	11	7	0.0017755333	1.0E-15	-2.5457140257
91	11	14	6	5	10	47	0.0030947987	0.0203014294	-5.0449448118
92	4	5	13	10	10	50	0.0089676400	0.0280211294	-5.3627097009
93	4	9	8	8	6	50	0.0006841236	0.0220417565	-4.8853528155
94	8	7	12	9	8	50	0.0013477223	0.0202261676	-4.8916472827
95	3	5	8	7	10	50	0.0002133381	0.0213690217	-4.6369387477
96	13	12	6	8	13	50	1.0E-15	0.0164041754	-5.0513515289
97	17	13	14	13	8	50	0.0023125726	0.0173784914	-5.2877475149
98	8	5	13	12	14	50	0.0030103222	0.0194806474	-5.3101620340
99	10	16	11	9	4	50	0.0174025898	0.0355354629	-5.9689907364
100	4	5	9	12	10	50	0.0088960119	0.0283118357	-5.2762178079

C.2.6 parameters: tMRHCD, $\alpha = 10$, $\lambda = 0.002$, $\mu = 0.008$

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	6	4	4	5	7	34	1.0E-15	0.0196559163	-3.9053085660
2	5	6	5	4	3	10	1.0E-15	0.0090039443	-3.4959727084
3	5	5	4	3	3	5	1.0E-15	0.0030194518	-2.8964920083
4	5	4	8	7	6	13	1.0E-15	0.0084576696	-3.6770819193
5	5	6	3	2	4	14	1.0E-15	0.0133623515	-3.6497281756
6	4	5	7	7	4	8	1.0E-15	0.0047771119	-2.9815004874
7	2	3	3	2	6	50	1.0E-15	0.0284097827	-3.7513915442
8	6	7	8	7	5	11	1.0E-15	0.0060503583	-3.5470362743
9	4	7	5	6	5	50	1.0E-15	0.0239982565	-4.0551193959
10	6	8	3	5	3	50	0.0028176123	0.0279078719	-4.6313269727
11	5	6	5	6	4	9	1.0E-15	0.0064575054	-3.3642935367
12	3	4	5	6	5	10	1.0E-15	0.0083933913	-3.4450188479
13	4	4	6	5	6	6	0.0003118592	0.0020931690	-3.0524285836
14	8	9	7	7	8	7	0.0013773091	1.0E-15	-2.2494850541
15	5	4	7	7	8	9	1.0E-15	0.0035725887	-3.1145103294
16	3	4	5	5	3	5	1.0E-15	0.0030848972	-2.4684297389
17	3	3	4	4	4	4	1.0E-15	0.0009092487	-1.0282040626
18	4	5	6	6	5	6	1.0E-15	0.0018293371	-2.2165936587
19	7	8	6	6	7	6	0.0015924244	1.0E-15	-2.2525595648
20	7	6	7	7	8	8	1.0E-15	0.0013109313	-2.2302012358
21	3	3	8	9	5	3	0.0066331166	1.0E-15	-3.6866774559
22	4	4	2	2	4	4	1.0E-15	0.0018608929	-1.7346635678
23	4	5	4	3	6	10	1.0E-15	0.0083199145	-3.5084958324
24	5	4	6	6	9	7	0.0026196391	0.0034779826	-3.4963547790
25	6	6	6	6	4	6	1.0E-15	0.0011896123	-1.6051976043
26	5	4	3	3	7	9	0.0003306432	0.0069929005	-3.2755778963
27	7	8	6	6	4	9	1.0E-15	0.0046115684	-3.1242818438
28	8	7	3	3	4	3	0.0052738065	1.0E-15	-3.2315393409
29	6	4	5	5	6	7	1.0E-15	0.0032582567	-3.1676337405
30	8	6	5	6	4	27	1.0E-15	0.0169700731	-3.9583908891
31	1	0	8	8	4	10	0.0135463158	0.0232468008	-4.6582312771
32	6	5	8	8	4	10	1.0E-15	0.0058026963	-3.2784525074
33	6	7	4	5	6	9	1.0E-15	0.0051221216	-3.4534085493
34	6	5	6	6	3	7	1.0E-15	0.0040271671	-2.9073113286
35	9	8	5	5	5	5	0.0024377489	1.0E-15	-2.9823423567
36	5	6	6	7	8	11	1.0E-15	0.0055430117	-3.5508183166
37	7	7	3	6	4	50	0.0009093774	0.0250153070	-4.9396154079
38	4	7	8	8	4	32	1.0E-15	0.0182815234	-4.2514929285
39	5	4	5	5	8	9	0.0002896261	0.0049696905	-3.2103867237
40	9	7	8	8	4	15	1.0E-15	0.0086637642	-3.9301738173
41	4	5	7	7	5	7	1.0E-15	0.0027643874	-2.7619546118
42	6	6	5	5	5	5	0.0006548976	1.0E-15	-1.0742526114
43	7	7	3	2	6	11	1.0E-15	0.0080286985	-4.0267793374
44	4	4	5	6	7	4	0.0036551653	1.0E-15	-3.1708454343
45	5	6	4	4	4	4	0.0015984131	1.0E-15	-2.0853340061
46	4	5	4	4	6	6	1.0E-15	0.0024597302	-2.5077957845
47	6	5	7	8	8	11	1.0E-15	0.0050421363	-3.6019829056
48	7	8	6	6	6	6	0.0010873244	1.0E-15	-2.0974883561
49	5	6	3	3	4	3	0.0038007113	1.0E-15	-2.5843846682
50	2	4	7	7	6	14	1.0E-15	0.0105545654	-3.9429076298
51	8	8	5	5	4	9	1.0E-15	0.0047271273	-3.0427248862
52	4	4	5	4	5	4	0.0016229455	1.0E-15	-2.4976396171
53	6	6	5	4	4	6	1.0E-15	0.0024669011	-2.8943192412
54	5	5	5	4	6	6	1.0E-15	0.0017874119	-2.6747083363

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
55	6	8	7	7	9	11	1.0E-15	0.0040197955	-3.3894931890
56	5	5	5	5	8	5	0.0018570127	1.0E-15	-2.0485308816
57	6	3	6	4	8	50	0.0045883636	0.0277796824	-4.8494313649
58	9	5	7	6	6	50	1.0E-15	0.0218490695	-4.4934439242
59	6	4	4	4	4	9	1.0E-15	0.0079353070	-3.1816327408
60	4	5	6	5	6	8	1.0E-15	0.0046346851	-3.3426924087
61	7	9	7	7	2	6	0.0064102241	0.0070061187	-4.4686098916
62	2	1	4	4	8	5	0.0078570644	0.0091217948	-4.0527610215
63	7	6	6	6	4	7	1.0E-15	0.0026857961	-2.7514910829
64	8	7	3	3	6	7	0.0025512886	0.0049853707	-3.5607780802
65	5	6	7	7	8	8	1.0E-15	0.0018105933	-2.6271215201
66	2	1	4	5	6	27	0.0020706416	0.0229630302	-4.0240075947
67	5	5	8	8	6	5	0.0023995040	1.0E-15	-2.2837073261
68	3	2	7	7	5	10	1.0E-15	0.0079425677	-3.5677740825
69	6	6	6	6	4	6	1.0E-15	0.0011896123	-1.6051976043
70	8	8	7	6	6	8	1.0E-15	0.0018063303	-2.8923250962
71	5	4	6	8	3	49	0.0009845295	0.0258178494	-4.3066534190
72	4	4	8	9	7	4	0.0053193983	1.0E-15	-3.6322243130
73	6	5	9	7	5	50	0.0000293141	0.0224511497	-4.2574647086
74	1	2	1	1	5	1	0.0108846561	1.0E-15	-2.9560446924
75	9	7	6	5	7	21	1.0E-15	0.0120462993	-3.9666308363
76	3	6	5	5	4	27	1.0E-15	0.0192323153	-3.7839566172
77	8	7	6	5	9	13	1.0E-15	0.0062163322	-3.7682599327
78	6	5	7	7	3	9	1.0E-15	0.0059467605	-3.2497915922
79	8	4	6	6	5	50	1.0E-15	0.0233246300	-4.2820950268
80	8	6	6	6	6	6	0.0011154028	1.0E-15	-2.9905594023
81	6	7	6	8	6	50	1.0E-15	0.0219082847	-4.1306763847
82	6	6	7	7	4	7	1.0E-15	0.0021782334	-2.2443525847
83	5	5	3	4	5	5	1.0E-15	0.0014058519	-2.5853177953
84	6	6	2	3	6	8	1.0E-15	0.0054032359	-3.7317552728
85	7	6	8	8	4	9	1.0E-15	0.0040860866	-3.2319112948
86	7	6	5	5	3	8	0.0003365663	0.0056281025	-3.1247339401
87	8	11	5	5	4	3	0.0081679961	1.0E-15	-4.2768212087
88	5	7	10	9	8	23	1.0E-15	0.0116926001	-4.1487743891
89	8	6	8	7	10	21	1.0E-15	0.0102502846	-4.0772709748
90	5	6	7	7	11	5	0.0049498556	0.0000684730	-3.6747168328
91	5	7	8	10	7	50	1.0E-15	0.0207008997	-4.4837000324
92	9	11	6	5	4	23	0.0032541048	0.0166108146	-4.6237924183
93	6	6	10	9	5	5	0.0033741237	1.0E-15	-3.5341986058
94	6	5	2	1	11	20	0.0158149524	0.0288883936	-5.0833704583
95	12	9	6	6	8	5	0.0057425415	1.0E-15	-4.2016578161
96	3	3	11	10	4	3	0.0072072516	1.0E-15	-4.1157190193
97	10	10	4	4	8	13	1.0E-15	0.0059321913	-3.9199151067
98	7	6	10	10	6	12	1.0E-15	0.0050757730	-3.4283780133
99	6	8	11	12	8	28	1.0E-15	0.0124881526	-4.3556065587
100	11	12	8	5	5	50	0.0092639820	0.0290756083	-5.4852938209

C.2.7 parameters: tMRHCD, $\alpha = 20$, $\lambda = 0.001$, $\mu = 0.009$

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	10	10	4	4	14	3	0.0119341018	1.0E-15	-4.6066934503
2	12	11	8	6	14	50	0.0005344933	0.0170302126	-5.0952314582

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
3	10	12	6	6	11	5	0.0068230274	1.0E-15	-4.1552525514
4	9	10	8	7	7	12	1.0E-15	0.0044252294	-3.6095913776
5	8	10	6	6	10	16	1.0E-15	0.0069988776	-3.8421106885
6	8	10	8	9	11	18	1.0E-15	0.0069908426	-4.0211368908
7	10	9	5	5	5	5	0.0029400847	1.0E-15	-3.4054688724
8	7	8	6	4	9	50	1.0E-15	0.0207262767	-4.5212867542
9	10	11	12	13	9	17	1.0E-15	0.0051400344	-3.8242300996
10	9	7	10	11	6	26	1.0E-15	0.0124022122	-4.2732430747
11	9	8	11	11	8	8	0.0015752786	1.0E-15	-2.9953643147
12	13	10	7	8	10	45	1.0E-15	0.0164538947	-4.5415347225
13	9	8	9	9	12	8	0.0022993741	1.0E-15	-3.0962442212
14	10	12	7	9	12	50	1.0E-15	0.0170439687	-4.7453399392
15	12	11	14	14	10	15	1.0E-15	0.0025790761	-3.2620706426
16	14	15	9	10	13	20	1.0E-15	0.0051771052	-4.2019755456
17	10	11	10	12	8	26	1.0E-15	0.0105545969	-4.4424584291
18	9	8	7	7	4	11	1.0E-15	0.0055978237	-3.4894937151
19	11	8	12	11	12	32	1.0E-15	0.0115071412	-4.4417827192
20	13	12	7	8	11	19	1.0E-15	0.0065220473	-4.2108733393
21	6	6	4	5	10	14	0.0014235202	0.0091737880	-4.1536025714
22	12	13	4	4	10	3	0.0113998996	1.0E-15	-4.5935774802
23	4	7	11	9	11	50	0.0045261132	0.0232940651	-5.2074947972
24	9	9	8	8	15	7	0.0043333404	1.0E-15	-3.6456203256
25	12	12	10	10	11	12	1.0E-15	0.0008908182	-1.7781011160
26	10	8	9	11	12	47	1.0E-15	0.0163292489	-4.6007993527
27	7	6	9	9	12	10	0.0022637343	0.0032513887	-3.8081582775
28	9	10	16	15	11	8	0.0045493635	0.0001118996	-4.3460069544
29	9	8	13	13	12	15	1.0E-15	0.0032209437	-3.5343449533
30	8	8	11	12	13	8	0.0033132151	1.0E-15	-3.9016958562
31	11	12	11	12	11	11	0.0006222483	0.0000331343	-3.2033256147
32	6	6	9	9	12	6	0.0041371533	1.0E-15	-3.4309528678
33	11	9	10	10	8	8	0.0019646966	0.0000591035	-3.4157237998
34	11	13	9	10	16	39	1.0E-15	0.0122075453	-4.6781429323
35	12	14	13	14	7	35	1.0E-15	0.0121897268	-4.8323638314
36	6	7	8	7	7	9	1.0E-15	0.0029510399	-3.3044876977
37	6	6	12	10	10	50	0.0000689270	0.0187020255	-4.6570370260
38	10	9	11	11	12	12	1.0E-15	0.0011714969	-2.6142742089
39	6	9	10	9	7	41	1.0E-15	0.0175737044	-4.3418763711
40	7	6	7	7	13	5	0.0060941658	1.0E-15	-3.8485580550
41	11	10	10	11	8	14	1.0E-15	0.0040740234	-3.6460649755
42	9	7	6	7	7	17	1.0E-15	0.0093556913	-3.8674198621
43	6	8	9	10	19	4	0.0122439784	0.0004134378	-5.4108764277
44	12	14	13	13	9	18	1.0E-15	0.0046707937	-3.9210917862
45	9	14	10	8	8	50	0.0052624054	0.0229719160	-5.3321655578
46	12	11	9	9	9	9	0.0010880988	1.0E-15	-2.5823393039
47	7	8	10	11	6	18	1.0E-15	0.0087497806	-4.0283489304
48	10	11	9	9	12	9	0.0017411616	1.0E-15	-2.8048143094
49	9	10	11	11	10	11	1.0E-15	0.0009692348	-2.2230325735
50	12	11	14	11	10	50	1.0E-15	0.0158793990	-4.9452534637
51	13	10	12	12	8	23	1.0E-15	0.0083844973	-4.3414186711
52	9	9	6	5	13	16	0.0029999505	0.0089975952	-4.6353022577
53	11	13	12	11	15	22	1.0E-15	0.0059166118	-4.2093072301
54	11	11	7	7	8	7	0.0022544836	1.0E-15	-2.8995923566
55	7	9	12	11	11	22	1.0E-15	0.0083844973	-4.2276589745
56	13	13	9	10	11	14	1.0E-15	0.0024370731	-3.4676666939

no.	M	R	H	C	D	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
57	13	11	10	10	9	9	0.0017194715	1.0E-15	-3.4070819132
58	10	10	11	11	5	13	1.0E-15	0.0042419533	-3.8311433481
59	9	8	9	8	12	16	1.0E-15	0.0055098811	-3.9042659238
60	11	11	11	10	10	11	1.0E-15	0.0006295908	-2.4491476645
61	10	10	14	14	14	15	1.0E-15	0.0017926925	-3.1679990810
62	9	10	11	11	8	8	0.0018622148	1.0E-15	-2.9193638296
63	15	17	9	10	7	5	0.0083805126	0.0000261100	-5.1041541207
64	7	8	8	8	13	6	0.0048828683	1.0E-15	-3.6504489579
65	5	8	9	9	8	17	1.0E-15	0.0083844973	-4.0827319321
66	9	10	6	7	16	4	0.0106991938	1.0E-15	-4.9768573142
67	10	9	8	8	10	8	0.0015860366	1.0E-15	-2.5079370967
68	12	9	14	13	7	46	1.0E-15	0.0160028993	-4.9057323817
69	11	13	8	7	8	28	1.0E-15	0.0118181641	-4.4055762216
70	6	6	8	9	16	4	0.0100000000	1.0E-15	-4.7010367161
71	9	9	7	8	13	9	0.0032243533	0.0022806075	-4.1593706541
72	8	10	7	6	9	20	1.0E-15	0.0097254234	-4.0472213490
73	11	11	10	9	11	11	1.0E-15	0.0006192561	-2.5482873621
74	10	9	12	12	13	13	1.0E-15	0.0013851773	-3.0849481022
75	10	12	10	9	3	10	0.0086022408	0.0111105138	-5.1384589521
76	7	5	17	19	6	2	0.0173781048	1.0E-15	-5.7846316624
77	9	9	4	5	14	11	0.0078803100	0.0097999999	-5.0278625004
78	11	12	12	12	12	12	1.0E-15	0.0002794041	-1.6741179052
79	10	11	11	10	13	15	1.0E-15	0.0031196597	-3.6482091567
80	8	9	9	9	11	8	0.0019659004	1.0E-15	-2.7897784000
81	10	9	13	15	14	45	1.0E-15	0.0138605721	-4.7611498077
82	4	8	9	9	10	50	1.0E-15	0.0193189958	-4.7975216828
83	10	8	5	5	11	4	0.0080906030	1.0E-15	-4.1754567443
84	13	10	10	9	9	33	1.0E-15	0.0128016973	-4.3953406431
85	6	10	9	9	12	50	1.0E-15	0.0179244775	-4.7296230382
86	8	9	11	11	12	14	1.0E-15	0.0030819221	-3.1520789409
87	12	12	9	10	8	14	1.0E-15	0.0037277575	-3.6699550844
88	10	12	10	11	8	20	1.0E-15	0.0076988548	-4.1006878626
89	14	11	9	11	12	50	1.0E-15	0.0159045383	-4.8282564304
90	10	7	11	10	11	32	1.0E-15	0.0125423119	-4.4234497277
91	10	9	10	10	11	11	1.0E-15	0.0009465588	-2.2348356969
92	12	13	12	11	9	16	1.0E-15	0.0041034417	-3.8060748984
93	9	10	14	15	7	6	0.0058065147	1.0E-15	-4.5664311601
94	8	9	5	5	9	11	1.0E-15	0.0040483950	-3.4191466104
95	9	8	14	13	8	6	0.0055721285	1.0E-15	-4.2351069151
96	11	10	11	11	7	12	1.0E-15	0.0025001670	-3.3031202761
97	8	7	11	11	13	16	1.0E-15	0.0045392237	-3.8367140759
98	7	7	6	6	7	7	1.0E-15	0.0004913621	-0.9852911638
99	10	9	10	10	5	13	1.0E-15	0.0049513660	-3.7330229735
100	13	12	10	9	8	18	1.0E-15	0.0062185366	-4.0593050027

C.2.8 parameters: tMAMMALS, $\alpha = 10$, $\lambda = 0.002$, $\mu = 0.008$

no.	H	C	Ma	M	R	D	Cat	Ho	S	Ca	P	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
1	4	7	7	6	6	6	8	6	4	4	6	11	1.0E-15	0.00704	-9.56431
2	8	6	6	3	2	6	11	6	8	7	2	39	0.00696	0.02802	-11.0647
3	6	6	6	6	8	6	10	5	6	5	3	5	0.00426	0.00210	-8.67286
4	8	9	9	4	4	5	4	6	4	3	7	3	0.00729	1.0E-15	-8.61382

no.	H	C	Ma	M	R	D	Cat	Ho	S	Ca	P	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
5	2	3	3	6	6	3	4	4	1	1	3	7	1.0E-15	0.00717	-7.76684
6	6	6	6	5	5	5	7	8	5	6	3	8	1.0E-15	0.00402	-7.28653
7	6	7	7	7	8	4	4	5	5	4	4	8	1.0E-15	0.00381	-7.53912
8	9	10	8	6	4	3	4	7	9	8	3	33	1.0E-15	0.01854	-10.0535
9	10	10	12	7	6	8	8	7	6	6	9	6	0.00317	1.0E-15	-7.35235
10	8	8	8	3	8	3	6	5	1	3	4	38	1.0E-15	0.02183	-10.0910
11	9	9	10	4	6	0	2	6	8	6	7	8	0.00727	0.01060	-10.9712
12	4	4	4	6	4	9	7	4	5	3	5	4	0.00431	0.00153	-8.26363
13	5	5	5	5	5	3	3	8	8	7	3	3	0.00609	1.0E-15	-7.89439
14	4	4	5	4	5	5	3	3	5	5	6	6	1.0E-15	0.00351	-6.15223
15	6	6	6	4	4	4	4	8	4	3	6	4	0.00336	0.00072	-7.08212
16	5	4	8	6	4	7	8	10	7	7	4	26	1.0E-15	0.01500	-9.77562
17	11	10	13	8	7	6	7	8	3	3	7	6	0.00585	0.00329	-9.98667
18	6	5	4	8	6	6	3	3	6	7	6	17	1.0E-15	0.01262	-8.92567
19	5	5	5	9	7	4	4	3	3	2	5	4	0.00401	0.00212	-7.90924
20	6	7	6	5	5	6	5	11	6	7	4	4	0.00522	1.0E-15	-8.53118
21	5	3	5	4	5	8	6	7	5	4	4	12	1.0E-15	0.00924	-8.92346
22	5	5	4	4	5	10	6	7	8	6	5	4	0.00472	1.0E-15	-8.46430
23	4	4	4	3	4	7	5	5	5	5	6	4	0.00202	0.00062	-5.82137
24	7	7	8	5	1	8	5	6	7	7	7	12	1.0E-15	0.00764	-9.30735
25	5	6	7	11	5	7	5	10	6	5	7	50	1.0E-15	0.02124	-9.98872
26	5	5	7	5	3	4	5	3	5	5	8	15	0.00207	0.01181	-9.51553
27	6	6	7	7	8	5	6	8	8	8	6	8	1.0E-15	0.00186	-6.04936
28	4	4	5	4	6	3	1	3	8	7	6	8	0.00217	0.00828	-8.73397
29	6	6	5	8	5	7	7	2	4	4	3	8	0.00186	0.00677	-8.63734
30	7	7	7	4	4	5	5	6	7	7	6	7	1.0E-15	0.00206	-5.03700
31	5	5	7	5	3	4	5	3	5	5	8	3	0.00586	1.0E-15	-7.85320
32	11	11	13	4	6	6	6	7	3	3	4	3	0.00791	1.0E-15	-8.87948
33	6	5	6	5	3	8	11	11	3	4	2	2	0.01234	1.0E-15	-10.3686
34	4	4	5	7	7	5	4	4	9	8	9	4	0.00462	1.0E-15	-7.76323
35	10	10	6	5	3	7	7	2	4	3	3	2	0.01076	1.0E-15	-9.89199
36	9	9	10	5	6	6	5	2	12	11	5	5	0.00691	0.00394	-10.4806
37	5	7	5	6	7	6	5	8	8	7	5	14	1.0E-15	0.00892	-9.00811
38	6	6	6	5	7	5	3	6	8	8	4	9	1.0E-15	0.00506	-7.60236
39	6	6	6	4	7	9	4	4	5	4	4	4	0.00426	0.00056	-8.50878
40	7	7	9	5	8	7	8	9	9	8	6	12	1.0E-15	0.00514	-8.17013
41	8	9	8	3	6	8	4	8	8	4	7	50	1.0E-15	0.02232	-10.2187
42	3	3	3	3	3	6	6	8	7	7	7	3	0.00481	1.0E-15	-6.69219
43	8	8	9	5	6	5	5	7	7	6	8	9	0.00018	0.00352	-7.15660
44	3	3	2	4	5	10	8	7	8	9	9	7	0.00410	0.00522	-9.27383
45	8	7	8	8	7	6	6	5	8	5	10	15	1.0E-15	0.00833	-9.09155
46	5	5	4	8	7	6	2	5	9	8	5	14	1.0E-15	0.00972	-9.02663
47	7	8	8	5	8	6	7	1	3	3	1	1	0.01661	1.0E-15	-10.2908
48	6	6	4	5	4	5	5	4	10	6	4	50	1.0E-15	0.02449	-9.14159
49	8	8	7	6	5	3	2	2	1	2	5	7	0.00329	0.00845	-8.91546
50	5	5	5	3	6	3	7	7	7	5	5	13	1.0E-15	0.00986	-8.57195
51	3	3	5	8	6	5	4	4	11	8	7	3	0.00786	1.0E-15	-9.37939
52	5	5	5	4	7	8	9	7	6	9	6	14	1.0E-15	0.00845	-8.71819
53	6	7	6	5	7	11	7	8	8	8	6	15	1.0E-15	0.00806	-8.85940
54	5	5	7	4	4	4	4	9	4	4	6	4	0.00353	1.0E-15	-6.86512
55	2	2	2	5	6	5	7	10	5	4	7	2	0.00993	1.0E-15	-8.68218
56	6	6	4	4	5	7	6	3	4	3	4	9	1.0E-15	0.00739	-7.52324
57	5	5	7	7	7	3	4	5	7	6	11	3	0.00812	1.0E-15	-8.90195
58	2	2	3	6	6	11	13	7	5	8	7	5	0.00821	0.00504	-10.3358

no.	H	C	Ma	M	R	D	Cat	Ho	S	Ca	P	$\widehat{\alpha}$	$\widehat{\lambda}$	$\widehat{\mu}$	loglik
59	4	3	3	7	7	7	6	6	3	6	5	13	1.0E-15	0.00966	-8.91551
60	3	3	3	7	6	7	8	7	5	5	6	8	1.0E-15	0.00336	-6.71433
61	4	3	3	7	5	5	7	6	8	7	9	13	1.0E-15	0.00845	-8.88523
62	3	2	5	5	9	2	4	4	8	6	5	50	0.00075	0.02575	-9.88002
63	7	7	5	3	5	4	5	6	4	5	3	9	1.0E-15	0.00706	-7.52748
64	7	7	10	6	5	5	3	5	8	9	6	5	0.00472	0.00222	-9.05410
65	3	3	5	8	6	8	7	7	10	8	11	17	1.0E-15	0.00922	-9.41170
66	4	4	7	9	8	8	10	8	6	4	6	17	1.0E-15	0.00941	-9.35917
67	6	6	6	8	6	5	4	6	7	7	4	8	1.0E-15	0.00344	-6.68337
68	3	3	2	6	5	6	5	10	4	3	5	5	0.00461	0.00435	-8.65295
69	3	3	5	6	7	6	4	9	8	10	8	18	3.44E-6	0.01096	-9.31254
70	9	8	8	6	4	3	2	4	2	3	2	2	0.00845	1.0E-15	-8.33246
71	6	7	6	6	6	10	8	8	8	8	7	6	0.00188	1.0E-15	-6.55443
72	3	3	1	4	5	4	4	6	3	3	9	9	0.00352	0.01124	-9.05043
73	5	5	5	2	3	4	5	2	5	6	4	6	1.0E-15	0.00481	-6.75844
74	6	5	6	2	4	5	8	6	8	8	3	17	1.0E-15	0.01268	-9.39038
75	4	4	6	4	7	7	5	8	4	5	5	12	1.0E-15	0.00838	-8.30332
76	6	4	6	5	6	5	7	6	3	3	6	8	1.0E-15	0.00472	-8.55618
77	7	7	5	5	7	6	5	3	6	5	6	9	1.0E-15	0.00554	-7.46442
78	5	5	6	9	7	3	5	4	5	5	5	5	0.00257	0.00170	-7.32579
79	7	8	12	5	2	2	2	1	3	6	5	1	0.01780	1.0E-15	-10.8353
80	4	4	8	7	4	5	5	5	3	5	3	50	1.0E-15	0.02503	-9.10666
81	6	6	6	4	5	3	7	3	6	6	5	7	0.00049	0.00437	-7.47617
82	8	6	7	6	5	5	8	4	6	6	7	14	1.0E-15	0.00933	-8.88609
83	7	7	5	5	3	8	4	6	6	4	6	15	1.0E-15	0.01122	-8.66420
84	4	4	4	6	5	7	7	5	8	10	7	4	0.00424	1.0E-15	-7.31805
85	5	6	6	5	5	4	3	6	4	4	6	6	1.0E-15	0.00236	-6.48154
86	2	3	2	3	3	7	9	4	7	8	8	2	0.00901	1.0E-15	-9.29516
87	8	8	8	5	6	5	4	5	7	7	3	8	1.0E-15	0.00372	-7.27017
88	6	6	6	5	4	7	10	7	6	6	6	6	0.00160	0.00098	-6.69709
89	6	6	5	3	3	9	6	7	4	4	7	3	0.00589	1.0E-15	-7.80766
90	4	3	6	10	4	7	6	7	9	8	3	50	1.0E-15	0.02266	-10.4697
91	3	3	4	4	6	6	5	6	5	6	3	7	1.0E-15	0.00445	-7.10032
92	7	7	5	8	5	5	6	7	9	7	9	13	1.0E-15	0.00711	-8.48825
93	12	9	11	5	5	1	3	8	6	6	9	21	0.00885	0.02157	-11.6224
94	4	4	4	8	4	5	9	5	7	7	8	15	1.0E-15	0.00994	-8.91390
95	6	5	5	8	10	7	13	7	5	4	4	4	0.00655	1.0E-15	-10.0918
96	8	8	9	6	7	6	9	4	3	4	6	10	0.00142	0.00641	-8.80282
97	7	7	9	8	4	3	4	5	3	4	4	3	0.00655	1.0E-15	-8.42196
98	7	7	7	3	5	7	6	9	5	4	9	12	1.0E-15	0.00692	-8.68577
99	8	9	7	6	6	4	7	5	3	4	8	19	1.0E-15	0.01265	-9.08371
100	8	8	6	8	6	5	3	7	8	9	5	13	1.0E-15	0.00767	-8.67911

APPENDIX D

Abbreviations

α	...	ancestral gene copy number
BD	...	Birth and Death
B.C.	...	Before Christ
C / chimp	...	Chimpanzee
D	...	Dog
DNA	...	DeoxyriboNucleic Acid
F	...	fruit Fly
gcn	...	Gene Copy Number
H	...	Human
HC	...	ancestor of Human and Chimpanzee
iid	...	Independent Identically Distributed
λ	...	duplication (or birth) rate
μ	...	deletion (or death) rate
M	...	Mouse
Mb	...	Million Bases
Mbp	...	Million Base Pairs
ML	...	Maximum Likelihood
MLE	...	Maximum Likelihood Estimator
MOM	...	Method Of Moments
MR	...	ancestor of Mouse and Rat
MRCA	...	Most Recent Common Ancestor
MRH	...	ancestor of Mouse, Rat, and Human
MRHC	...	ancestor of Mouse, Rat, Human, and Chimpanzee
MRHCD	...	ancestor of Mouse, Rat, Human, Chimpanzee, and Dog
MRHF	...	ancestor of Mouse, Rat, Human, and fruit Fly
mya	...	Million Years Ago
myr	...	Million Years
OF	...	Olfactory Receptor

R	. . .	rat
tMRHCD	. . .	Tree including Mouse, Rat, Human, Chimp, and Dog
tMRHF	. . .	Tree including Mouse, Rat, Human, and fruit Fly
tMRHF2R	. . .	Tree including Mouse, Rat, Human, and fruit Fly and 2 Rounds of WGD
WGD	. . .	Whole Genome Duplication

APPENDIX

E

Curriculum vitae

Andrea Führer
Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories (MFPL)
Dr. Bohr Gasse 9, A-1030 Vienna, Austria
E-Mail: andrea.fuehrer@univie.ac.at

Date of birth: December, 27th 1979
Place of birth: Wolgast, Germany
Citizenship: German

Education

2006-2007 PhD student at the Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, Vienna

2003-2006 PhD student at the the Bioinformatics Department, Heinrich-Heine University, Düsseldorf

1998-2003 Diploma in Biomathematics at the Ernst-Moritz-Arndt University, Greifswald

1992-1998 Abitur at the Runge Gymnasium, Wolgast

Related conference presentations

Maximum-likelihood estimation of gene duplication- and deletion-rates. Andrea Fuehrer, Ingo Paulsen & Arndt von Haeseler. Annual Meeting of the Society for Molecular Biology and Evolution (SMBE), June 2005, Auckland, NZ

A maximum-likelihood framework to infer gene duplication and deletion rates for gene families. Andrea Fuehrer & Arndt von Haeseler. Mathematical and Statistical Aspects of Molecular Biology (MASAMB), March 2007, Manchester, UK

Using maximum likelihood to infer gene family specific gene duplication and deletion rates and the ancestral gene copy number. Andrea Fuehrer & Arndt von Haeseler. Annual Meeting of the Society for Molecular Biology and Evolution (SMBE), June 2007, Halifax, Canada

Vienna, September 2007