Development and Applications of Neutral Models for Evolution of Gene Expression

Inaugural – Dissertation

zur
Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Michael Roßkopf

aus Bochum

Mai 2007

Aus dem Institut für Bioinformatik
der Heinrich-Heine-Universität Düsseldorf

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

Referent: Prof. Dr. Arndt von Haeseler
Korreferent: Prof. Dr. Michael Leuschel

Tag der mündlichen Prüfung: 22.06.2007

Acknowledgments

First and foremost, I wish to thank my supervisor Arndt von Haeseler for his excellent advise, collaborations, and his friendly behaviour. I want to thank Gunter Weiss for the idea for this thesis and the mentoring in the first year of my PhD studies. Also I want to thank Michael Leuschel for accepting the task to read this thesis as a second reviewer.

I thank Ralf Kronenwett from the University Hospital Düsseldorf for the close collaboration and the medical data sets. I want to thank Philipp Khaitovich, Michael Lachmann, Wolfgang Enard, Ines Hellmann, and Svante Pääbo for the primate data sets, fruitful discussions, and new impulses during my visits at the Max-Planck Institute for Evolutionary Anthropology in Leipzig. Furthermore, I thank Chris Voolstra from the University of Cologne for discussions and his mice data sets.

Special thanks to Heiko Schmidt for help on several stuff and to Lutz Voigt for keeping the computers running. Finally, I would like to thank Thomas Laubach, Simone Linz, Jochen Kohl, Stefan Zanger, Gabriel Gelius-Dietrich, Steffen Kläre, Claudia Kiometzis, Anja Walge, Le Sy Vinh, Bui Quang Minh, Ricardo de Matos Simoes, Nicole Scherer, Thomas Schlegel, Tanja Gesell, Andrea Führer, Sascha Strauss, Jutta Buschbom, Ingo Ebersberger, and all other colleagues and former members of the Bioinformatics Institute in Düsseldorf and the Center for Integrative Bioinformatics Vienna (CIBIV) in Vienna. Ultimately, I am grateful to my family, my friends, and Christin.

Financial support from the rectorate of Düsseldorf University, from the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF), and from the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

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

1. M. Rosskopf, A. von Haeseler (2006) Testing the neutral evolution hypothesis for gene expression data, *Proc. Mathematical and Statistical Aspects of Molecular Biology (MASAMB 2006).*

2. M. Rosskopf, A. von Haeseler (2007) A gene expression evolution model with mutational and non-mutational effects, *submitted to Genetics.*

3. M. Rosskopf, G. Weiss, A. von Haeseler (2007) A neutral model for evolution of gene expression with gamma-distributed mutation effects, *in preparation.*

4. M. Rosskopf, A. von Haeseler (2007) A Tajima-type test to detect selection in gene expression data, *in preparation.*

The EMOGEE software package presented in this thesis is freely available from `http://www.cibiv.at/software/emogee`.

Other publications:

1. U.-P. Rohr, A. Rohrbeck, H. Geddert, S. Kliszewski, M. Rosskopf, A. von Haeseler, A. Schwalen, U. Steidl, R. Fenk, R. Haas, R. Kronenwett(2005) Primary human lung cancer cells of different histological subtypes can be distinguished by specific gene expression profiles, *Onkologie 2005*, 28(suppl 3):127.

2. I. Bruns, U. Steidl, J.-C. Fischer, S. Raschke, G. Kobbe G, R. Fenk, M. Rosskopf, S. Pechtel, U.-P. Rohr, A. von Haeseler, P. Wernet, D. Tenen, R. Haas, R. Kronenwett (2006) Pegylated G-CSF mobilizes CD34+cells with different stem and progenitor cell subsets and distinct functional properties in comparison with unconjugated G-CSF (2006) *Blood*, 108, 965A-966A 3382 Part 1.

3. E. Diaz-Blanco, I. Bruns, F. Neumann, J.-C. Fischer, T. Graef, M. Rosskopf, B. Brors, S. Pechtel, S. Bork, A. Koch, A. Baer, U.-P. Rohr, G. Kobbe, A. von Haeseler, N. Gattermann, R. Haas, R. Kronenwett (2007) Molecular signature of CD34+ hematopoietic stem and progenitor cells of patients with CML in chronic phase, *Leukemia*, 21, 494-504.

Abstract

Recent studies describe that the level of gene expression between species is positively correlated with the time that has passed since the species split from a common ancestor (Ranz and Machado, 2006). Moreover, Khaitovich *et al.* (2004) found a linear relationship between divergence time and expression differences. This linearity can be explained by the neutral theory (Kimura, 1983). Consequently, a neutral model for gene expression evolution was suggested (Khaitovich *et al.*, 2005b). The model describes mutations in the regulatory region of a gene by a compound Poisson process. The strength of changes in the expression level is described by a continuous distribution, the so-called mutation effect distribution. That is, whenever a mutation occurs, the gene expression level changes according to the mutation effect distribution.

In this thesis the model by Khaitovich *et al.* (2005b) is extended in two ways. In a first extension a gamma distribution is used to describe mutation effects which is more flexible than the distributions used in the original model. In a second extension, non-mutational effects are taken into account. These effects (e.g., metabolism and environmental effects) overlay mutational changes of gene expression. To describe them a new parameter is introduced which provides a better fit to evolutionary data. This makes it possible to estimate influences of mutational and non-mutational changes of the gene expression level. According to this, a Bayesian method to detect genes with mutations in their regulatory regions is suggested. Furthermore, a non-neutrality test is presented which can be applied to gene expression data sampled from individuals of a population. Based on this test one can detect those genes that show a significant deviation from expression levels under neutrality. The test is an adaptation of the widely used Tajima's D test (Tajima, 1989). Finally, a medical application is applied in which carcinogenesis is considered as an evolutionary process. All models and methods described in this thesis are evaluated with synthetic data and applied to biological data.

Contents

1. Introduction

1.1. Motivation

It has been first proposed in the 1970s that evolution occurs on two levels, since Wilson *et al.* (1974) observed that rates of morphological evolution are weakly correlated with rates of protein evolution. An explanation is that mutations can affect a phenotype by altering coding regions of gene products or by altering regulatory regions which control the level of gene expression. Reasoning from the study by Wilson *et al.* (1974), it was suggested that morphological evolution depends mainly on changes in gene regulation rather than changes in coding sequences. However, most of the geneticists in that time focused their research on the evolution of deoxyribonucleic acid (DNA) sequences, since molecular techniques to explore gene expression on a large scale were not available.

Hence, today evolution of DNA sequences on genome level is widely understood, but mechanisms and evolution of gene regulation which affects the transcript abundance of all genes referred to as the transcriptome are still in its infancy. A difficulty is that the expression of a gene is a continuous trait which depends on several influences, for example, the developmental state, the tissue examined or the environment. Thus, it has to be measured many times under different conditions. Expression of some genes is also influenced by *trans*-effects, resulting from activation or repression by products of other genes. Thus, a single mutation might change the expression level of several genes, since the transcriptome has a very complex structure of dependencies.

Fortunately, new techniques arose in the last decade. Since microarray technology is available, it is possible to measure levels of gene expression for a large proportion of genes of a genome (Baldi and Hatfield, 2002; Speed, 2003). Thus, it is possible to quantify results of gene regulation at a time-point in a tissue. Since diseases like cancer affect the transcriptome, a large number of medical studies were carried out. For example, gene expression between normal tissues and tumour tissues or gene expression between untreated tissues and tissues under drug response were compared (cf. Driscoll *et al.* (2003); Dudoit *et al.* (2002); Golub *et al.* (1999); Li *et al.* (2001); Ramaswamy *et al.* (2001)). An important goal is to discover the mechanisms of cancer and other diseases to enable improved diagnoses and to find new methods of treatment. Beside these medical applications, microarrays are also an appropriate tool to address the pre-discussed problem of exploring the evolution of gene expression. Thus, the technology has been applied in a rich variety of studies to identify gene expression variation within species and expression divergence between species to infer mode and rate of evolution on the level of transcriptome.

Within species variation was observed, for example, for yeast (Cavalieri *et al.*, 2000), *Drosophila* (Jin *et al.*, 2001; Nuzhdin *et al.*, 2004; Gibson *et al.*, 2004; Wayne *et al.*, 2004), teleost fishes (Oleksiak *et al.*, 2005), mice (Enard *et al.*, 2002; Schadt *et al.*, 2003), and human (Enard *et al.*, 2002; Morley *et al.*, 2004; Storey *et al.*, 2007) (cf. a review by Ranz and Machado (2006)). In some cases a large proportion of genes showed significant differences in gene expression among individuals, for example, Storey *et al.* (2007) observed that 83 % of the genes in human individuals are differentially expressed, while 17 % of the genes between human populations are differentially expressed. Oleksiak *et al.* (2005) observed in heart tissue of teleost fish *Fundulus heteroclitus* that 94 % of the genes are significantly different among individuals. Further, it was suggested that differing life conditions can cause gene expression differences, for example, adaptation of teleost fish species to different water temperatures (Oleksiak *et al.*, 2002). A fraction of measured variation in species is the result of reactions to environmental and internal influences. Variation can correlate with phenotypic differences or can be heritable. It is a great challenge to identify the non-mutational effects and to distinguish them from gene expression changes resulting from mutations on DNA sequence level. This is also important when studying differences between different species in order to observe changes caused by evolution.

Divergence between species was examined, for example, between different *Drosophila* species (Rifkin *et al.*, 2003), different teleost fish species (Oleksiak *et al.*, 2002), different mice species, and different primate species (Enard *et al.*, 2002). A frequent observation is that expression divergence between species differ the more the more time has passed, since species split from an ancestor. Rifkin *et al.* (2003), for instance, reported for *Drosophila* species during metamorphosis that the number of genes with significant changes in developmental expression between two lineages are consistent with the genetic distance. In a study by Khaitovich *et al.* (2004) it was observed that gene expression differences between human, chimpanzee, orangutan, and rhesus macaque accumulate approximately linearly with time in brain and liver tissues. The same results were obtained for three mice species (Khaitovich *et al.*, 2004). However, Gilad *et al.* (2006b) used multi-species arrays containing probes of the same primate species used in Khaitovich *et al.* (2004) without observing a linear trend. Evidence for a non-constant gene expression evolution between different tissues was found, since an acceleration in human brain in comparison to chimpanzee brain was observed (Enard *et al.*, 2002; Khaitovich *et al.*, 2005b). Nevertheless, most studies indicate an approximately constant increase of expression differences with time (cf. review by Khaitovich *et al.* (2006)) which can be explained by a neutral model. According to the neutral theory by Kimura (1983) the majority of genetic changes on DNA sequence level are selectively neutral so that most of the genetic variability is the result of an equilibrium between mutations and genetic drift. Indeed, the theory alludes to the genome, but as a result of the previous cited studies it seems to be reasonable to apply it to the transcriptome. However, it is assumed that the expression of some genes evolved non-neutral, but under selection. Thus, since the neutral theory constitutes a null hypothesis, it can be used to identify those genes which are under selection.

The identification of gene expression differences under selection was addressed in numerous studies. Rifkin *et al.* (2003) found evidence for stabilising selection as the henpecking mode in *Drosophila*, but they found also genes which indicate directional selection or neutrality. Lemos *et al.* (2005) analysed different data sets of mice, *Drosophila*, and primates with an ANOVA (= Analysis of variance) based method (Kerr *et al.*, 2000) and came to the conclusion that the majority of genes evolves under stabilising selection. As stated above, analysis of different primate species indicates a faster evolution in human brain in comparison to chimpanzee brain (Enard *et al.*, 2002; Khaitovich *et al.*, 2005b). Further, an overall up-regulation of expression was observed in human in comparison to chimpanzee (Caceres *et al.*, 2003; Hsieh *et al.*, 2003). These results can be explained by directional selection affected some genes during evolution. However, other reasons were discussed, for example, effects of differential hybridisation. Furthermore, positive selection in primate testes was suggested by Khaitovich *et al.* (2005a). It is noted that selection does not reject the applicability of the neutral theory (Kimura, 1983).

For a better understanding of gene expression evolution it is indispensable to describe it by appropriate models. To this end, a simple neutral stochastic model has been developed by Khaitovich *et al.* (2005b). This thesis discusses the development of advanced stochastic models to describe the evolution of gene expression in more detail. The model by Khaitovich *et al.* (2005b) is used as a starting point. Based on extensions of that model like including non-mutational effects or more complex descriptions of mutation effects, new analysis methods for gene expression data are suggested. The performance of the extended models is evaluated with synthetic data and applied to biological samples of different species. This thesis is focused on the evolution of nearly related species and also on the processes leading to variability within species which is related to the field of population genetics. Furthermore, a medical application of the models is presented in an additional chapter. All models and methods are implemented into two software package termed EMOGEE and EMOGEE Tools.

1.2. Organisation of the thesis

Chapter 2: This chapter comprises an introduction to the biological and mathematical background. First, the neutral theory is explained and motivated. Subsequently, gene expression is described followed by a section about the microarray technology which is used to measure the level of gene expression on a large scale. The third part explains the relevant mathematical theory. It starts with basic facts about stochastic models and describes specially the basic evolutionary models used in bioinformatics. Additionally, methods to estimate model parameters from data are illustrated. Furthermore, the Wright-Fisher model and the coalescent process which are necessary to understand chapter 5 are explained. Finally, optimisation methods used in this thesis are explicated.

Chapter 3: The gamma distribution is introduced as an alternative characterisation of the mutation effects. The gamma distribution is defined by two parameters (shape and scale) and allows a more flexible description of observed data than the basic model. Since analytic estimation of the parameters is not possible in this extended model, an optimisation method is presented.

Chapter 4: Here so-called non-mutational effects are included into the analysis of gene expression. These effects are caused by metabolism in the cells and measurement errors. They overlay expression changes depending on mutations. The problem is addressed by adding a normal distributed error which summarises all non-mutational effects. For parameter estimation two methods are suggested: (1) a χ^2-fit method and (2) a maximum-likelihood (ML) method. The incorporation of non-mutational effects provides a better fit to the data than the basic model. Furthermore, a simple detection method for single genes mutated in their regulatory region is suggested.

Chapter 5: The neutral theory assumes that the majority of new variants arisen from mutations are not affected by selection. However, in some cases these variants can have real advantages or disadvantages. For homologous gene sequences sampled from a natural population statistical tests exist to find out if genes evolved under the influence of selection. One of such tests, the Tajima's D test, is adapted to be used for the analysis of gene expression data from individuals sampled from populations. To this end, the underlying stochastic model which is based on discrete sequence evolution in the origin test is replaced by the gene expression evolution model.

Chapter 6: While chapters 3–5 address evolutionary questions, a medical application is considered in this chapter. The emergence and development of cancer which depend on mutations on DNA sequence level change the level of gene expression of many genes. Thus, the carcinogenesis is regarded here as an evolutionary process which is described by the same models as the evolutionary data in the previous chapters. The mutation detection method described in chapter 4 is applied to medical data. The genes with mutations in their regulation are used for clustering. The results are compared with the SAM (= Significance analysis for microarrays) method widely used for gene expression analysis.

2. Background

2.1. The Neutral Theory

2.1.1. Definition

“The neutral theory asserts that the great majority of evolutionary changes at the molecular level, as revealed by comparative studies of protein and DNA sequences, are caused not by Darwinian selection but by random drift of selectively neutral or nearly neutral mutants. The theory does not deny the role of natural selection in determining the course of adaptive evolution, but it assumes that only a fraction of DNA changes in evolution are adaptive in nature, while the great majority of phenotypically silent molecular substitutions exert no significant influence on survival and reproduction and drift randomly through the species.” (Kimura, 1983)

Today, the *neutral theory* (or rather the neutral mutation-random drift hypothesis) has been accepted by most scientists. This was not always the case. When the theory was proposed in the 1960s, it led to a great dispute known as the ’neutralist-selectionist controversy’, since it was different to the former view of the *synthetic theory* (Campbell and Reece, 2005). To explain this development of evolutionism, it is necessary to go back more than one century in time.

2.1.2. Formation of the synthetic theory

At the beginning of the 19th century it was generally assumed that all species had been once created and do not evolve. Lamarck was the first to consider evolution. In 1801 he suggested that individuals lost traits which are not used and develop traits by strengthening through use. The use and disuse depends on the environment. Further, he assumed that acquired attributes are inherited to the offspring which induces a direction in evolution. The modern theory of evolution goes back to Charles Darwin (Darwin, 1859). He discovered a large number of fossils and species and recognised variations between individuals of the same species. He came to the conclusion that some individuals of a species have a better chance to survive than others which is due to small advantages depending on environmental factors (e.g., white rabbits have an advantage to grey rabbits in snow landscapes to hide out from predators). Thus, these individuals have a higher chance to breed. The offspring inherits the advantages and the population fits to the environment better and better over generations. This process is referred to as *natural selection*. By separation into different isolated sub-populations with independent development new species can emerge. However, in the time when the theory was suggested, it was neither known how variations within species arise nor how inheritance works.

Darwin's theory was doubted strongly after publication and it took decades until the basics of mutation and natural selection were approved. More and more people believed in these mechanisms, since new evidences were found. As an example, Weldon (1901) first reported *stabilising selection* instancing the number of turns in snail shells (cf. chapter 2.1.5). More evidence came from cell biology and genetics, for example, the rediscovery of *Mendelian rules* of inheritance in 1900 or the discovery that genes are located on the chromosomes by Thomas Hunt Morgan in 1908. All these findings have been merged with Darwin's evolution theory to the so called 'synthetic theory'. The claim of that theory is that *directional selection* is the mainspring of evolution (Campbell and Reece, 2005).

From the beginning of the 20th century biologists and mathematicians started to develop models and methods to study evolution. To this end, Pearson, influenced by Weldon, started to develop statistical methods like the χ^2*-method*. Hardy (1908) and Weinberg (1908) established science which was called *population genetics* later with an equation to describe the equilibrium of genotypic frequencies under the assumption of random mating and Mendelian inheritance. Fisher (1922) first used stochastic methods to describe fluctuations in gene frequencies by random sampling of gametes. He also developed the *diffusion equation* method. In contrast to others in due time, Wright recognised the importance of the disappearance of alleles in evolution by chance. He developed the later called *shifting balance theory* of evolution (Wright, 1932). Wright and Fisher also introduced the important *Wright-Fisher model* to describe genealogical relations within populations (Fisher, 1930; Wright, 1931) (this model is discussed in detail later in this chapter). Haldane suggested a method to estimate the time for a dominant allele to spread in a population from a small ratio of carriers to a larger one, but without applying random effects. Later, he extended his work by addressing various factors on the change of gene frequencies (see Haldane (1932)). Together, Fisher, Wright, and Haldane nearly completed the field of classical population genetics in the 1930s.

2.1.3. The emergence of the neutral theory

In the 1960s the *molecular biology* arose and new techniques were developed: It became possible to compare amino acid sequences of closely related organisms and to estimate substitution rates (Zuckerkandl and Pauling, 1965). Furthermore, the variability of enzymes among individuals could be detected by *electrophoretic techniques* (Harris, 1966; Lewontin and Hubby, 1966). These improvements made it possible to explore evolution on a large scale by estimating rates of amino acid substitutions and genetic variability. In that time it was expected that these new findings would confirm directional selection as the most important factor in evolution.

But disbelief emerged by applying the mathematical theory. An approximately uniform rate of amino acid substitutions per year for each protein between different lineages was observed. Substitutions seemed to have random patterns and the rate seemed to be very high. Kimura discovered that an unusually high rate of production of advantageous mutants would be required to explain the high rate of molecular evolution. However, this conflicted with the observation of constant substitution rates over long time periods, since a population would fit the environment perfectly someday. Furthermore, no visible correlation between the high variability within populations and environmental factors was found.

Due to these facts Kimura suggested the later-called 'neutral theory' (first published in Kimura (1968a,b)). He concluded that the majority of nucleotide substitutions must be the result of random fixation of selectively neutral mutants rather than Darwinian selection. Kimura got great encouragement for his thesis, but also much criticism. Kimura's arguments for the neutral theory depended on results from the analysis with population models. It can be shown by simulations that for a large population also a large number of mutants arise. Admittedly, the majority of these mutants are lost by random drift within a few generations. Thereby, it does not matter whether a mutant is deleterious, neutral or advantageous unless the advantage is exceedingly large.

In evolutionary models the fitness of an allele is described by a numerical value, the so-called *fitness value.* Kimura developed equations to calculate a critical value for the effect of a mutation on the fitness value: Mutations which change the fitness value with a magnitude below this critical value are called *nearly neutral*, since they have no significant effect on the fate of the allele. If a mutant has a fitness value which differs below the critical value from the fitness value of its ancestral allele which is fixed in the population, the fate of the mutant to establish in the population or to disappear is determined by chance.

It is assumed that the allele composition of a population changes approximately constantly over time. This explains the correlation of the number of substitutions between species and the time since these species emerged from a common ancestor. This continuously changing gene pool is known as *random genetic drift.*

2.1.4. Further cases for the neutral theory

A lot of criticism of the neutral theory depends on misunderstandings. The neutral theory does not state that genes which evolve neutrally have no function. It merely claims that the mutant forms of these genes are selectively nearly equivalent to the precursors. An explanation for the assumption that a majority of mutations have no major effect is the *physiological homeostasis.* Especially in higher organisms this equilibrium of bodily functions is a buffer against external dysfunctions from the environment and also against internal incidents (Kimura, 1983). Thus, a slightly disadvantageous mutation could be regarded as an internal incident which can be adjusted by the homeostasis.

The neutral theory also does not contradict the existence of selection. It merely constitutes that the majority of mutated alleles is not affected by selection effects, since the influence on the fitness is too small. However, there might be a large number of mutations which are deleterious, but these mutations are weeded out from the population and therefore evolution is not affected.

Some mutations on sequence level are apparently neutral. This is due to the redundancy of the genetic code which maps three bases of a gene sequence to one amino acid of the corresponding protein (cf. chapter 2.2.1). The sequence is composed of four types of bases: Adenine (A), cytosine (C), guanine (G), and uracil (U). The genetic code is shown in table 2.1. Particularly, mutations in the third base do not lead to an amino

Table 2.1.: The genetic code. Three bases encode one amino acid, when a protein is built. The bases are abbreviated: U = uracil, C = cytosine, A = adenine, G = guanine. The three bases for methionine symbolise the starting position at which the translation begins. The sequences UAA, UGA, and UAG stop the translation (cf. chapter 2.2.1).

1st base	2nd base	3rd base U	3rd base C	3rd base A	3rd base G
U	U	Phenylalanine	Phenylalanine	Leucine	Leucine
	C	Serine	Serine	Serine	Serine
	A	Tyrosine	Tyrosine	STOP	STOP
	G	Cysteine	Cysteine	STOP	Tryptophan
C	U	Leucine	Leucine	Leucine	Leucine
	C	Proline	Proline	Proline	Proline
	A	Histidine	Histidine	Glutamine	Glutamine
	G	Arginine	Arginine	Arginine	Arginine
A	U	Isoleucine	Isoleucine	Isoleucine	Methionine (START)
	C	Threonine	Threonine	Threonine	Threonine
	A	Asparagine	Asparagine	Lysine	Lysine
	G	Serine	Serine	Arginine	Arginine
G	U	Valine	Valine	Valine	Valine
	C	Alanine	Alanin	Alanine	Alanine
	A	Aspartic acid	Aspartic acid	Glutamic acid	Glutamic acid
	G	Glycine	Glycine	Glycine	Glycine

acid substitution in many cases. Thus, these mutations have no effect on the amino acid sequence. They are termed as *synonymous mutations* (in comparison to *non-synonymous mutations* which change the amino acid sequence). However, the speed of synthesis of the sequence is influenced by the quantity of so-called *transfer ribonucleic acid* (tRNA) molecules which are used to translate the genetic code (cf. chapter 2.2.1). Some tRNAs are more frequent than others which depends on the organism. This phenomenon is referred to as *codon bias* and might cause selection effects (Bulmer, 1991).

The neutral theory is widely used as a ”null model” for statistical tests to detect selection during evolution. Although the majority of mutations are neutral, some mutations have significant positive or negative effects on the fitness of an individual so that the fate of the mutated allele cannot be explained by chance. Please note that extremely deleterious effects cannot be recognised, since carriers of such mutated alleles became extinct. In most of the tests within or between species variation in sequences or other traits are estimated. Thereupon, this variation is compared to the expected variation estimated from a neutral model. A significant difference in this comparison rejects the neutral hypothesis and indicates selectional effects. Examples of tests are the followings: The *HKA test* (Hudson *et al.*, 1987) compares the ratio of within species diversity and between species divergence. The *McDonald-Kreitman test* (McDonald and Kreitman, 1991) considers the ratio of synonymous and non-synonymous base substitutions. *Fu's test* (Fu and Li, 1993) compares the number of mutations on internal and external branches of a genealogy. Finally, the *Tajima's D test* (Tajima, 1989) compares two different estimators for the mutation rate within a population. It will be explained in chapter 5.

2.1.5. Modes of selection

It is assumed that a gene can be under different modes of selection: Stabilising selection (also termed as negative, normalising, purifying or centripetal selection) occurs if the trait corresponding to the gene has reached the optimal characteristic concerning the *ecological niche* which means that the carrier of that allele fits the environment perfectly. All other alleles of that trait are inferior. Thus, their frequencies fall off over the time and the chance to become fixed in the population is poor. If environmental influences change, the adaptation is getting suboptimal which leads to a trend explained by directional selection (also termed positive selection). Related to the new environment, a mutation leading to a new allele might be superior in comparison to the former allele. Thus, the frequency of the new allele is increased and the trait changes towards an optimum. In some cases the combination of two different alleles of a gene provides the highest fitness in a diploid population in which individuals contain two sets of chromosomes. Although one of the alleles might have a weak fitness in the homozygous case in which an individual contains two identical alleles. However, both alleles become fixed in the population at stable frequencies which is called *balancing selection.*

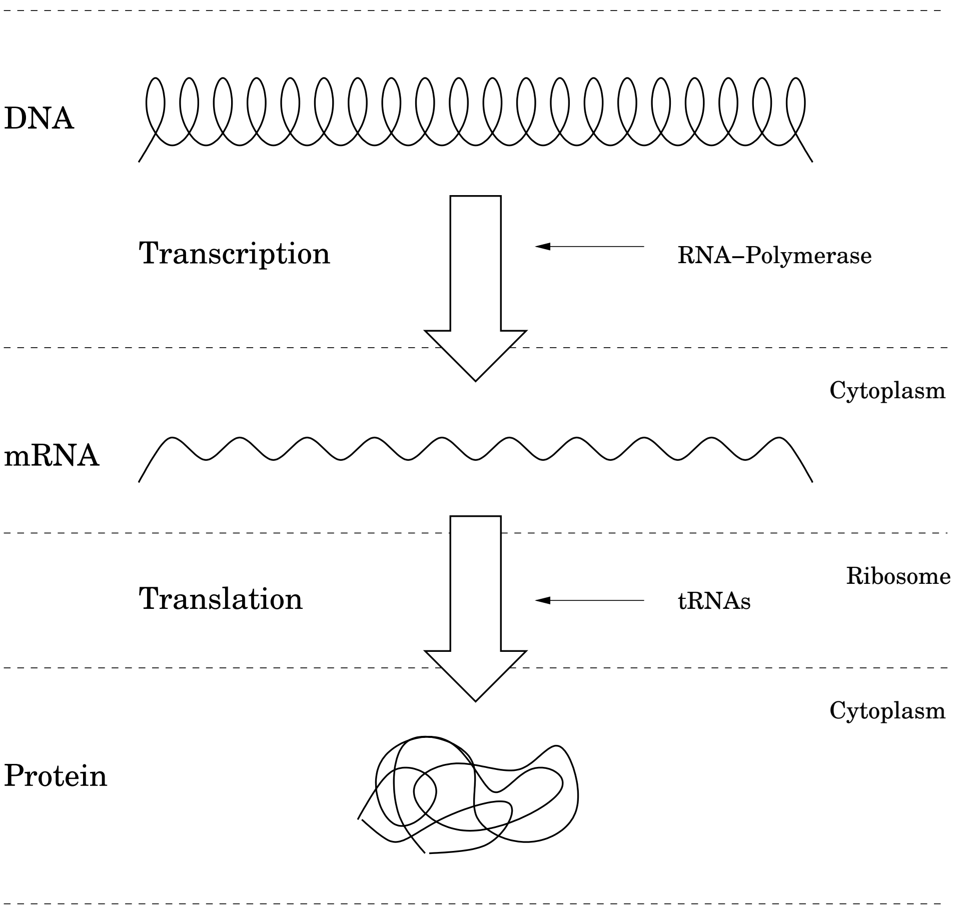


Figure 2.1.: The process of gene expression.

2.2. Gene expression

2.2.1. The process of gene expression

Gene expression is the process in which the DNA sequence of a gene is used to provide cell structures and functions. This complex process consists of different steps depending on the gene product and the type of the cell, for example, in cells with a nucleus (*eukaryotes*) the process is more complex than for cells without a *nucleus* (*prokaryotes*). The gene products can be very different, since a gene can encode a structure protein, an enzyme protein or a special RNA molecule which is necessary for different cell functions. For proteins the process is also called *protein biosynthesis*. The steps of gene expression are illustrated in figure 2.1. In all cases the first step is the *transcription*.

In the transcription a DNA region containing a gene is copied into an complementary single-stranded *ribonucleic acid* (RNA) molecule by an *RNA polymerase* enzyme. The process begins by the binding of a RNA polymerase to the so-called *promoter* region of that gene which should be transcribed. From the start the RNA polymerase moves base-by-base along the DNA and catalyses covalent linking of ribonucleotides which match with the current position of the DNA sequence. A is transcribed into U, T (= thymine) into A, C into G, and G into C (in comparison to DNA, in RNA uracil is used instead of thymine). The process of transcription stops if a the RNA polymerase reaches a specific termination sequence. The resulting RNA molecule is called *messenger RNA* (mRNA). In eukaryote cells it is referred to as *pre-mRNA*, since the so-called *splicing* is performed as an additional step which results in the (mature) mRNA. During the splicing some parts, the *introns*, are removed from the RNA sequence, while the *exons* remain there. In some cases different variants of mRNA are produced by skipping or shuffling exons in the sequence which is referred to as *alternative splicing*.

The mRNA is used for the *translation* which takes places at small cell organelles called *ribosomes*. In the translation mRNA is decoded to produce a specific polypeptide according to the genetic code. In that code, three bases which are called *codon* encode one amino acid (cf. table 2.1). A sequence of non-overlapping codons encodes a series of amino acids which are bind to polypeptides. Important for the decoding are transfer RNAs (tRNAs) which bind to specific amino acids. A tRNA has a three-base-pairs long region, the *anti-codon* which is specific for the bonded amino acid. The anti-codon matches with a codon of the mRNA by complementary base pairing so that a tRNA molecule translates the sequences of three bases into a corresponding amino acid. The translation process starts with a special codon on the mRNA with the sequence AUG, the *starting symbol*. Then tRNAs with bonded amino acids bind to the matching codons in series. The amino acids of adjacent tRNAs form peptide bonds and are released from the tRNA which leads to a growing amino acid chain. The translation stops at a so-called *stop codon* which can be UAA, UGA or UAG.

The amino acid sequence adopts a three-dimensional structure which enables its function, for example, to be a structure protein in the cell membrane or an enzyme protein with a special role in metabolism. The level of gene expression which results in the level of protein production or directly used RNAs depends on many different influences like metabolic pathways or epistatic effects, but also on the tissue, the developmental stage or the physiologic state of the cell. It is controlled by complex mechanisms like different affinity of the RNA polymerase to the promoter region, inhibitors or activators which influence the chance of transcription. Inhibitors and activators are even proteins so that expression levels of different genes are not independent from each other. It is an important issue to understand the regulation of gene expression, since changes in the expression levels of some genes play an important role in different diseases.

Detailed introductions into the process of gene expression are described in the books Alberts *et al.* (2002), Campbell and Reece (2005), and Griffiths *et al.* (2002).

2.2.2. Measuring the level of gene expression with microarrays

The amount of mRNA copies of a gene present in the cell is a quantitative measure for the level of expression of that gene, since mRNAs are digested in the cell plasma (Campbell and Reece, 2005). The amount is also an indirect measure for the corresponding protein, but this is very imprecise, since some proteins, for example, structure proteins, exist in the cells for a long time after the mRNAs have been digested.

Microarrays are a valuable tool to measure the level of gene expression. Typical microarrays are assemblies of thousands of small spots of DNA on a solid surface (e.g., glass). The spots are arranged in a rectangular array. Each spot contains a large number of immobilised copies of a particular sequence which are referred to as *probes*. They exclusively correspond to a segment of one gene of a species. The expression level of that gene can be measured. Depending on the large number of spots, the expression of thousands of genes can be measured simultaneously with one microarray. The resulting data is termed *expression profile*. When a microarray is used to measure the level of gene expression in a tissue, the mRNAs are extracted from the tissue. These mRNAs are typically converted to *complementary DNAs* (cDNA) which are the reverse transcripts of the mRNAs. The cDNAs are amplified and used to produce *complementary RNAs* (cRNA) which are RNAs transcribed from cDNAs. The cRNAs are cut into small fragments and labelled with fluorescent dye. Subsequently, the labelled cRNAs are attached to the microarrays. Sequences of cRNA which are complementary to the DNA in the spots hybridise with them. After being cleaned from remaining unbound cRNAs, the microarrays are scanned with special scanners to produce digital images of the spots. The fluorescent intensity of a spot is a measure for the amount of cRNAs complementary to the gene which is represented by the spot. Therefore, it is a measure for the level of expression of that gene.

Microarrays differ in array surface, number of measurable genes and labelling methods.

Two widely used technologies are the *oligonucleotide* arrays from Affymetrix (Affymetrix, 2004) and *spotted arrays*. Affymetrix produces ready to use chips for different organisms. The probes are synthesised by a masking technology *in situ* on the chip. The arrays have 25 bases long oligonucleotides as probes which are representative for one gene. For each gene between 11 and 20 spots of oligonucleotides from different sections of the gene exist which are distributed over the chip. These spots are called *perfect matches* (PM), since the oligonucleotides are complementary to the subsequences of the gene. For each oligonucleotide exits one additional spot with mismatching oligonucleotides (*mismatches*, MM). They differ in the 13th base and are used to detect non-specific hybridisation. All probes for one gene (PM and MM) form a *probe set*. Another technology are spotted arrays which are typically custom-designed. A robot spotter fills small quantities of oligonucleotides or longer sequences on a glass slide. The quality of the spots is not as good as those from Affymetrix chips. An advantage is that one chip can be used to measure the expression in two tissues simultaneously, for example, one tissue of a special phenotype of interest against one of its normal counterpart. In this case a different dye is used for the cRNAs of each tissue (e.g., green and red). The colour mixture of the result show which genes have increased or decreased their expression in the phenotype in comparison to the normal case.

A more detailed description of microarray technologies is described in the books by Baldi and Hatfield (2002) and Schena (2003).

2.2.3. Analysis of microarray data

Since the number of spots on a microarray is very large, visual inspection is infeasible. Thus, computer analysis is necessary. Therefore, a large number of methods have been developed. Their application depends on the type of experiment. In most cases two or more phenotypes, for example, normal tissue and tumour tissue or tissues of different subtypes of a disease are profiled with microarrays to compare the difference in gene expression intensities between the phenotypes. Another application is the measurement of the change of gene expression during the course of a disease or during drug treatment. Figure 2.2 shows the typical workflow of a microarray analysis. At first, the microarrays are scanned after hybridisation to get digital images of the fluorescence intensities. Figure 2.3 shows an enlarged image of a typical Affymetrix microarray surface after hybridisation.

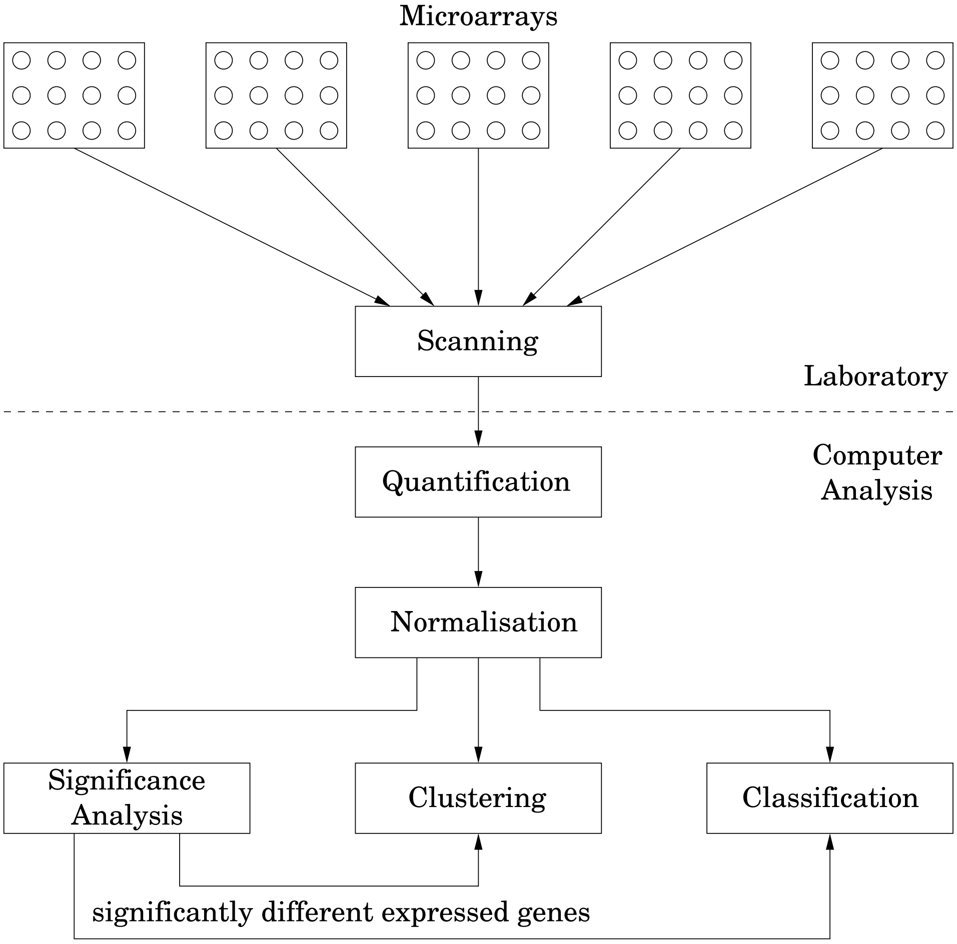


Figure 2.2.: Typical workflow of a microarray analysis. After scanning, quantification, and normalisation the expression profiles can be analysed in different ways. Oftentimes only the significantly expressed genes detected by a significance analysis are used for clustering or classification.

After scanning, it is necessary to convert the spot intensities from digital images into numerical quantities. This process is called *quantification* (Schena, 2003). Since scanning is performed with high resolution, each spot of a microarray is represented by a large number of pixels. A common method is to take the mean colour intensity of the inner spot area as the *signal* which represents the gene expression level. The mean intensity of the diffuse border area is called *background*. It results from inferring biochemical events like substrate reflection and is regarded as noise. The ratio of signal and noise can be used as a quality criterion. Some methods subtract the background from the signal to correct the intensity, since spurious events also take place in the inner spot area. Saturated spots which exceeds an upper limit of accurate intensity detection are problematic. Thus, all

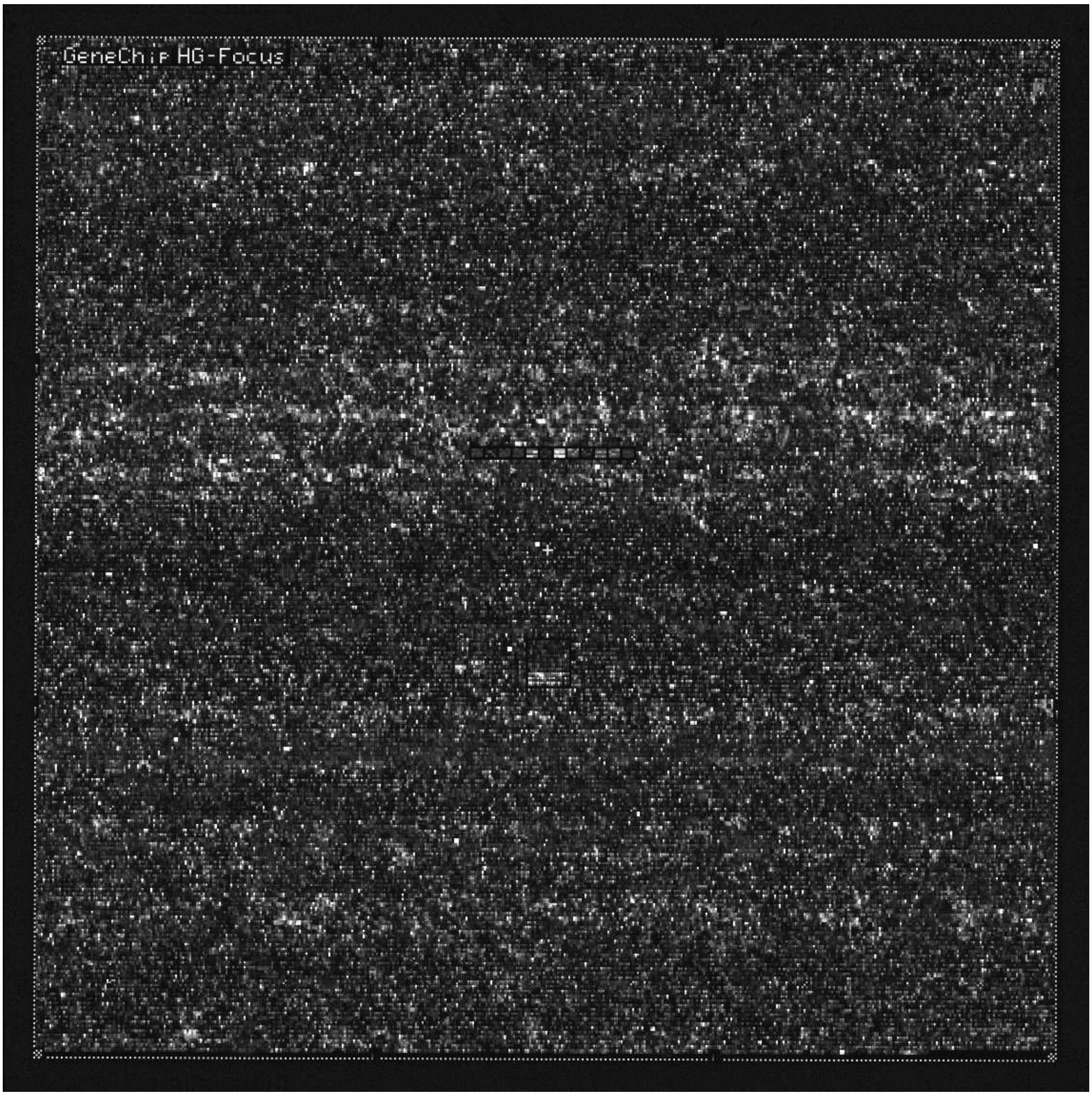


Figure 2.3.: Scanned image of an Affymetrix HG-Focus microarray after hybridisation (the image originates from the University Hospital Düsseldorf). The picture has been enlarged. The scanned area of a HG-Focus array has a size of about 0.9 cm × 0.9 cm.

saturated spots are identical in appearance and cannot be compared. Ways to avoid saturation are the limitation of quantity of RNA which is put on the chip and the choice of a shorter time interval for hybridisation.

After quantification one can compare the data from different microarray experiments. To this end, *normalisation* of spot intensities is necessary to make the expression values comparable, for example, to detect those genes which changed their expression level between two observed phenotypes. If the expression values are distributed linearly, a linear function can be applied to set the mean intensities of all expression profiles to the same level. For non-linear data it is necessary to separate the range of the distribution in different parts and scale these parts with different linear functions or alternatively to apply a polynomial. Further, for Affymetrix arrays it is necessary to combine the intensities of all probes of a probe set to one absolute measurement which represents the expression of the corresponding gene. A very simple method takes the mean value of the differences of the PM and the corresponding MM probes. However, some complex model-based normalisation methods have been developed (e.g., by Chu *et al.* (2002)). Widely used methods performing absolute measurement calculation and normalisation for Affymetrix microarray data which are implemented in software packages are the RMA (= Robust Multichip Average) method (Bolstad *et al.*, 2003; Irizarry *et al.*, 2003a,b) and the VSN method (= Variance stabilisation and calibration for microarray data) (Huber *et al.*, 2002). The former one uses quantile normalisation on the individual probe intensities with a different form of background correction not taking MM probes into account. The latter one uses the method of variance stabilising transformations for normalisation.

After normalisation users often want to detect genes which show significant expression differences between the examined phenotypes, since this might indicate a biological reason for the phenotypical difference. A simple *significance analysis* is the *t-test.* The null hypothesis of this test is that the mean values of two distributions are identical (the distributions are assumed to be normal distributions). To apply it to gene expression data it is assumed that the distributions of expression values of a gene in a number of expression profiles of two phenotypes have the same mean. Then the t-test is used to calculate a critical value on a specified significance level. If this limit is exceeded, the null-hypothesis is rejected and the considered gene expression levels are significantly different. A more sophisticated analysis is implemented in the *SAM* package (Significance Analysis for Microarrays) (Tusher *et al.*, 2001). This method is based on a permutation test and can be applied also to multi-class data sets which contain more than two different phenotypes. In case of multi-class data also the *ANOVA* method (= Analysis of variance) is widely used (Kerr *et al.*, 2000).

While the significance analysis filters out those genes which might be non characteristic for the examined phenotypical difference, *clustering* methods are used to show the relations between the expression profiles. Clustering belongs to *unsupervised learning.*

Thereby, it is assumed that one has no *a priori* knowledge of the data analysed. Thus, for microarray analysis the phenotypes of the tissues are not taken into account. An expression profile is commonly regarded as a vector in which each dimension represents the expression level of one gene. Different clustering methods are used: In *hierarchical clustering* (Johnson, 1967) the expression profiles are ordered in a tree-like structure. In the beginning, each expression profile is regarded as one cluster. The two most similar clusters are then summarised to one cluster. This procedure is iterated until one cluster (the root of the tree) remains. The *k-means* method (McQueen, 1967) maps the set of expression profiles to k groups which are represented by the *centroids* (average vector) of its members. The method is iterative and starts with a random mapping and subsequent centroid calculation. Afterwards, each expression profile is mapped to the class which is equal to that one of the nearest centroid, and the centroids are calculated again. The method stops if no class membership changes during one iteration. More complex are *self-organising maps* (Kohonen, 2001). Thereby, the expression profiles are mapped to a two-dimensional grid. Similar expression profiles are mapped to the same or nearby positions. Before any clustering method is performed, the dimension of the gene space is typically reduced by significance analysis to lower the noise in the data. Instead of clustering the expression profiles in the gene space it is also possible to cluster the genes in the space of expression profiles.

In comparison to clustering, methods of *classification* belong to *supervised learning*: A training set of expression profiles is compiled in which each expression profile is labelled with a so-called *class label* representing its phenotype. Then a classification algorithm is trained onto this set to build a *classifier*. The classifier is then applied to expression profiles of unknown state to make a class prediction. Different classification methods are common. Like in clustering methods the expression profiles are regarded as vectors. In *k-nearest neighbours* (Cover and Hart, 1967) a new expression profile is assigned to the class to which the majority of the k nearest neighbouring expression profiles belongs. *Support Vector Machines* (Christianini and Shawe-Taylor, 2000) use a training phase to calculate a hyperplane which separates the two classes of expression profiles (the basic form is used for binary classification, but more complex variants exist). To classify an expression profile, its position to the hyperplane is calculated and thereupon a decision is made. Another method to build a classifier is *Genetic Programming* (Banzhaf *et al.*, 1998; Koza, 1992), a machine learning method inspired by the biological evolution. The classifiers are represented by computer programs which evolve many iterations based on evolutionary principles. Starting from random programs, those programs with the best classification accuracy on the training expression profiles are selected for the next iteration. These programs are copied and changed slightly by random (mutation). Furthermore, program code from different programs is merged to new programs (recombination). Then the accuracy on the training set is checked again. An important application of classification in future might be the subtype prediction of similar cancers which is often difficult with histological inspections, but necessary to enable the best possible therapy.

More detailed introductions into the field of microarray analysis are Baldi and Hatfield (2002), Knudsen (2002), and Speed (2003).

2.3. Stochastic models for evolutionary processes

2.3.1. Mathematical background of models and parameter estimation

Stochastic models are widely used in bioinformatics to describe different processes (e.g., growth of a population or evolutionary processes). Stochastic models are mathematical models which consist of equations to describe quantities. Additionally, they take random effects into account. These effects are specified by random variables with particular probability distributions which are incorporated into the equations describing the model.

Stochastic models can be used to analyse experimental data. In this case it is assumed that the model is an adequate description of the corresponding observed data. The goal is to estimate an assignment for the model parameters which describe the real data best. In other words, the intention is to find the parameter assignment so that data generated by the model has similar properties than the observed real data. To this end, one common way is the *method of moments* (Pearson, 1894). *Moments* are characteristics of a distribution which describe their shape and scale, for example, the *mean* (first moment), the *variance* (second central moment which is a measure for the width), the *skewness* (third central moment which is a measure for symmetry), and the *kurtosis* (fourth central moment which is a measure for the concavity). Moments can be described by equations. These equations characterise the expectations for the moments depending on the model parameters. To estimate the model parameters for real data, the moment equations are equated with the observed moments from the data. Then the equations are transformed so that the parameters can be computed by the observed moments.

The equations for the moments of a stochastic process can be derived by the so-called *characteristic functions.* A characteristic function for a continuous random variable X is a mapping $\varphi_X : \mathbb{R} \to \mathbb{C}$ which is defined by

$$\varphi_X(t) = E[e^{itX}], \tag{2.1}$$

whereas $t \in \mathbb{R}$ and E is the expectation (i is the imaginary unit of the complex numbers). For discrete random variables the characteristic function is given by

$$\varphi_X(t) = \sum_{k=1}^{\infty} e^{itx_k} P(X = x_k). \tag{2.2}$$

If the characteristic function of a continuous random variable X can be differentiated m times, the m-th moment will be generated (details are described in Feller (1957)):

$$E[X^m] = (-i)^m \varphi_X^{(m)}(0). \tag{2.3}$$

Another way to estimate model parameters from data is the *Maximum-likelihood* (ML) method (cf. Bickel and Doksum (2001)). The method maximises the so-called *likelihood* function. The likelihood is the probability for the observed data given a parameter assignment. Assume that a model with parameters Θ from a parameter space T generates data x. This induces a family of probability density functions $x \to p(x|\Theta)$. The likelihood function is

$$L(\Theta|x) = p(x|\Theta). \tag{2.4}$$

Thus, the ML is

$$max(L(\Theta|x)) = max\{p(x|\Theta)|\Theta \in T\}. \tag{2.5}$$

To find the maximum, an optimisation method is used (cf. chapter 2.4).

Parameters of a model can also be estimated by a *Bayesian method* which depends on the *Bayes' theorem* for two stochastic events A and B with probabilities $P(A)$ and $P(B)$, respectively:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.6}$$

$P(A|B)$ is the so-called *posterior probability.* $P(B|A)$ is the conditional probability of B given A which is the likelihood function for B given A. $P(A)$ is the *prior* probability of A. It does not take B into account. The prior probability $P(B)$ acts as a normalising constant. Let now Θ be the parameters of a model (corresponding to A in equation 2.6) and x be observed data (corresponding to B in equation 2.6). To estimate the parameters for observed data by a Bayesian method, those values for Θ have to be chosen which maximises the posterior probability $P(\Theta|x)$. To this end, an optimisation method is used (cf. chapter 2.4).

2.3.2. The Poisson process and the compound Poisson process

To model evolutionary processes it is necessary to describe genetic changes caused by mutations. Since mutations are typically random events, the *Poisson process* (cf. Taylor and Karlin (1998) or Ewens and Grant (2001)) is used. The Poisson process describes the number of random events occurring within a time interval. The Poisson process is based on the *Poisson distribution*

$$p_k = \frac{d^k e^{-d}}{k!} \tag{2.7}$$

with the parameter d and $k = 0, 1, \ldots$. The expectation and the variance of the Poisson distribution are both equal to μ.

A Poisson process M is defined by its rate $\mu > 0$ (e.g., the mutation rate) and is referred to as $M(t)$, whereas t is a time point with $t \geq 0$. The expected number of events up to time t is Poisson distributed with parameter μt. The process increments $M(t_1) - M(t_0), M(t_2) - M(t_0), \ldots, M(t_m) - M(t_{m-1})$ are independent random variables for any time points $t_0 = 0 < t_1 < t_2 < \ldots < t_m$. Thus, for $s \geq 0$ and $t > 0$ the random variable $M(s+t) - M(s)$ has the Poisson distribution

$$Pr\{M(s+t) - M(s) = k\} = \frac{(\mu t)^k e^{-\mu t}}{k!} \tag{2.8}$$

for $k = 0, 1, \ldots$ and $M(0) = 0$. Time can be considered on an evolutionary scale that is real time t scaled by the mutation rate μ which is $d = \mu t$. In this case the process is referred to as a Poisson process $M(d)$.

A more complex process is the *compound Poisson process* (Taylor and Karlin, 1998). Given a Poisson process $M(t)$, an independent random variable is used to describe each event of $M(t)$. These random variables $X_1, X_2, \ldots$ share the same distribution function. They are added up to describe the change of a value over time. Thus, the value at time t is defined by

$$Y(t) = \sum_{k=1}^{M(t)} X_k, \tag{2.9}$$

with $t \geq 0$. If ϕ and σ^2 are the common mean and variance of $X_1, X_2, \ldots$, then mean and variance of $Y(t)$ are

$$E[Z(t)] = \phi\mu t, \tag{2.10}$$

$$Var[Z(t)] = (\phi^2 + \sigma^2)\mu t. \tag{2.11}$$

2.3.3. The Wright-Fisher model and the coalescent process

A simple stochastic model to describe genealogical relations of individuals within a population is the *Wright-Fisher model* (Fisher, 1930; Wright, 1931). This important model for population genetics describes the gene transmission in an idealised population from one generation to the next. It is assumed that the population consists of N diploid or $2N$ haploid individuals. Here, haploid individuals are regarded. The entity 'individual' is defined by its gene, so it is referred to as 'gene' here. The basic Wright-Fisher model makes the following simplifications:

1. Discrete and non-overlapping generations

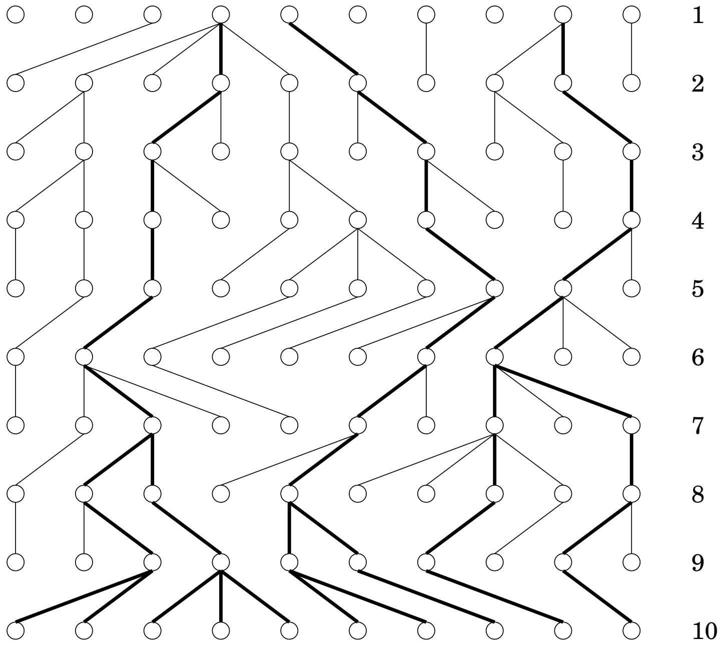


Figure 2.4.: Simulations on the Wright-Fisher model with 10 individuals over 10 generations. The lineages are sorted to get a tree-like structure. Those lineages which have not become extinct are drawn in bold face.

2. Haploid individuals

3. Constant population size

4. Equal fitness of all individuals (this makes it a neutral model)

5. No geographical or social structure

6. No recombination of genes

The model can be simulated as follows: For a generation at time t the next generation at $t+1$ is simulated by randomly selecting $2N$ new genes from the parental genes (without replacement). An example of this process for 10 generations is presented in figure 2.4. The chance to choose one specific gene from generation t as a parent is $1/2N$. Since the number of genes in generation $t+1$ is also $2N$, the chance of a gene to produce v successors is described by a binomial distribution:

$$P(v=k) = \binom{2N}{k}\left(\frac{1}{2N}\right)^{k}\left(1-\frac{1}{2N}\right)^{2N-k} \tag{2.12}$$

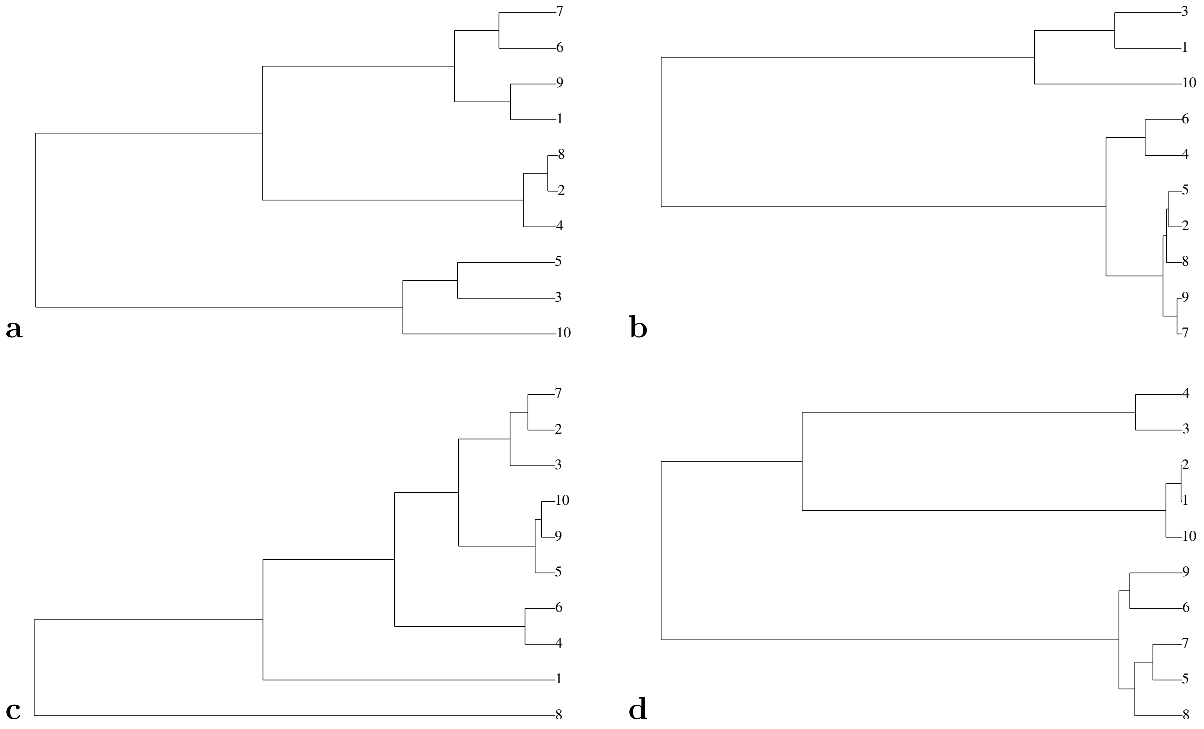


Figure 2.5.: Examples of coalescent trees with 10 genes.

If $2N$ is large, v is approximately Poisson distributed with mean of one and variance of one:

$$P(v = k) \approx \frac{1}{k!} e^{-1} \tag{2.13}$$

The genealogy of a population which has been generated by a Wright-Fisher model can be described by the *coalescent process* (Kingman, 1982; Hudson, 1991). In contrast to the Wright-Fisher model the coalescent process goes back in time. It describes the time to a so-called *coalescent event* in the past at which the lineages of two or more genes from the population descended from a *most recent common ancestor* (MRCA). The coalescence time T_2 for two genes to find a MRCA is described by the geometric distribution

$$P(T_2 = j) = \left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N}, \tag{2.14}$$

whereas $j = 1, 2, \ldots$ is the number of generations to the MRCA back in time. This equation can be extended, for example, to describe the coalescence time T_k for a coalescent event of two genes out of k genes. Since the time is measured in discrete units, the model is called *discrete time coalescent*. In contrast, the *continuous time coalescent* uses the average time that two genes find a common ancestor in one unit of time, whereas one unit of time is $2N$ generations. The distribution function of the waiting time T_k in the continuous representation for k genes to have $k-1$ ancestors is

$$P(T_k \leq t) = 1 - e^{-\binom{k}{2}t} \tag{2.15}$$

for $t = j/(2N)$, whereas j is the time measured in generations. Hence, the waiting time follows an exponential distribution with parameter $\binom{k}{2}$ (referred to as $\texttt{Exp}\left(\binom{k}{2}\right)$). Thus, the expectation of $P(T_k \leq t)$ is $E[T_k] = 1/\binom{k}{2}$. With this equation the simulation of a genealogy of size n which shows the ancestral relations of the n genes can be performed. For this purpose the time to merge two lineages is drawn from $\texttt{Exp}\left(\binom{k}{2}\right)$ for $k = n$. The two genes which merge are randomly chosen from all genes in the sample. $k-1$ lineages remain and the process is iterated with $k \rightarrow k-1$ until the last two lineages merge. Then the coalescent tree is complete. Figure 2.5 shows examples of four coalescent trees with 10 genes. Different properties of coalescent trees have been derived: The expected height H_n of a tree with n genes is

$$E[H_n] = \sum_{j=2}^{n} E[T_j] = 2\sum_{j=2}^{n} \frac{1}{j(j-1)} = 2\left(1 - \frac{1}{n}\right). \tag{2.16}$$

If n tends to infinity, the expected height tends to 2 for a scaling of $2N$ generations. This is only twice of the expected waiting time for two genes to find their common ancestor. The expected total branch length L_n of a coalescent tree with n genes is

$$E[L_n] = \sum_{j=2}^{n} jE[T_j] = 2\sum_{j=1}^{n-1} \frac{1}{j}. \tag{2.17}$$

Coalescent trees can be used together with mutation models to describe the sequence evolution in a population. In this case mutations are simulated along coalescent trees. Since the underlying Wright-Fisher model does not consider selection, simulations of genealogies are useful for statistical tests to create a null-hypothesis distribution. The Tajima's D test (Tajima, 1989) which is used to detect genes under selection in a population is important in this context (cf. chapter 5). A detailed introduction into the Wright-Fisher model and the coalescent process is given in Hein *et al.* (2005).

2.3.4. Mutation models

Two important mutation models are the *infinite alleles model* and the *infinite sites model*. Both contain a parameter θ describing the *population mutation rate*. That is the expected number of mutations between two individuals of a Wright-Fisher population. The infinite alleles model assumes that each mutation creates a new allele. The only piece of information about two alleles is whether they are identical or different. If they are different, nothing is stated about the quantity of the difference. The population mutation rate θ can be estimated from the number of different alleles a in a sample of size n (Ewens, 1972):

$$a = \sum_{j=0}^{n-1} \frac{\theta}{j+\theta} \tag{2.18}$$

The infinite sites model is more complex. It is assumed that each site in a sequence can be subject to not more than one substitution in its entire history. It is a reasonable approximation for slowly evolving sequences and/or short time scales. Hence, each mutation occurs at a different site. Thereby, not the actual nucleotide is regarded, but its state "not mutated" or "mutated". However, in comparison to an infinite alleles model it is possible to count the number of mutations which have occurred between two alleles. The population mutation rate θ can be estimated from the number of *segregating sites* s which are variable positions due to mutations (Watterson, 1975):

$$\theta = \frac{s}{\sum_{j=1}^{n-1} \frac{1}{j}} \tag{2.19}$$

When mutation models are used with genealogies, the number of mutations on each branch is drawn. To this end, a Poisson distribution with parameter $l\theta/2$ is used for a branch length l. The division by two is necessary, since the individuals are connected over two branches by a MRCA. More details are described in Hein *et al.* (2005).

Another group of models are the *finite site models* which describe substitutions in the loci of a finite string. The first finite site model was developed by Jukes and Cantor (1969). In this model all possible mutations of a position are equally likely. Kimura (1980) extended this model to take into account the observation that *transition* events (a substitution from a purine to another purine nucleotide (A↔G) or from a pyrimidine to another pyrimidine nucleotide (C↔T)) occur at a faster rate than *transversion* events (substitution from a purine to a pyrimidine or the other way round). Thus, in the model the probability for transitions is greater than for transversions. Felsenstein (1981) introduced unequal base frequencies which are superior in describing real DNA sequences. Hasegawa *et al.* (1985) combined the two models of Kimura (1980) and Felsenstein (1981).

2.3.5. Models for continuous traits

While DNA sequences have discrete states which are changed by mutations, other traits, for example, body size or extremity length, are described by continuous quantities. The expression of a gene can be regarded as a continuous trait as well. Lande (1976) developed a model for phenotypic selection which acts on quantitative characters. He described selection effects and random genetic drift. Later, Lynch and Hill (1986) described a general neutral model of genetic variance within populations and the rate of divergence of quantitative traits under random drift and mutation.

Some models were originally designed to describe evolution of gene expression. Gu (2004) developed a statistical framework for phylogenomic analysis of gene family expression profiles using Brownian motion: Expression changes are assumed to follow a normal distribution with variance proportional to the time which has passed since the evolutionary process has started. Gu (2004) introduced a number of random variables to describe lineage-specific evolutionary rates, directional trends resulting from directional selection, and dramatic shift which may happen after a gene duplication. Unfortunately, the number of parameters is greater than the degree of freedom which makes an accurate parameter estimation impossible. Thus, the different random variables were summarised by their sum which is possible according to the additive nature of the model. Thereupon, a ML approach was used for parameter estimation.

Khaitovich *et al.* (2005b) used a compound Poisson process to describe the evolution of gene expression: The Poisson process is used to model the number of mutations in the regulatory region of a gene which occur in a time period. For each mutation a random variable following a so-called *mutation effect distribution* (MED) is used to describe the mutation effect on the level of expression of that gene. Since the chosen MED can be arbitrary, the model is more flexible than a Brownian motion model. The work by Khaitovich *et al.* (2005b) represents the fundamentals of this thesis. Thus, it is described in detail later.

2.4. Optimisation

2.4.1. Overview

Optimisation aims to compute the optimal parameters of a complex system or model. "Optimal" means here the minimum or maximum of an objective function. Optimisation is used, when an analytic solution cannot be found. One of the simplest optimisation problems is the search for the root of a one-dimensional function.

2.4.2. Bracketing

A commonly used optimisation method is the *bracketing method* (Brent, 1972). It finds the root of a one-dimensional continuous and monotonic function f. The algorithm starts with the search for an interval with the limits x_1 and x_2 which brackets the root. This is performed by iteratively increasing the width of a randomly chosen initialisation interval. The searched interval is detected if the signs of $f(x_1)$ and $f(x_2)$ are different. In this case, the root is, due to the intermediate value theorem, in-between x_1 and x_2.

Subsequently, the first iteration of the optimisation starts. A split point x with $x_1 < x < x_2$ is chosen (e.g., randomly or by selecting the mean value of x_1 and x_2). If the signs of $f(x_1)$ and $f(x)$ are identical, the root is in-between the interval $[x, x_2]$. Otherwise the root is in-between the interval $[x_1, x]$. The new interval is used for the next iteration step recursively. The method stops if the range of the bracketed interval falls below a predefined value. After this, the mean value of the interval limits is presented as the estimate for the root of f.

2.4.3. The Brent's method

Another optimisation method is *Brent's method* (also referred to as *Golden Section Search*) (Brent, 1972). The method finds a minimum or maximum of a one-dimensional function f without using derivatives. It is assumed that the algorithm searches for a minimum (when searching for a maximum the $\leq$-signs have to be replaced by $\geq$-signs below). Before starting, it is essential to bracket a minimum in an interval which is then downsized iteratively. A minimum is known to be bracketed for three points x, y, and z

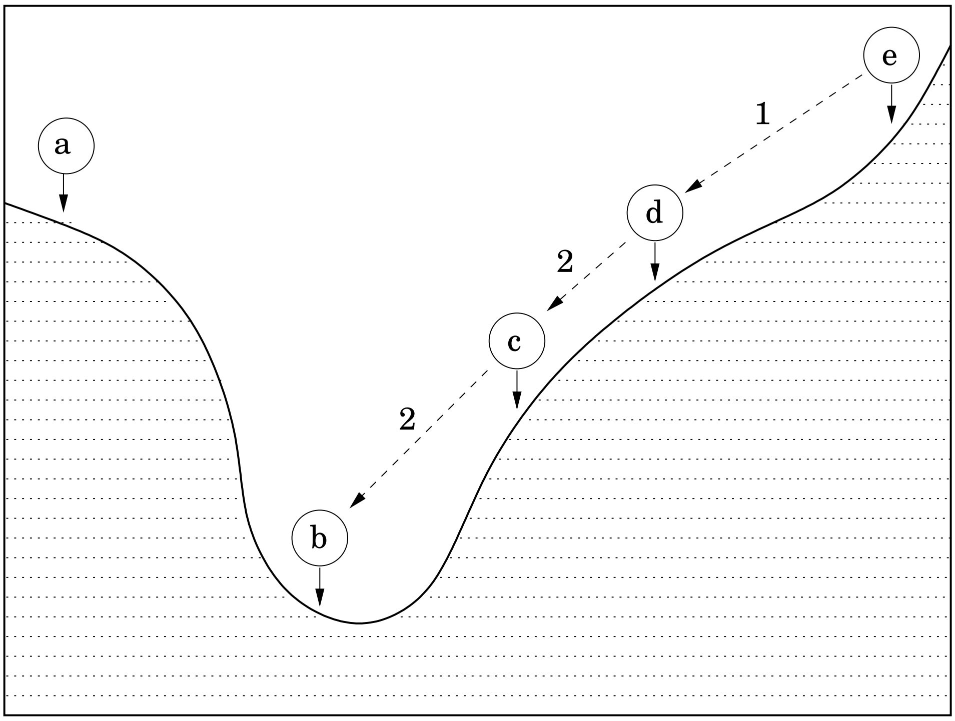


Figure 2.6.: Search for the minimum of a one-dimensional function. In the beginning the minimum is bracketed by a, c, and e. In step 1 the function is evaluated at d which replaces e. In step 2 the function is evaluated at b which is a better minimum. Thus, d is replaced by c as the new upper limit of the interval which brackets the minimum.

with $x < y < z$ so that $f(y) \leq min\{f(x), f(z)\}$. Thus, the best minimum found so far is $f(y)$.

The algorithm starts in one of the two intervals (x, y) or (y, z). The interval can be chosen either by random or by size. In the latter case the larger interval is chosen. It is assumed that the algorithm selects the interval (x, y). Then a new point v is chosen which separates (x, y) into two intervals by the ratio of golden section so that $v := y - g(y - x)$, whereas $g = 0.38197$ is the golden section constant (if the interval (y, z) has been chosen, it is separated by $v := y + g(z - y)$). It was shown that the choice of the golden section constant provides the fastest convergence (Press *et al.*, 1992).

If $f(v) \leq f(y)$, a better minimum has been found and the procedure is started again with the points x, v, and y. Otherwise the points v, y, and z are used, since y is still minimal. Thus, the range containing the minimum is downsized. The method is iterated until a sufficiently small interval has been found. Then the point with the smallest function

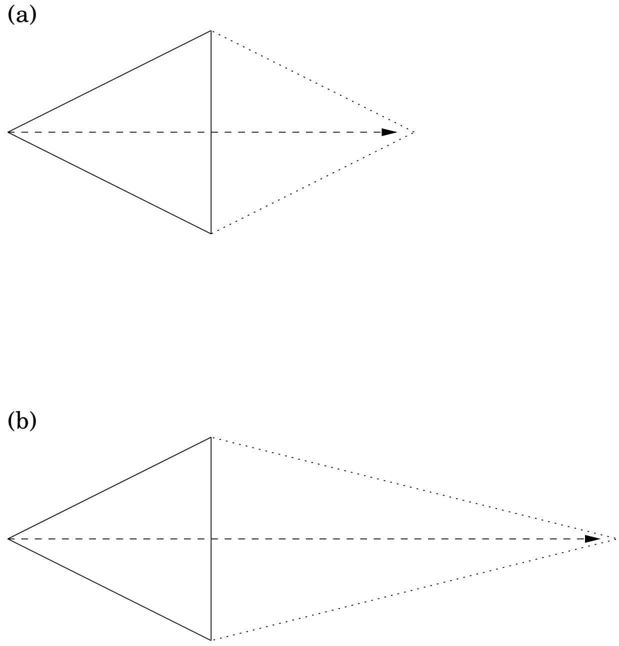


Figure 2.7.: Downhill Simplex Method in two dimensions. All four types of steps are demonstrated which are (a) reflection, (b) reflection and expansion, (c) contraction, and (d) multiple contraction. The arrows symbolise the change of vector-positions during a step and the spotted lines show the resulting new shape.

value is presented as the minimum (cf. figure 2.6 for an example of the Brent's method).

For minimising functions of higher dimension a generalised Brent's method is available. It minimises the first dimension using the one-dimensional Brent's method, then the second and so forth. Subsequently, a new iteration starts with the first dimension again. This continues until the so far best minimum is not changed more than a predefined value within one iteration over all dimensions. A problem of the Brent's method is that it is possible to get stuck in local minima. However, this is a problem of all optimisation methods.

2.4.4. The Downhill Simplex Method

The *Downhill Simplex Method* by Nelder and Mead (1965) is a multi-dimensional optimisation method which can be used to find the optimum of a non-linear function f of more than one independent variable. Here it is assumed that the method searches for

Algorithm 2.1: Downhill simplex method

```
Data: Multi-dimensional function $f$
Result: Optimum of $f$
Initialisation;
while stop criterion is false do
    Calculate the function values and the best, worst and second-worst vector;
    Try a reflection from the worst vector;
    if the reflection leads to a vector better than the best vector then
        try an additional expansion;
    if the reflection leads to a vector worse than the second-worst vector then
        try a contraction from the worst vector;
        if the contraction leads to a vector worse than the worst vector then
            make a multiple contraction;
```

a minimum. The method merely requires function evaluation, but no derivatives. However, it needs more computational time than methods which can make use of derivatives (e.g., the Newton-Raphson method (Whittaker and Robinson, 1967)). The method uses a geometric structure, the *simplex*, which manoeuvres through the search space towards the optimum. The shape of the simplex depends on the dimension of the search space. For a search space with k dimensions the simplex consists of $k+1$ vectors with k dimensions each and all edges which connect the vectors. Thus, a simplex in a 2-dimensional space is a triangle, a simplex in a 3-dimensional space is a tetrahedron *et cetera*. Please note that the simplexes used by the method have to be non-degenerated so that for each vector of the simplex, all other vectors span the vector space of k dimensions.

In the beginning an initial simplex from a vector P_0 is chosen by adding unit vectors e_i, scaled by constants λ_i with $1 \leq i \leq k$, to define the remaining k vectors P_i. The constants λ_i can be chosen randomly by using a distribution depending on the length scale of the problem. Then the function value for each vector is calculated. After that a series of steps is performed. The steps are of four different types: (a) reflection, (b) reflection and expansion, (c) contraction, and (d) multiple contraction. Figure 2.7 illustrates these four cases for a search in a 2-dimensional space. In a reflection the vector with the largest function value which is the worst vector in order to find the minimum is moved through the opposite face of the simplex. If the new vector has a smaller function value, an expansion is tried by expanding the distance orthographically to the axis of reflection in order to find a better vector and to make larger steps towards the optimum.

If the function values describe a valley in a region of the search space which is reached by the simplex, a contraction of the simplex is performed to seep through the valley. The simplex can also contract around its best vector to pass through “bottlenecks” within a region surrounded by larger function values. The method terminates if the distance bridged in a reflection step and the decrease of the function value are smaller than predefined constants. An informal description of the method is algorithm 2.1. Like other optimisation methods the downhill simplex method may get stuck in a local optimum. To avoid this, the method should be restarted many times with different initial vectors. After accomplishing the procedure the best vector is returned as the result.

3. A model with gamma-distributed mutation effects

In this chapter a gene expression evolution model is presented which uses a gamma distribution to describe the mutation effects on the level of gene expression. The gamma distribution is more complex and flexible than the distributions used in the basic gene expression evolution model.

3.1. Introduction

Kimura's neutral theory constitutes that the majority of genetic changes on DNA level are selectively neutral so that the future of most mutations whether to establish in a population or to disappear is less the result of selection, but the result of random genetic drift (Kimura, 1983). This theory alludes to the genome, but many studies addressing gene expression evolution make it reasonable to apply it to the transcriptome which describes the set of all mRNA molecules in a cell. It was characterised many times that gene expression between species differ the more the more time has passed, since the taxa split from a common ancestor (cf. reviews by Ranz and Machado (2006) and Khaitovich *et al.* (2006)). Moreover, Khaitovich *et al.* (2004) found a positive linear correlation between evolutionary time between taxa and the divergence of gene expression strength which can be explained by the neutral theory.

Thereupon, Khaitovich *et al.* (2005b) suggested a neutral model in which a changing of the expression level of a gene is induced by a mutation in the regulatory sequence (on DNA level) of that gene. Here, this model is referred to as the *M model* (M means mutation). Whenever a mutation happens, the expression level changes according to the so-called *mutation effect distribution* (MED) with expectation zero. Two types of MEDs were used to test their applicability to describe expression changes: (1) a normal distribution and (2) a positively skewed extreme value distribution. Both distributions are governed by only one parameter. Because mutations are modelled by Poisson events that occur with rate μ per time unit t the expression level of a gene at a given time point is distributed according to a compound Poisson process. Using this model the authors estimated the parameters of the model from comparative expression array studies from primate liver and brain. Khaitovich *et al.* (2005b) asserted that the extreme value distribution is superior to describe the gene expression changes in all examined data sets. Due to the asymmetry of the extreme value distribution, it was suggested that upward changes in expression during evolution are less frequent, but of greater average magnitude than downward changes.

In this chapter the *gamma distribution* is suggested as the MED. The gamma distribution is determined by two parameters for shape and scale and therefore a better fit of the predicted data to the measured data seems plausible. Unfortunately, the gamma distributed MED makes an analytical solution to estimate the model parameters impossible. Thus, an optimisation method is applied. In the following the theory is explained in detail. Subsequently, applicability of the optimisation method based on synthetic data is displayed and finally a biological example is discussed.

3.2. Materials and methods

3.2.1. The M-gamma model

A sample 1 is assumed. Consider the fate of the expression level of a single gene from sample 1: The expression level is influenced by mutations in the regulatory region of the gene. Following standard assumptions, the number of mutations $M(d_1)$ follows a Poisson process with time unit t_1 scaled by the rate μ. Thus, $d_1 = \mu t_1$ which denotes the expected number of mutations. Conditioned on a mutation, the level of expression of a gene changes according to a MED X. Here a gamma distribution with density

$$g_{\alpha,\beta}(x) = \frac{(x+\alpha\beta)^{\alpha-1}e^{-\frac{x+\alpha\beta}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, \alpha > 0, \beta > 0, x > -\alpha\beta \tag{3.1}$$

is used as MED. With this setting the mean is zero, the variance is $\alpha\beta^2$, and skewness is $2/\sqrt{\alpha}$. If a negatively skewed distribution is required, the mirrored version of the gamma distribution is used. The Poisson process and the MED are combined to the compound Poisson process

$$Y_1(d_1) = Y_1(0) + \sum_{i=1}^{M(d_1)} X_i \tag{3.2}$$

which describes the logarithm of the expression level for a gene after $M(d_1)$ mutations. Thereby, $X_1, \ldots, X_{M(d_1)}$ follow to the same gamma distribution as X. $Y(0)$ describes the logarithm of expression level at time zero. Depending on the used MED, the model is referred to as *M-gamma model.* Assume another sample 2, whereas the samples 1 and 2 descended from a common ancestor independently with d_1 and d_2 expected mutations, respectively. Let $Z_{1,2}$ be the random variable describing the difference in expression of the gene between the samples 1 and 2. Thus, $Z_{1,2}$ is the difference of two independent compound Poisson processes $Y_1(d_1)$ and $Y_2(d_2)$, described by

$$Z_{1,2} = Y_1(d_1) - Y_2(d_2) = \sum_{i=1}^{M(d_1)} X_i - \sum_{j=1}^{M(d_2)} X_j. \tag{3.3}$$

The moments of the distribution $Z_{1,2}$ can be derived using characteristic functions (cf. chapter 2.3.1 or Feller (1957) for details):

$$\text{Variance } v_{1,2}^{(gamma)} = \mu_2(Z_{1,2}) = \alpha\beta^2(d_1 + d_2) \tag{3.4}$$

$$\text{Coefficient of skewness } s_{1,2}^{(gamma)} = \gamma_1(Z_{1,2}) = \frac{2(d_1 - d_2)}{\sqrt{\alpha}(d_1 + d_2)^{3/2}} \tag{3.5}$$

$$\text{Coefficient of kurtosis } k_{1,2}^{(gamma)} = \gamma_2(Z_{1,2}) = 3 + \frac{3\alpha + 6}{\alpha(d_1 + d_2)} \tag{3.6}$$

To avoid notational difficulties $v_{1,2}$, $s_{1,2}$, and $k_{1,2}$ are used in this chapter to refer to the moments of $Z_{1,2}$.

3.2.2. Parameter estimation

Applying the model to data: If the M-gamma model is applied to real data taken from microarray experiments, it is assumed that all genes on the array are independent of each

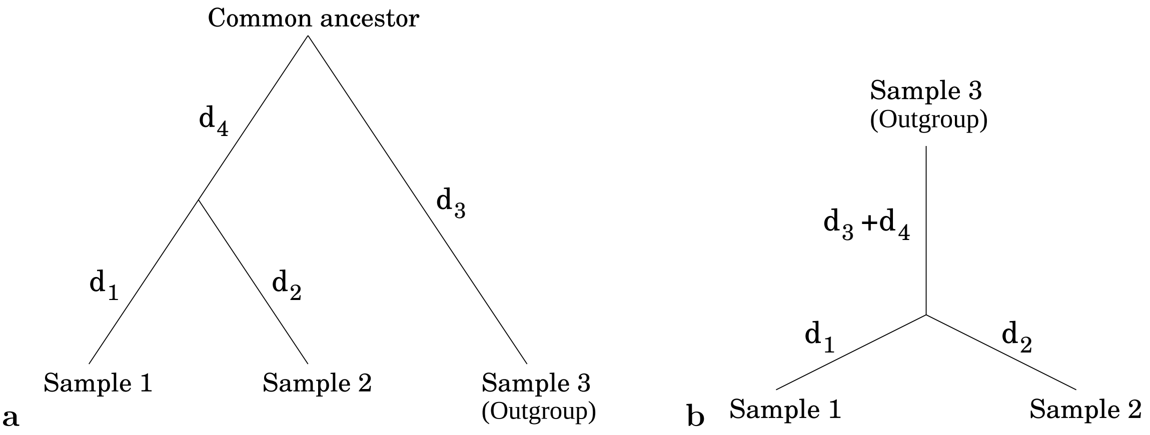


Figure 3.1.: Trees used by the model. The rooted tree (a) distinguishes between the branch to the ancestor of sample 1/2 with parameter d_4 and to the outgroup with parameter d_3. The unrooted tree (b) combines the branches with the parameters d_3 and d_4 to one branch with parameter $d_3 + d_4$.

other and followed the same evolutionary process described by equations 3.2. Therefore, gene expression changes caused by *trans*-effects are neglected. In order to find out those values for the parameters α, β, d_1, and d_2 which describe the data best, the moments from the distribution of gene expression differences of the real data are estimated. These ones, referred to as $\hat{v}_{1,2}$, $\hat{s}_{1,2}$, and $\hat{k}_{1,2}$, are equated with the equations 3.4–3.6. Unfortunately, it is not possible to yield a unique solution by transformation, since the model contains four parameters α, β, d_1, and d_2, but only three equations for the moments $v_{1,2}$, $s_{1,2}$, and $k_{1,2}$. However, one could use a higher moment to obtain a fourth equation. But for practical reasons this is not advisable, since single outliers in the data would lead to large variation in the estimates. This is due to large exponents in equations of high moments.

Use of an outgroup: Another way to derive additional equations is to add a third sample which acts as an outgroup. This can be represented by the tree illustrated in figure 3.1. In comparison to the use of two samples, the use of an outgroup leads to two additional branches representing compound Poisson processes with parameters d_3 and d_4, respectively. Hence, the model has six parameters. The procedure provides two additional random variables $Z_{1,3}$ and $Z_{2,3}$ describing the differences between sample 1 or sample 2, respectively, and the outgroup sample 3. Thus, the following system is obtained:

$$\text{Variance } v_{1,3}^{(gamma)} = \mu_2(Z_{1,3}) = \alpha\beta^2(d_1 + d_3 + d_4) \tag{3.7}$$

$$\text{Coefficient of skewness } s_{1,3}^{(gamma)} = \gamma_1(Z_{1,3}) = \frac{2(d_1 - d_3 + d_4)}{\sqrt{\alpha}(d_1 + d_3 + d_4)^{3/2}} \tag{3.8}$$

$$\text{Coefficient of kurtosis } k_{1,3}^{(gamma)} = \gamma_2(Z_{1,3}) = 3 + \frac{3\alpha + 6}{\alpha(d_1 + d_3 + d_4)} \tag{3.9}$$

$$\text{Variance } v_{2,3}^{(gamma)} = \mu_2(Z_{2,3}) = \alpha\beta^2(d_2 + d_3 + d_4) \tag{3.10}$$

$$\text{Coefficient of skewness } s_{2,3}^{(gamma)} = \gamma_1(Z_{2,3}) = \frac{2(d_2 - d_3 + d_4)}{\sqrt{\alpha}(d_2 + d_3 + d_4)^{3/2}} \tag{3.11}$$

$$\text{Coefficient of kurtosis } k_{2,3}^{(gamma)} = \gamma_2(Z_{2,3}) = 3 + \frac{3\alpha + 6}{\alpha(d_2 + d_3 + d_4)} \tag{3.12}$$

For clarity, $v_{1,3}$, $s_{1,3}$, $k_{1,3}$, $v_{2,3}$, $s_{2,3}$, and $k_{2,3}$ are used to refer to the moments in this chapter. Estimates from real data are referred to as $\hat{v}_{1,3}$, $\hat{s}_{1,3}$, $\hat{k}_{1,3}$, $\hat{v}_{2,3}$, $\hat{s}_{2,3}$, and $\hat{k}_{2,3}$. Unfortunately, the parameter α appears as a pure scaling parameter of the parameters d_1, d_2, d_3, and d_4. Thus, it is not possible to determine α, and therefore, also the collective of nine equations cannot be used to yield an analytic solution. However, it is possible to derive partial solutions. The variances $v_{1,2}$, $v_{1,3}$, and $v_{2,3}$ are proportional. Thus, d_1, d_2, and $d_3 + d_4$ can be estimated directly from the three variances except for the factor $\alpha\beta^2/2$ that is

$$d_1 = \frac{v_{1,2} + v_{1,3} - v_{2,3}}{2\alpha\beta^2}, \tag{3.13}$$

$$d_2 = \frac{v_{1,2} - v_{1,3} + v_{2,3}}{2\alpha\beta^2}, \tag{3.14}$$

$$d_3 + d_4 = \frac{-v_{1,2} + v_{1,3} + v_{2,3}}{2\alpha\beta^2}. \tag{3.15}$$

In other words the unrooted tree in figure 3.1 b reflects the evolutionary distance between the three samples, but depending on $\alpha\beta^2/2$ the tree either shrinks or grows in total length.

To determine this scaling factor, the coefficient of kurtosis is used to scale the branch lengths. Pertaining to the data, $k_{1,2}$ is applied, since it is assumed that this is the most robust estimate of the coefficient of kurtosis. This trick yields a solution depending on the parameter α. Now, it is obtained that

$$\hat{\beta} = \sqrt{\frac{\hat{v}_{1,2}(\hat{k}_{1,2} - 3)}{3\alpha + 6}}, \tag{3.16}$$

$$\hat{d_1} = \frac{(1.5\alpha + 3)(\hat{v}_{1,2} + \hat{v}_{1,3} - \hat{v}_{2,3})}{\alpha\, \hat{v}_{1,2}(\hat{k}_{1,2} - 3)}, \tag{3.17}$$

$$\hat{d_2} = \frac{(1.5\alpha + 3)(\hat{v}_{1,2} - \hat{v}_{1,3} + \hat{v}_{2,3})}{\alpha\, \hat{v}_{1,2}(\hat{k}_{1,2} - 3)}, \tag{3.18}$$

$$(d_3 \hat{+} d_4) = \frac{(1.5\alpha + 3)(-\hat{v}_{1,2} + \hat{v}_{1,3} + \hat{v}_{2,3})}{\alpha\, \hat{v}_{1,2}(\hat{k}_{1,2} - 3)}. \tag{3.19}$$

Thus, an estimate for α leads to a unique solution for the remaining parameters. It is possible to split the branch with parameter $d_3 + d_4$ to estimate d_3 and d_4 unique (illustrated by the tree in 3.1 a). In this case, one coefficient of skewness from a comparison with the outgroup is essential, since it is a measure for the ratio of mutations between two branches (Khaitovich *et al.*, 2005b). If, for example, the coefficient of skewness $s_{1,3}$ is used to split the path over the root into two branches with parameters d_3 and d_4, respectively, it follows

$$\hat{d_3} = \frac{c\, \hat{s}_{1,3}\sqrt{(2+\alpha)^3 \hat{v}_{1,3}^3}}{\alpha\, \hat{v}_{1,2}^{3/2}(\hat{k}_{1,2} - 3)^{3/2}} + \frac{(1.5\alpha + 3)\hat{v}_{1,3}}{\hat{v}_{1,2}(\hat{k}_{1,2} - 3)}, \tag{3.20}$$

$$\hat{d_4} = \frac{c\, \hat{s}_{1,3}\sqrt{(2+\alpha)^3 \hat{v}_{1,3}^3}}{\alpha\, \hat{v}_{1,2}^{3/2}(\hat{k}_{1,2} - 3)^{3/2}} + \frac{(-1.5 - \frac{3}{\alpha})\hat{v}_{1,2} + 1.5\hat{v}_{2,3} + \frac{3\hat{v}_{2,3}}{\alpha}}{\hat{v}_{1,2}(\hat{k}_{1,2} - 3)}, \tag{3.21}$$

whereas $c =\approx 1.29904$ (Note: c results from transformations during solving the system of equations for $v_{1,2}$, $v_{1,3}$, $v_{2,3}$, $s_{1,3}$, and $k_{1,2}$).

Sample-i-intermediate genes to estimate α: To estimate the parameter α, the relation of *sample-i-intermediate genes* is used here which was suggested in Khaitovich *et al.* (2005b). This relation subdivides the collection of genes into three classes C_1, C_2, C_3. The genes in C_i with $i = \{1, 2, 3\}$ are called sample-i-intermediate. Let $x_j^{(i)}$ be the gene expression value of gene j in sample i. A gene j is in C_i if its expression value $x_j^{(i)}$ lies in between the expression level of the other two samples:

$$C_1 = \{j | x_j^{(2)} < x_j^{(1)} < x_j^{(3)} \vee x_j^{(2)} > x_j^{(1)} > x_j^{(3)}\} \tag{3.22}$$

$$C_2 = \{j | x_j^{(1)} < x_j^{(2)} < x_j^{(3)} \vee x_j^{(1)} > x_j^{(2)} > x_j^{(3)}\}$$

$$C_3 = \{j | x_j^{(1)} < x_j^{(3)} < x_j^{(2)} \vee x_j^{(1)} > x_j^{(3)} > x_j^{(2)}\}$$

The class C_1 is enriched with genes in which changes on the branch with parameter d_2 caused the difference in expression between sample 1 and 2, while the class C_2 rather contains genes which changed their expression level on the branch with parameter d_1. It is assumed that the expression values of the majority of genes in C_1 and C_2 are closer to the expression values of the corresponding genes in the ancestor of sample 1 and sample 2 in comparison to genes in C_3. A reason for this is that sample 1 and sample 2 evolve together on the branch with parameter d_4. The distributions of differences between expression values in sample 1 and sample 2 for genes in C_1, C_2, and C_3 are denoted with Z_{S1I}, Z_{S2I}, and Z_{S3I}, respectively.

The distributions Z_{S1I} and Z_{S2I} were used by Khaitovich *et al.* (2005b) to show that positively skewed MEDs are superior than symmetric MEDs in primates for an M model. If a positively skewed MED is used, Z_{S1I} is expected to be negatively skewed and Z_{S2I} is expected to be positively skewed. If the MED is negatively skewed, it is *vice versa*. In contrast, for symmetric MEDs the coefficients of skewness of Z_{S1I} and Z_{S2I} are expected to be zero (cf. Khaitovich *et al.* (2005b) for more details). Thus, the coefficients of skewness of Z_{S1I} and Z_{S2I} are affected by the skewness of the MED. The skewness of the used gamma distributed MED is $2/\sqrt{\alpha}$ (the skewness of the mirrored variant is $-2/\sqrt{\alpha}$). Hence, both distributions Z_{S1I} and Z_{S2I} can be used to estimate α, since the skewness of the gamma distributed MED depends only on α. However, there are no equations known to describe the coefficients of skewness of Z_{S1I} and Z_{S2I}. Hence, computer simulations of the model are performed to estimate them from the simulated data. In this study the more robust second coefficient of the Pearson's skewness [3(`mean`−

`median)/standard deviation`] is applied to estimate the skewnesses of Z_{S1I} and Z_{S2I} annotated as s_{S1I} and s_{S2I}, respectively.

Simulation of the process: A simulation of the evolutionary process according to the M-gamma model takes place as follows: At first values for the parameters α, β, d_1, d_2, d_3, and d_4 are chosen. Then the expression value at the root (or the inner node in case of an unrooted tree) is initialised with 0 (other values can be used as well without consequence, since only the differences between the values in the leaves are used after simulation). Subsequently, the following procedure is performed for the whole tree according to a depth first search: To simulate the process from an ancestor to a child along the branch with parameter d_i with $i \in \{1, 2, 3, 4\}$, the number of mutations occurring on that branch is drawn from a Poisson distribution with parameter d_i. The resulting $M(d_i)$ changes in gene expression are drawn from the gamma distribution with parameters α and β (cf. equation 3.1) (or its mirrored variant for negatively skewed mutation effects). These changes are added to the expression value of the ancestor to calculate the expression value in the child. Expression values in the leaves are stored as the results. The whole simulation is repeated many times to get a distribution of simulated expression values (e.g 10^7 genes evolving according to the process).

A particular case of simulation is referred to as *simulation depending on* α. Here at first a value for α is chosen. Then the parameter estimates $\hat{\beta}, \hat{d}_1, \hat{d}_2, \hat{d}_3$, and $\hat{d}_4$ (or $\widehat{(d_3 + d_4)}$ instead) are calculated using equations 3.16–3.21. Accordingly, the simulation process is started with these parameter values. As a result of computer simulation depending on α, the second coefficients of Pearson's skewness s_{S1I} and s_{S2I} referred to as $s_{S1I}(\alpha)$ and $s_{S2I}(\alpha)$, respectively, are estimated.

Optimisation of α: If one wants to estimate parameters for real data, the goal is to find an estimate $\hat{\alpha}$ so that $s_{S1I}(\hat{\alpha}) = \hat{s}_{S1I}$ and $s_{S2I}(\hat{\alpha}) = \hat{s}_{S2I}$. Thereby, $\hat{s}_{S1I}$ and $\hat{s}_{S2I}$ are estimated from the real data. The estimated $\hat{\alpha}$ together with the parameter estimates from equations 3.16–3.21 are a good description of the evolutionary process leading to the observation. Thus, an optimisation strategy is used. Let α' be an estimate for α. The distance of α' to the optimal solution is defined by the following objective functions:

$$\delta_1(\alpha') := s_{S1I}(\alpha') - \hat{s}_{S1I} \tag{3.23}$$

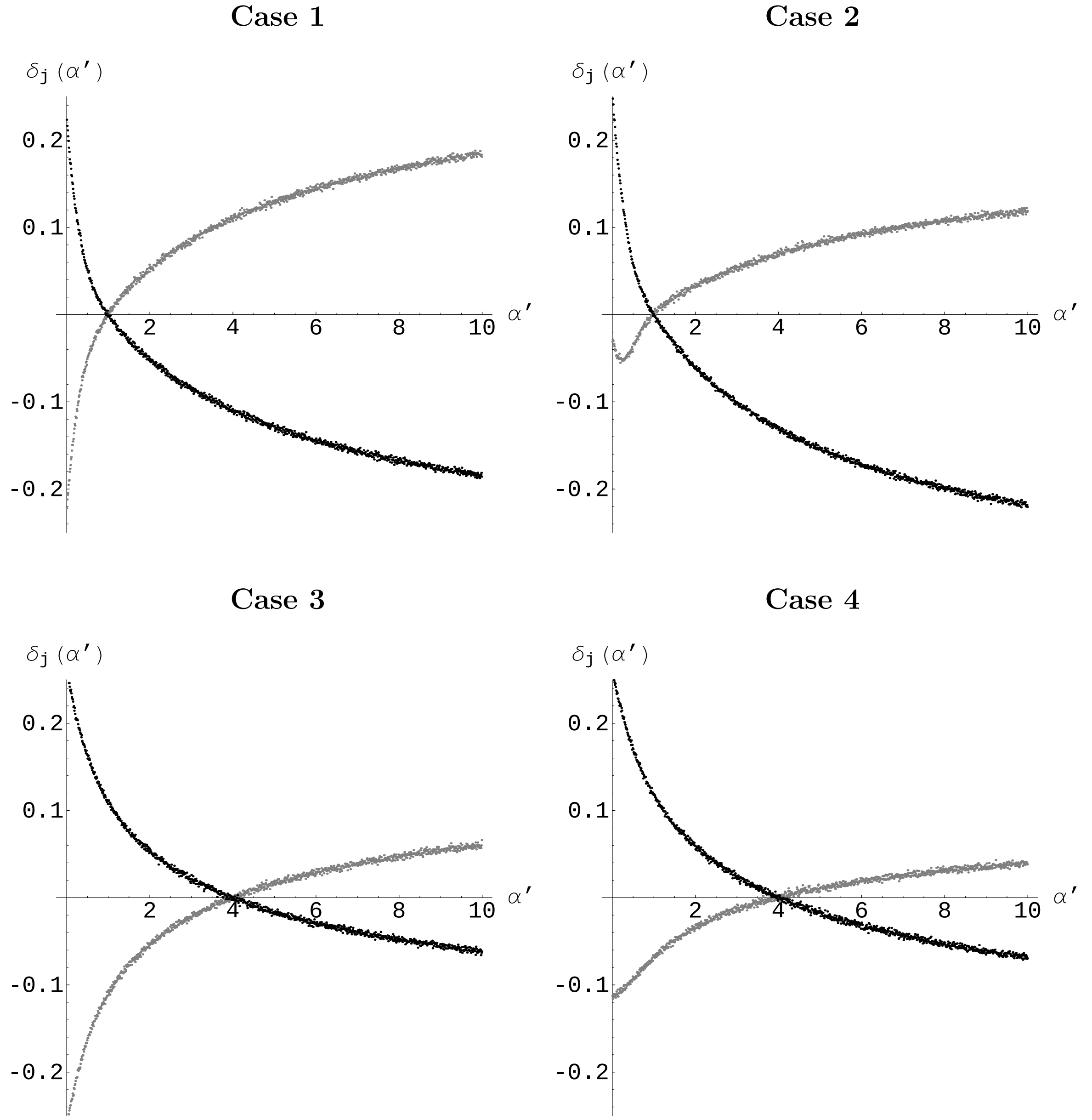


Figure 3.2.: α' plotted against $\delta_1(\alpha')$ (grey) and $\delta_2(\alpha')$ (black) in 0.01-steps. The plot shows the cases 1 to 4 described in table 3.1. Each point is based on 10^6 simulations.

$$\delta_2(\alpha') := s_{S2I}(\alpha') - \hat{s}_{S2I} \tag{3.24}$$

Thus, the nulls of equations 3.23 and 3.24 have to be found. Figure 3.2 shows $\delta_1(\alpha')$ and $\delta_2(\alpha')$ each in the interval $[0.01, 10.00]$ for four different parameter assignments (the cases are described in table 3.1). The nulls of the two functions are determined

Table 3.1.: Test cases for the parameter estimation method.

Case	α	β	d_1	d_2	d_3	d_4
1	1.0	1.0	1.0	1.0	2.0	1.0
2	1.0	1.0	2.0	1.0	2.0	1.0
3	4.0	0.5	1.0	1.0	2.0	1.0
4	4.0	0.5	2.0	1.0	2.0	1.0

independently with the bracketing method (cf. chapter 2.4.2 or (Brent, 1972)). Before this method is started, it is checked whether to choose the gamma distribution or its mirrored variant as the MED: If $\hat{s}_{S1I} < 0$ and $\hat{s}_{S2I} > 0$ the gamma distribution is used. If $\hat{s}_{S1I} > 0$ and $\hat{s}_{S2I} < 0$ the mirrored gamma distribution is used. For each step of the bracketing method a series of 10^7 simulations is performed to generate $s_{S1I}(\alpha')$ and $s_{S2I}(\alpha')$, respectively. Optimisation stops if the width of the bracketed interval falls below 0.001. The best α-estimates from the optimisation of $\delta_1(\alpha')$ and $\delta_2(\alpha')$ are called $\hat{\alpha}_{\delta_1}$ and $\hat{\alpha}_{\delta_2}$, respectively. Please note that $\hat{\alpha}_{\delta_1} = \hat{\alpha}_{\delta_2}$ for idealised data. To address noise in the data, the final estimate $\hat{\alpha}$ is calculated by the mean value of $\hat{\alpha}_{\delta_1}$ and $\hat{\alpha}_{\delta_2}$ weighted by the sizes of the sample-i-intermediate subsets:

$$\hat{\alpha} = \frac{\hat{\alpha}_{\delta_1} \cdot |C_1|}{|C_1| + |C_2|} + \frac{\hat{\alpha}_{\delta_2} \cdot |C_2|}{|C_1| + |C_2|} \tag{3.25}$$

The weighting is used, since the size of the subsets depends on the ratio of d_1 and d_2. In extreme cases, subsets might get very small and therefore become very sensitive against outliers.

3.3. Experiments and results

3.3.1. Evaluation of the parameter estimation method

In order to validate the parameter estimation method, synthetic data sets were generated by simulation of the process corresponding to the tree in figure 3.1a (cf. chapter 3.2.2). For each gene in a synthetic data set one simulation was performed to generate the gene expression values in the three samples. This was repeated numerous times to simulate

Table 3.2.: Parameters estimates for synthetic data sets. The table shows the mean values and 95 % confidence intervals of the estimates of 1, 000 data sets generated with the parameters in table 3.1.

Case	#Genes	$\hat{\alpha}$	$\hat{\beta}$	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_4$
1	10^4	1.273	0.968	1.030	1.031	1.957	1.142
		(0.570 , 2.435)	(0.776 , 1.135)	(0.499 , 1.913)	(0.506 , 1.938)	(0.906 , 3.680)	(0.534 , 2.177)
1	10^5	1.057	0.992	0.986	0.986	1.940	1.019
		(0.834 , 1.359)	(0.935 , 1.044)	(0.774 , 1.252)	(0.771 , 1.252)	(1.505 , 2.488)	(0.792 , 1.288)
2	10^4	1.177	0.990	2.220	1.109	2.239	1.099
		(0.388 , 2.613)	(0.860 , 1.125)	(1.107 , 3.876)	(0.551 , 1.917)	(1.043 , 4.152)	(0.564 , 2.007)
2	10^5	1.020	0.999	2.026	1.013	2.030	1.011
		(0.771 , 1.372)	(0.954 , 1.042)	(1.640 , 2.470)	(0.823 , 1.233)	(1.620 , 2.523)	(0.817 , 1.243)
3	10^4	5.280	0.475	1.011	1.011	1.876	1.163
		(2.257 , 11.302)	(0.308 , 0.637)	(0.697 , 1.558)	(0.708 , 1.571)	(1.164 , 2.950)	(0.764 , 1.836)
3	10^5	4.171	0.494	0.996	0.966	1.947	1.041
		(3.341 , 5.133)	(0.442 , 0.542)	(0.881 , 1.129)	(0.883 , 1.136)	(1.694 , 2.234)	(0.905 , 1.187)
4	10^4	5.300	0.501	2.087	1.041	2.112	1.020
		(2.003 , 16.903)	(0.377 , 0.621)	(1.439 , 3.059)	(0.708 , 1.535)	(1.252 , 3.378)	(0.551 , 1.634)
4	10^5	4.110	0.499	2.006	1.003	2.004	1.005
		(3.154 , 5.385)	(0.460 , 0.536)	(1.782 , 2.273)	(0.891 , 1.135)	(1.700 , 2.325)	(0.857 , 1.168)

the large number of genes on a microarray. Afterwards, the estimation method was applied to the synthetic data sets and the estimates were compared to the parameter values used for generation.

Four different parameter assignments were selected to generate synthetic data sets. They are shown in table 3.1. Case 1 and 2 represent a situation in which the gamma distribution assumes the shape of an exponential distribution. In case 3 and 4 the MED is rather similar to a normal distribution, but nevertheless with a skewness of $2/\sqrt{\alpha} = 1$. Please note that there are no test cases with a negatively skewed MED, since the results are analogous (not shown). Case 1 and 3 represent clock-like trees with an equal number of expected mutations on each branch from the root to the leaves. Case 2 and 4 show accelerated evolution on the branch to sample 1. However, the parameters were chosen so that the variance $\alpha\beta^2$ of the gamma distributed MED is 1 in each case. Thus, the variance between two samples is assumed to be the same for both MEDs if the branches connecting them have equal parameter values. For each of the test cases series of $1,000$ data sets each with (1) $10,000$ and (2) $100,000$ genes were generated.

The parameters were estimated for each of the sets with the optimisation method (cf. chapter 3.2.2). Mean estimates and 95 % confidence limits of the $1,000$ estimations were

Table 3.3.: Comparison of the primate data sets. The table shows the sample sizes of the different species and the number of genes.

Data set	#Human	#Chimpanzee	#Orangutan	#Genes
Liver 95	6	6	2	1,971
Brain 95	6	3	1	1,998
Liver 133	6	5	5	8,036
Brain 133	6	5	1	10,444

calculated for each case. The results are presented in table 3.2. In each case the mean estimates correspond approximately to the true parameters. For $10,000$ genes the deviation from the true parameters is stronger than for $100,000$ genes. The deviation of the estimates for α is stronger than for the other parameters. However, the estimators look asymptotically unbiased overall. The confidence limits are decreased with an increase of the number of genes from $10,000$ to $100,000$. The confidence limits correlate roughly linearly with the values of the parameters, for example, the limits of $\hat{d_1}$ are twice as large in the cases in which $d_1 = 2$ in comparison to the cases in which $d_1 = 1$ (for the same number of genes). Further, the shape of the MED affects the variance of d_1, d_2, d_3, and d_4 estimates, since in case 1 and 2 ($\alpha = 1.0, \beta = 1.0$) the confidence limits are slightly greater than in case 3 and 4 ($\alpha = 4.0, \beta = 0.5$).

3.3.2. Analysis of primate data

The M-gamma model was applied to different primate data sets containing expression profiles from liver and brain of human, chimpanzee, and orangutan (Khaitovich *et al.*, 2005b). The common ancestor of human and chimpanzee lived about 6 million years ago and the common ancestor of all three species lived about 13 million years ago (Glazko and Nei, 2003). Thus, the orangutan species was used as the outgroup sample 3, while the human species and the chimpanzee species were regarded as sample 1 and sample 2, respectively. Table 3.3 gives an overview over all data sets which were collected with Affymetrix HG U95Av2 arrays (liver95 and brain95) and Affymetrix U133plus2 arrays (liver133,brain133). Both array types are designed for human samples. To avoid artifacts, only expression values of those genes were included in which the corresponding oligonucleotide sequences match between human and chimpanzee. However, all these expression

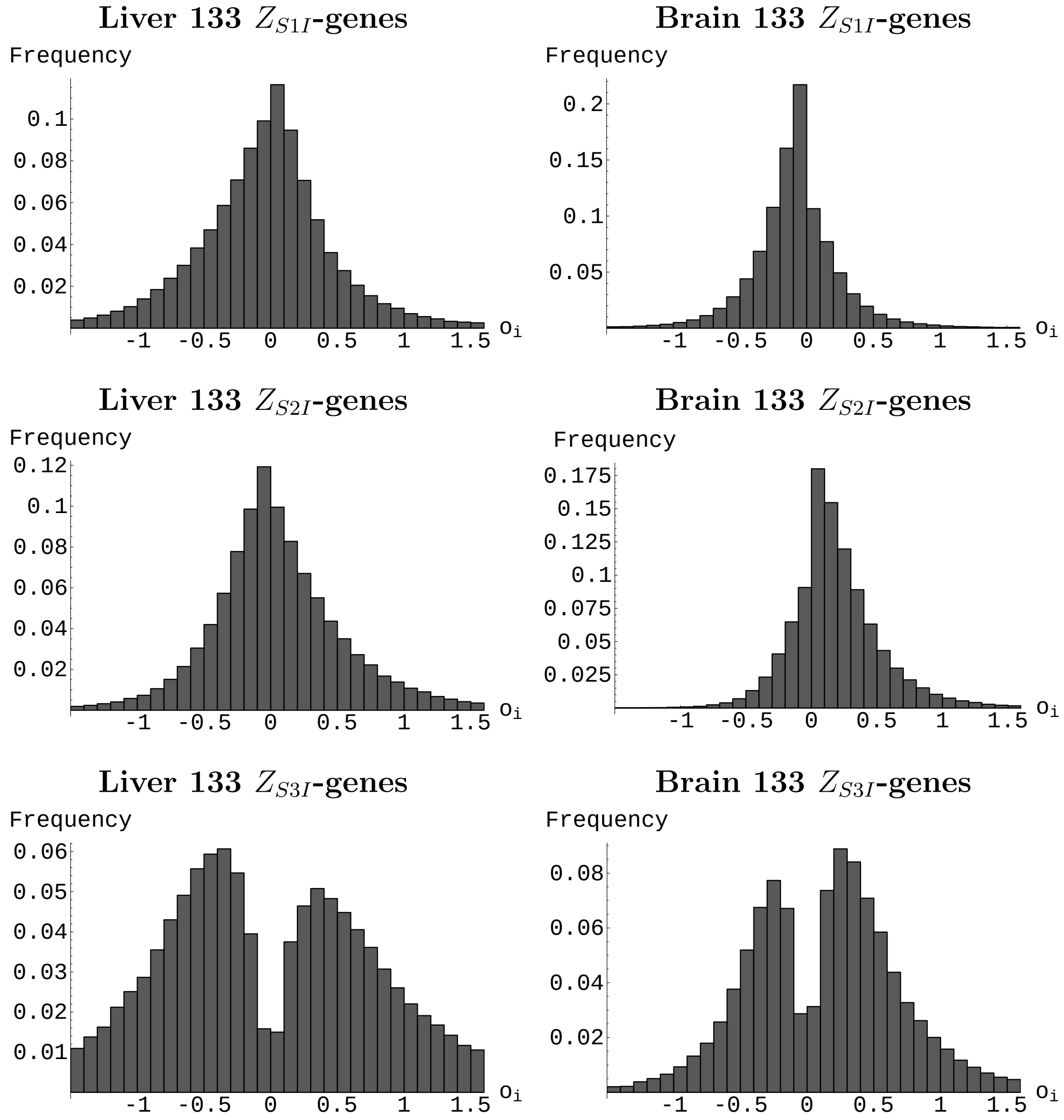


Figure 3.3.: The gene expression difference distributions Z_{S1I}, Z_{S2I}, and Z_{S3I} between human and chimpanzee.

values were also measured from orangutan without checking the match between the corresponding oligonucleotides from human and orangutan, since the orangutan genome data was not available. Thus, gene expression measured from orangutan is controversial, since measurements might be misleading as a result of weak hybridisation caused by mismatching human specific oligonucleotides.

After scanning, the raw data was normalised with the Bioconductor RMA function (Bolstad *et al.*, 2003; Irizarry *et al.*, 2003a,b). Thereby, only those probe sets were taken into account in which expression was significantly above background level in all samples

Table 3.4.: Parameter estimates from primate data. 95 % confidence limits of 1,000 bootstrap estimates are shown in brackets.

Data set	$\hat{\alpha}$	$\hat{\beta}$	$\hat{d}_1$(human)	$\hat{d}_2$(chimpanzee)	$\hat{d}_3$(orangutan)	$\hat{d}_4$
Liver 95	68.339	0.056	0.353	0.331	-3.221	3.974
932 valid	(5.256 , 1197.6)	(0.013 , 0.230)	(0.264 , 0.582)	(0.228 , 0.577)	(-15.83 , -0.406)	(1.174 , 16.54)
Brain 95	9.904	0.112	0.478	0.198	-1.909	3.007
926 valid	(1.515 , 1094.7)	(0.014 , 0.283)	(0.288 , 1.067)	(0.077 , 0.683)	(-18.531 , -0.352)	(1.494 , 20.010)
Liver 133	10.240	0.272	0.539	0.550	0.000	1.004
987 valid	(4.201 , 282.58)	(0.179 , 0.362)	(0.376 , 0.936)	(0.325 , 1.060)	(-0.100 , 0.285)	(0.589 , 1.670)
Brain 133	5.373	0.354	1.557	0.884	-0.176	4.476
991 valid	(0.458 , 523.80)	(0.296 , 0.392)	(0.826 , 2.449)	(0.399 , 1.760)	(-0.785 , 0.000)	(2.297 , 7.870)

of the corresponding data set. In the primate data sets the species are represented by more than one individual in most cases. Thus, the sample size of the three samples is greater than 1. Let n_1 be the size of sample 1, n_2 be the size of sample 2, and n_3 be the size of sample 3. Before starting the parameter estimation method, the moments $\hat{v}_{1,2}$, $\hat{v}_{1,3}$, $\hat{v}_{2,3}$, $\hat{s}_{1,3}$, $\hat{k}_{1,2}$, $\hat{s}_{S1I}$, and $\hat{s}_{S2I}$ were estimated from the data. To this end, these moments were first estimated for all pairs of individuals of the two respective species ($n_1 \cdot n_2$ comparisons between human and chimpanzee, $n_1 \cdot n_3$ comparisons between human and orangutan, and $n_2 \cdot n_3$ comparisons between chimpanzee and orangutan). Subsequently, for each moment the mean value from all pairwise comparisons was calculated and then used for the parameter estimation method. For all primate data sets a positively skewed distribution was assumed, since $s_{S1I} < 0$ and $s_{S1I} > 0$ in each set. Figure 3.3 shows exemplarily the distributions Z_{S1I}, Z_{S2I}, and Z_{S3I} for the sets "Liver 133" and "Brain 133" to illustrate the difference of the skewness. Obviously the Z_{S3I}-distributions consists mainly of larger gene expression differences (indicated by the minimum around zero), since there is a great chance for genes in C_3 that the gene expression in sample 1 and 2 drifted in different directions, while it stayed unchanged in sample 3. In addition to parameter estimation for the original data sets, the bootstrapping resampling method (Efron, 1979) was used over the genes and the individuals to construct 95 % confidence regions. For each data set 1,000 bootstrap data sets were generated and the parameter estimation method was applied to the resampled data.

Table 3.4 shows the parameter estimates and 95 % confidence intervals from bootstrapping. The results for α and β are very different in the data sets. However, within the two different types of microarrays (HG U95Av2 and U133plus2) the estimates for α are greater and for β are smaller in liver than in brain. Thus, for the same type of array the

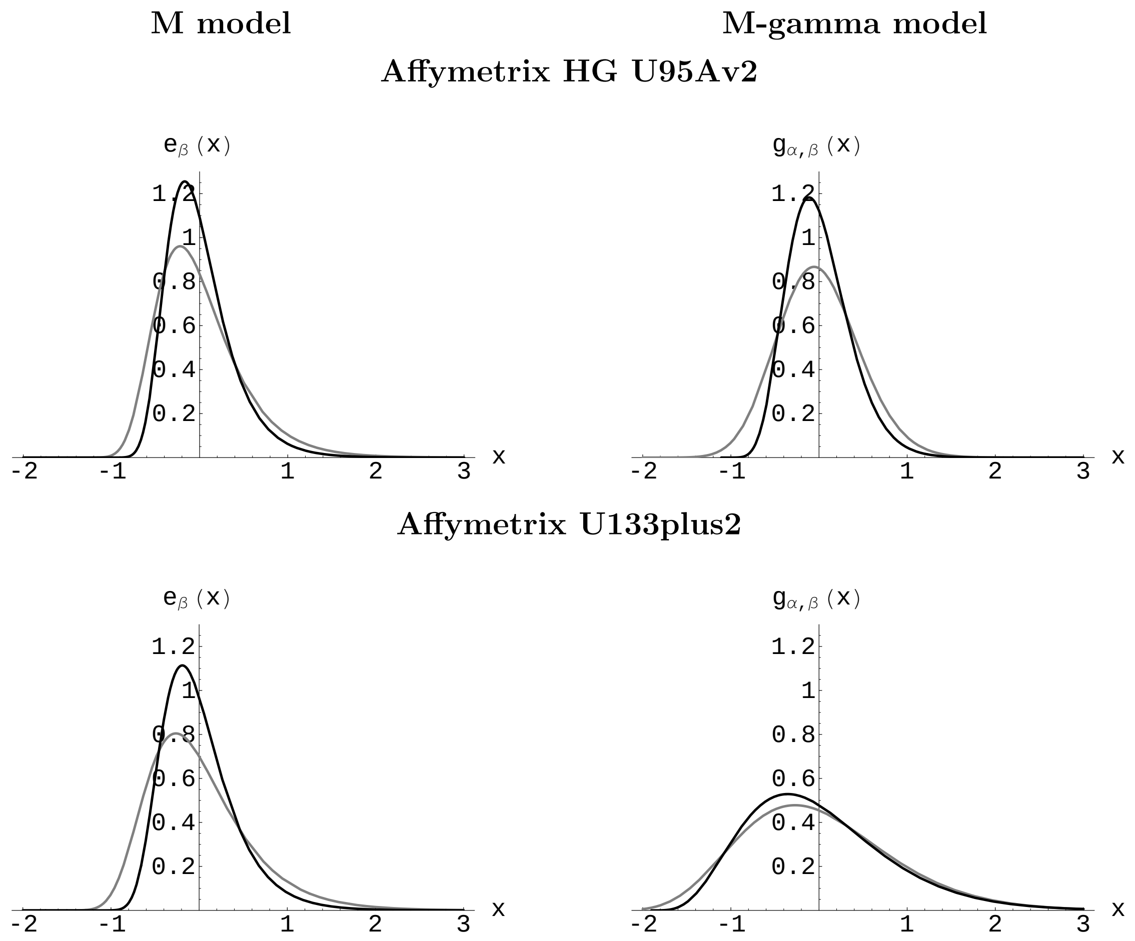


Figure 3.4.: Density of the mutation effects estimated from the liver (grey) and brain (black) data sets with the M model with extreme value distributed MED $e_\beta(x)$ (estimates for β were taken from Khaitovich *et al.* (2005b): Liver 95 ($\beta = 0.383$), Brain 95 ($\beta = 0.293$), Liver 133 ($\beta = 0.457$), Brain 133 ($\beta = 0.330$)) and with the M-gamma model.

skewness of the MED $2/\sqrt{\alpha}$ is greater in the brain sets than in the liver sets. However, the confidence limits have a wide range, especially for the parameter α. The ranges of the confidence intervals for α and β are greater in the "95"-data sets than in the "133"-sets which can be explained by the considerably smaller number of genes. To illustrate the shapes of the gamma distributed MEDs depending on the α- and β-estimates from the four data sets, figure 3.4 shows a comparison of the corresponding density functions $g_{a,b}(x)$ (equation 3.1). In order to compare these densities with previous results by Khaitovich *et al.* (2005b), figure 3.4 also shows density functions of the corresponding extreme value distributed MEDs $e_\beta(x)$ of the M model. The estimates for β were taken from Khaitovich *et al.* (2005b). The gamma distributions of the "133" data sets have a

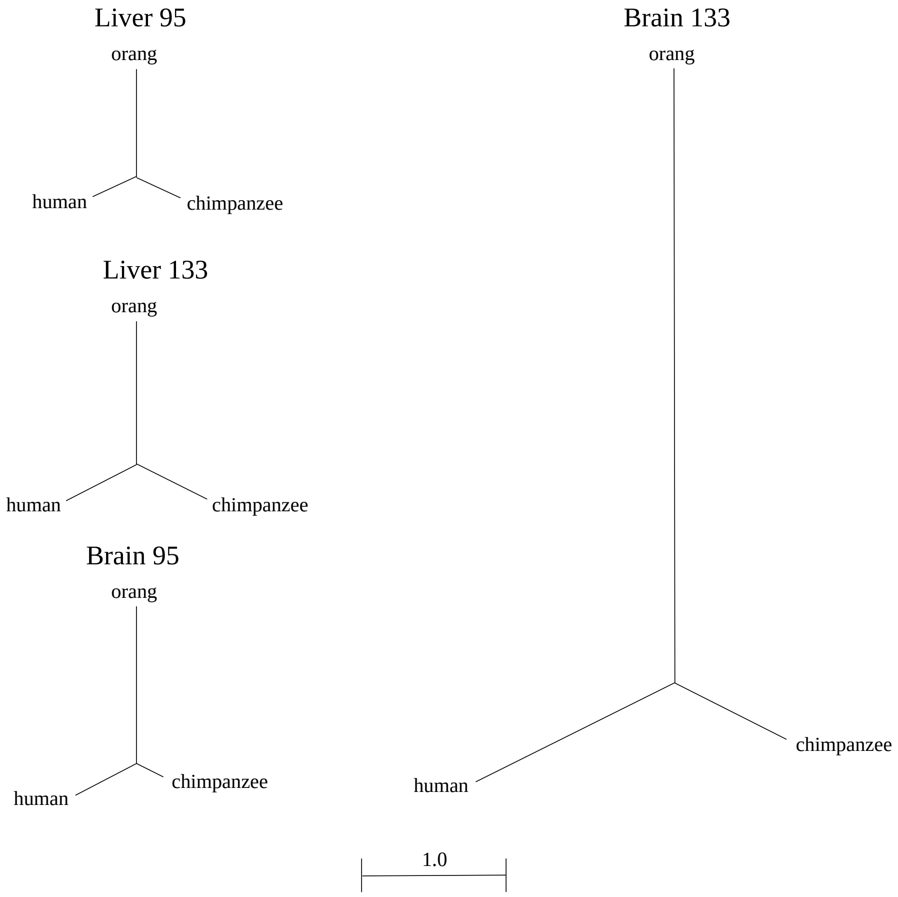


Figure 3.5.: Expected number of mutations represented by different branch lengths for the primate data sets.

greater variance $\alpha\beta^2$ (Liver 133: 0.758, Brain 133: 0.673) than the ones estimated from the considerably smaller "95" sets (Liver 95: 0.214, Brain 95: 0.124) and they also differ more greatly from the corresponding extreme value distributions (cf. figure 3.4).

When analysing the estimates for the expected number of mutations d_1, d_2, d_3, and d_4, it is noticeable that they are about three to four times greater for "Brain 133" than for the other three data sets. However, the ratios of these estimates between human and chimpanzee ($\hat{d}_1/\hat{d}_2$) are consistent in the different data sets: In the two liver sets the ratios $\hat{d}_1/\hat{d}_2$ are nearly equal (Liver 95: $0.353/0.331 = 1.066$, Liver 133: $0.539/0.550 = 0.98$). In contrast, the two brain sets show an acceleration on the human lineage (Brain 95: $0.478/0.198 = 2.414$, Brain 133: $1.557/0.884 = 1.761$). However, the confidence intervals of $\hat{d}_1$ and $\hat{d}_2$ overlap in both brain sets. While the ratio of $\hat{d}_1$ and $\hat{d}_2$ depends on the ratios between the three variances $\hat{v}_{1,2}$, $\hat{v}_{1,3}$, and $\hat{v}_{2,3}$ (cf. equations 3.13–3.15), the ratio between $\hat{d}_3$ and $\hat{d}_4$ depends on the coefficient of skewness $\hat{s}_{1,3}$ (cf. equations 3.20

Table 3.5.: Comparison of the mice data sets

Data set	#dom	#mus	#spretus	#Genes
Brain	6	6	3	19,406
Liver/Kidney	6	6	3	19,510
Testis	6	6	3	19,348

and 3.21) which is a more sensitive criterion against outliers in the empirical data than the variances. This would explain the negative $\hat{d}_3$ in all four data sets. However, if the position of the root of the tree is not considered, equation 3.19 can be used to estimate the sum $(\widehat{d_3 + d_4})$. The resulting unrooted trees which show the ratios between the three species are depicted in figure 3.5. Here the differences in size between the tree for "Brain 133" and the remaining ones is eye-catching.

3.3.3. Analysis of mice data

To present a second biological example, data sets of brain, testis and a mixture of liver and kidney tissues of mice were applied to the model. These data sets are part of the data analysed by Voolstra *et al.* (2007) using spotted arrays (OligoLibrary by Sigma-Genosys / Compugen spotted on Schott Nexterion Slides H). Table 3.5 gives an overview of the data sets. *Mus musculus domesticus* (dom) and *Mus musculus musculus* (mus) are subspecies of *Mus musculus.* They were regarded as sample 1 and 2, respectively. The species *Mus spretus* (spretus) was used here as the outgroup (regarded as sample 3). The split between *Mus musculus* and *Mus spretus* occurred about 1.1 million years ago (She *et al.*, 1990). The data sets are analysed in the same manner as the primate sets (cf. chapter 3.3.2).

The results are shown in table 3.6. Interestingly, a positively skewed MED could not fit the liver/kidney data, since $\hat{s}_{S1I} > 0$ and $\hat{s}_{S2I} < 0$. Thus, the mirrored version of the gamma distribution was used in this case (cf. chapter 3.2.2). Again, the estimates for α are very different between the data sets. The corresponding confidence intervals are extremely wide. For a small number of bootstrap data sets the optimisation algorithm could not find an estimate, since the upper limit of the search space for α' was reached which was set to 4,096. In these cases the algorithm stopped. The number of valid runs

Table 3.6.: Parameter estimates from mice data. 95 % confidence limits of 1,000 bootstrap estimates are shown in brackets. For the Liver/Kidney a mirrored version of the gamma distribution was applied. A small number of estimations on the 1,000 bootstraps per data set reached the limits of the search space of the optimisation method. Therefore, no results were obtained. These runs are invalid. The number of valid results is denoted in the first column.

Data set	$\hat{\alpha}$	$\hat{\beta}$	$\hat{d}_1$(dom)	$\hat{d}_2$(mus)	$\hat{d}_3$(spretus)	$\hat{d}_4$
Brain	13.155	0.350	0.108	0.192	0.100	0.103
953 valid	(3.534 , 2156.7)	(0.179 , 0.539)	(0.079 , 0.124)	(0.128 , 0.399)	(0.024 , 0.177)	(0.028 , 0.595)
Liver/Kidney	275.41	0.066	0.124	0.132	-0.040	0.449
833 valid*	(9.324 , 1428.4)	(0.012 , 0.227)	(0.103 , 0.187)	(0.102 , 0.195)	(-1.446 , 0.235)	(0.133 , 1.948)
Testis	3.586	0.301	0.107	0.079	0.046	0.157
965 valid	(1.698 , 309.76)	(0.035 , 0.475)	(0.078 , 0.124)	(0.051 , 0.125)	(-0.700 , 0.101)	(0.071 , 0.921)

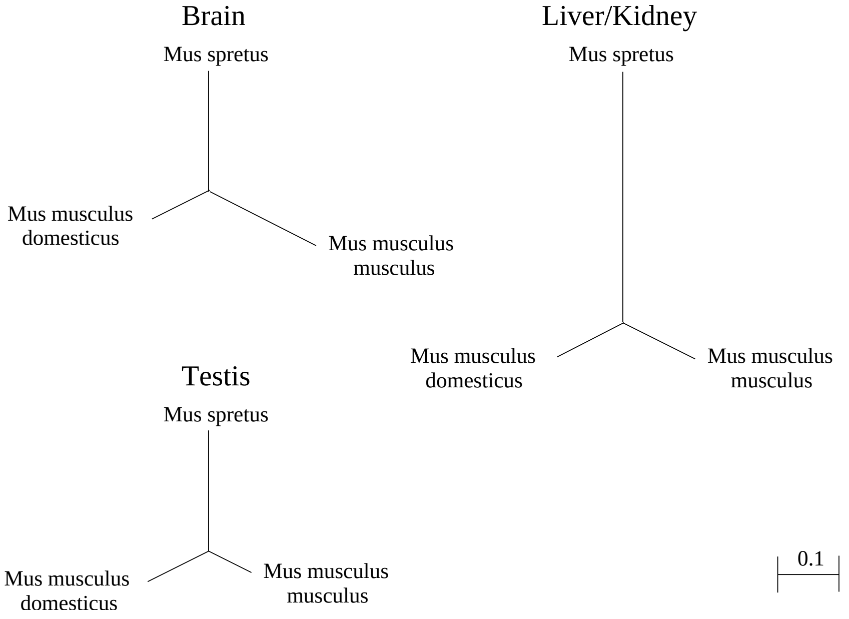


Figure 3.6.: Expected number of mutations represented by different branch lengths for the mice data sets.

in which the optimisation method terminates is denoted in the first column of table 3.6.

In comparison to the primate data sets, the d_i estimated from the mice sets are substantially smaller. In brain an acceleration on the mus-lineage (0.192) in comparison to the dom-lineage (0.108) with non-overlapping confidence intervals can be observed, while in liver/kidney the estimates $\hat{d}_1$ and $\hat{d}_2$ are very similar (0.124 and 0.132, respectively). In testis might be a slight acceleration in dom (0.107 against 0.079), but the confidence intervals overlap. Like in the primate sets, the ratios between $\hat{d}_3$ and $\hat{d}_4$ are problematic. Thus, unrooted trees were used for illustration (cf. figure 3.6).

3.4. Discussion

The presented M-gamma model provides a deeper analysis of gene expression evolution than the former model by Khaitovich *et al.* (2005b), since it uses a gamma distributed MED. Depending on its additional parameter, the gamma distribution is more flexible than the normal or the extreme value distribution. Moreover, the gamma distribution should summarise effects of different magnitude more accurately which was shown by Uzzel and Corbin (1971) for the discrete version of the gamma distribution. This might be useful when modelling evolution of gene expression, since it is reasonable to assume that the expression of genes evolve with different magnitude over all genes. However, further research is necessary to get a deeper view into these processes. By then, it is assumed that all genes evolve under the same model without any dependencies among each other which neglects *trans*-effects. A problem related to this simplification is that some genes might affect the expression levels of many other genes which would lead to imprecise estimates. However, it has been suggested that evolution of *cis*-effects and single gene affecting *trans*-effects are prevalent (Morley *et al.*, 2004).

To estimate the model parameters, a method of moments is used (Pearson, 1894). To this end, the relation of sample-i-intermediate genes is applied to obtain the distributions Z_{S1I} and Z_{S2I} whose skewnesses are necessary for the estimation. Unfortunately, closed equations for these characteristics are unknown. Thus, computer simulation is used to estimate the them from the simulated data. To adjust the best fitting parameter assignment a bracketing method is applied. Thereby, it is searched for the null of $\delta_1(\alpha')$ (equation 3.23) and $\delta_2(\alpha')$ (equation 3.24).

The method was validated with synthetic data sets which revealed large confidence intervals (cf. table 3.2). However, this depends on the number of genes. For $10,000$ genes, the typical number of genes on a microarray, variation of the estimates is greater than in the model by Khaitovich *et al.* (2005b). However, the number of model parameters is larger. While estimates for d_i with $i \in \{1, 2, 3, 4\}$ do not differ much from the former M model (Khaitovich *et al.*, 2005b), it is noticeable that α varies greatly (cf. table 3.2). A reason for the considerable variation in α is the nature of the gamma distribution. Its

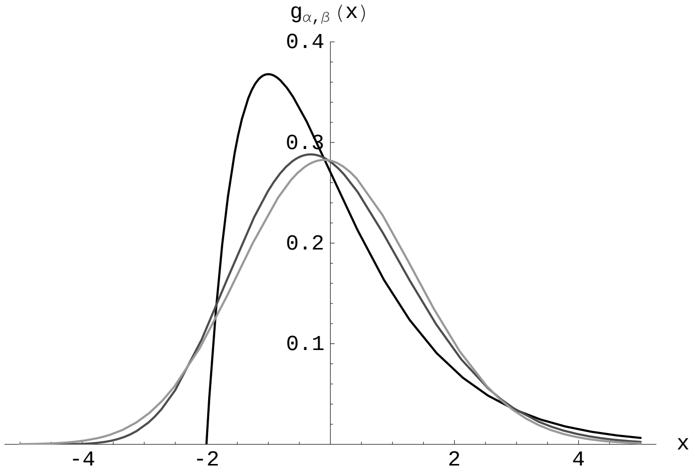


Figure 3.7.: Examples of the density of three gamma distributions with the same variance of 2 and $\alpha = 2, \beta = 1$ (black), $\alpha = 20, \beta = \sqrt{0.1}$ (dark grey), $\alpha = 200, \beta = \sqrt{0.01}$ (light grey).

variance is $\alpha\beta^2$ and its skewness is solely determined by α which is $2/\sqrt{\alpha}$. Thus, if α tends to ∞, the distribution tends to symmetry. For large α small changes in α would not dramatically affect the shape of the distribution if β is adjusted to fix the variance $\alpha\beta^2$. To illustrate this, figure 3.7 shows density functions of three gamma distribution with $\alpha = 2$, $\alpha = 20$, and $\alpha = 200$, but all with a variance $\alpha\beta^2 = 2$ adjusted by a corresponding value for β (1, $\sqrt{0.1}$, and $\sqrt{0.01}$, respectively). While a difference in the distributions for $\alpha = 2$ and $\alpha = 20$ can be observed, the distributions for $\alpha = 20$ and $\alpha = 200$ almost superimpose each other. Thus, for large numbers changes of α have practically no effect on the shape of the distribution, because β is adjusted automatically by the equation 3.16 if the optimisation method is used.

The estimation method restricts the search space by the use of the equations 3.16–3.21 which do not contain all moments. Naturally, the use of all moments derived from the three pairwise comparisons and the sample-i-intermediate distributions would be an improvement in order to estimate a parameter assignment which describes examined data best. Then an alternative is to search in the space of all 6 parameters α, β, d_1, d_2, d_3, and d_4. This was tried out (not shown here). To measure the quality of an estimate, the χ^2-score between the moments taken from simulations and the moments taken from the real data was used. In a different approach the likelihood was calculated and used as a criterion for the quality of an estimate (the likelihood function for a general model is introduced in chapter 4.2.3). However, the problem to find the optimum in this 6-dimensional search space is non-linear, the fitness-landscape is cliffy, and each query of the χ^2-score function or likelihood function as well results in a large number of simulations, because there are no closed equations for the moments of Z_{S1I} and Z_{S2I}. Another problem is to get stuck in local optima. To optimise the χ^2-score or likelihood function (1) the Downhill Simplex Method by Nelder and Mead (1965) and (2) an evolutionary algorithm (cf. Bäck and Schwefel (1993)) were tested. Unfortunately, both methods are extremely time consuming and therefore not efficient. Thus, the approach described in chapter 3.2.2 was used.

The M-gamma model was applied to microarray data from primates and mice. Thereby, one simplification was made: It was abstracted from the genealogies of the different individuals of the compared species. However, one can neglect this, since the divergence times between the compared species are exceedingly greater than the times to the common ancestors of the individuals within the species. The results for primates agree with previous results: Khaitovich *et al.* (2005b) observed an acceleration on the human lineage in brain but not in liver. However, another explanation for the different speed of evolution between human and chimpanzee brain is a decelerated evolution in chimpanzee brain instead of an acceleration in the human brain. Admittedly, these results depend on the ratios of the three robust and time-linear variances $\hat{v}_{1,2}$, $\hat{v}_{1,3}$, and $\hat{v}_{2,3}$. However, estimates for the expected number of mutations d_i with $i \in \{1, 2, 3, 4\}$ differ much among the two brain data sets (cf. figure 3.5). This might be due to the challenging estimation for α, since α scales the d_i parameters. Thus, one can conclude that these estimates are weak and more research is necessary.

The estimates for d_3 are problematic, since they are negative in all four primate data sets. The summary statistics used to estimate the ratio between the mutations on the branch to human/chimpanzee and the branch to the outgroup is $\hat{s}_{1,3}$. If this estimate differs strongly from the expectation, it is possible that one of the estimates $\hat{d}_3$ or $\hat{d}_4$ describing the mutations on the branches outgoing from the root is greater than expected for the variance between sample 1 and 3. Then the other parameter compensates it by a negative value. However, it is not necessary to regard this root placement problem if being only interested in the overall number of mutations d_3+d_4. Nevertheless, it is relevant to know why the problem of negative values for d_3 occurs in all primate data sets. A reason for this might be the used orangutan outgroup. The four different data sets consist of those genes whose probe oligonucleotide sequences match between human and chimpanzee. However, it was not checked whether they also match the orangutan sequences. Thus, the negative values might be the result of weak hybridisation depending on mismatching bases which led to a poor estimate for $s_{1,3}$. However, it is not easy to simulate such a situation in order to prove this assumption, since the M-gamma model simulates gene expression differences between samples, but no absolute gene expression levels, because the real ancestral states are unknown. Thus, it is critical to simulate weak hybridisation by decreasing the signal in simulated outgroup data to check whether it results in poor estimates for d_3 and d_4.

The estimates for α and β which describe the gamma distributed MED provide interesting results. However, all results for α have large confidence limits. Indeed, this can be explained by the nature of the gamma distribution as discussed previous. The differences in the results between the two array technologies Affymetrix HG U95Av2 and Affymetrix U133plus2 are greater than between the two different tissues. The MEDs estimated from the "133"-sets have a greater skewness and a greater variance than the MEDs estimated from the "95"-sets (cf. figure 3.4) for the same tissue. Since the "133"-sets have a four to five times greater number of genes than the "95"-sets, the results are statistically more powerful.

The results on the three mice data sets are not directly comparable with the results from Voolstra *et al.* (2007), since a different approach was used. Beside a SAM analysis (Tusher *et al.*, 2001) in which genes with differential expression between the species and subspecies were picked out, Voolstra *et al.* (2007) used the scaled divergence measure introduced by Lemos *et al.* (2005) that is the quotient of the between-species component of variance and the within-species component of variance. With this approach it was observed that testis has a large scaled divergence between species in comparison to the other tissues, but a small scaled divergence between subspecies. Indeed, 3-4 different subspecies were used. In the analysis with the M-gamma model only the subspecies dom and mus were used together with *Mus spretus* as the outgroup. However, the overall number of mutations between dom and mus is the smallest for testis. Since the variance of mutation effects is $\alpha\beta^2 = 0.325$ for testis (1.611 for brain and 1.200 for liver/kidney), it is in good agreement with the results by Voolstra *et al.* (2007). In contrast, the length of the branch to spretus is smaller than in liver which is twice as big there. However, a reason for this might be that the additional subspecies were not taken into account. Hence, a useful extension would be a model which can consider more than three samples together.

The use of a negatively skewed distribution for the liver/kidney mice data set extends the results by Khaitovich *et al.* (2005b). They suggested a positively skewed distribution for the mutation effects (indeed, only for the primate data). However, the estimated MED for liver/kidney is merely slightly negatively skewed with -0.121. Thus, an explanation for this result is noise in the data set, particularly all other data sets can be described by a positively skewed distribution. From the statistical point of view, the data sets are all critical anyway, because of the small sample size for each species.

3.5. Conclusion

The described M-gamma model provides additional information about the process of gene expression evolution, since it extends the MED by an additional parameter which makes it more flexible. The results from biological data sets indicate that a positively skewed MED is in most cases superior than a negatively skewed MED or a symmetric MED. Unfortunately, some model parameters estimated from applied data sets have large confidence intervals. Beside properties of the gamma distribution, this can be attributed to the small size of the data sets. Small sets are more sensitive against environmental effects, metabolic processes, and other non-mutation effects. Thus, a beneficial extension would be a method to estimate the influence of all non-mutational effects. This problem is addressed in the next chapter.

4. A model with mutational and non-mutational effects

In this chapter a gene expression evolution model is presented which includes effects that change the level of gene expression without mutating regulatory regions on DNA sequence level. These so-called non-mutational effects are caused, for example, by the environment and the cell cycle. The resulting model describes real data in a better way.

4.1. Introduction

In recent years several studies have been published with special focus on the evolution of gene expression. Typically, differences in gene expression between closely related species were compared (cf. reviews by Gilad *et al.* (2006b), Ranz and Machado (2006), and Khaitovich *et al.* (2006)). Accumulation of differences with time was reported frequently which led to the idea to apply the neutral theory by Kimura (1983) to the transcriptome. Thus, Khaitovich *et al.* (2005b) developed a neutral model for evolution of gene expression which is referred to as M model here. In this model (and also in the M-gamma model in chapter 3) random mutations occur in the regulatory region of a gene which increase or decrease the mRNA abundance of that gene. Other effects which change the level of gene expression are not addressed. Hence, all the differences between two samples (e.g., from two different species) are completely attributed to evolution.

This is a strong simplification, since mutations cause only a fraction of expression changes. Moreover, the expression level of a gene is influenced by non-mutational effects like all kinds of metabolic pathways, the cell cycle, epistatic effects, life history, and potential diseases. Furthermore, expression measurements by microarrays are noisy caused by variance in the hybridisation process and technical measurement errors while scanning (Baldi and Hatfield, 2002; Speed, 2003). Lemos *et al.* (2005) conjectured that half of the variance in gene expression within a population is caused by environmental variance. Khaitovich *et al.* (2004) estimated that the environmental component of variation is roughly three times greater than the genetic component.

For a better understanding of gene expression evolution it is indispensable to differentiate between mutational and non-mutational effects. A first step towards this goal is to estimate the impact of these two kinds of effects. The problem is addressed in this chapter by embedding all non-mutational effects into the M model by Khaitovich *et al.* (2005b). To this end, another random variable is introduced which summarises all influences not depending on mutations. This extended model is referred to as *M&E-model* (= Mutation and error). The non-mutational influences can be regarded as noise if one wants to observe evolutionary differences. To estimate the model parameters of this more complex model, an optimisation method is required. Two types of optimisation methods are discussed: (1) a χ^2-fit method which can be used for normal distributed and extreme value distributed mutation effects and (2) a maximum-likelihood (ML) method which provides a convenient solution for normal distributed mutation effects. The methods are applied to biological data to illustrate their applicability.

Finally, the M&E-model leads to a methodology that detects genes which changed their expression level due to mutations in regulatory regions during evolution. These genes are of special interest, since they might be the cause for phenotypic differences depending on regulatory differences. A simple Bayesian method is presented to detect these genes.

4.2. Materials and methods

4.2.1. The M&E model

The M&E-model is an extension of the M model by Khaitovich *et al.* (2005b). Corresponding to the M model, mutations in the regulatory region of a gene in a sample 1 affect the expression level of that gene. For time t_1 scaled by the mutation rate μ it is described by the compound Poisson process

$$Y_1(d_1) = Y_1(0) + \sum_{i=1}^{M(d_1)} X_i, \tag{4.1}$$

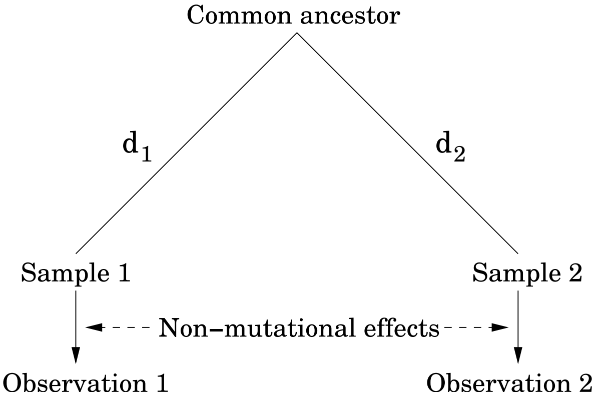


Figure 4.1.: Tree of two samples descended from a common ancestor. Mutation based variations in the difference of gene expression levels of the samples are not observable, since the mutational changes are overlayed by non-mutational effects.

whereas $d_1 = \mu t_1$. $M(d_1)$ is the Poisson distributed random variable to describe the number of mutations and X_i is a random variable which describes the effect of the i-th mutation on the log-scaled gene expression level. All X_i follow the same mutation effect distribution (MED) X. $Y_1(0)$ is the initial expression level (cf. chapter 3.2.1 for more detail). In the M&E-model non-mutational effects are modelled additionally by another random variable E_1 with a mean of zero which can be taken as statistical noise (E means error). The distribution of E_1 is referred to as *non-mutational effect distribution* (N-MED). With E_1 the process of gene expression evolution is described by

$$Y_1^E(d_1) = Y_1(d_1) + E_1. \tag{4.2}$$

Let $Z_{1,2}^E$ be a random variable describing the difference in expression of a gene between two samples 1 and 2 which diverged from a common ancestor independently with parameters d_1 and d_2, respectively. Let E_1 and E_2 be the corresponding non-mutational effects. E_1 and E_2 follow the same N-MED (cf. figure 4.1). Then $Z_{1,2}^E$ is defined by

$$Z_{1,2}^E = \left(\left(Y_1(0) + \sum_{i=1}^{M(d_1)} X_i \right) + E_1 \right) - \left(\left(Y_2(0) + \sum_{j=1}^{M(d_2)} X_j \right) + E_2 \right). \tag{4.3}$$

Please note that $Y_1(0) = Y_2(0)$. It follows that

$$Z_{1,2}^E = Y_1^E(d_1) - Y_2^E(d_2). \tag{4.4}$$

The parameters of this model are d_1 and d_2. Remaining parameters are those specifying the MED and the N-MED. The two types of MEDs used by Khaitovich *et al.* (2005b) are also applied here: (1) a normal distribution and (2) an extreme value distribution. The corresponding variants of the M&E model are referred to as M&E-normal and M&E-extreme model, respectively. The non-mutational effects are always assumed to follow a normal distribution with standard deviation σ_e. The moments of $Z^E_{1,2}$ for the model with both types of MEDs can be derived by using characteristic functions (cf. chapter 2.3.1 or Feller (1957) for details).

Moments of $Z^E_{1,2}$ of the M&E-normal model: Let σ_m be the standard deviation of the normal distributed MED. Then the moments are

$$\text{Variance } v_{1,2}^{(normal)} = \mu_2(Z_{1,2}^{E(normal)}) = \sigma_m^2(d_1 + d_2) + 2\sigma_e^2, \tag{4.5}$$

$$\text{Coefficient of skewness } s_{1,2}^{(normal)} = \gamma_1(Z_{1,2}^{E(normal)}) = 0, \tag{4.6}$$

$$\text{Coefficient of kurtosis } k_{1,2}^{(normal)} = \gamma_2(Z_{1,2}^{E(normal)}) = 3 + \frac{3\sigma_m^4(d_1 + d_2)}{(\sigma_m^2(d_1 + d_2) + 2\sigma_e^2)^2}. \tag{4.7}$$

In Khaitovich *et al.* (2005b) the coefficient of skewness of $Z_{1,2}$ is used to separate the two parameters d_1 and d_2. However, the coefficient of skewness is zero for normal distributed MEDs. Thus, one cannot estimate d_1 and d_2 separately. Therefore, $d = d_1 + d_2$ is estimated in this chapter, whenever a normal distributed MED is used.

Moments of $Z^E_{1,2}$ of the M&E-extreme model: Let β be the parameter of the extreme value distributed MED. Then the moments are

$$\text{Variance } v_{1,2}^{(extreme)} = \mu_2(Z_{1,2}^{E(extreme)}) = \frac{\pi^2}{6}\beta^2(d_1 + d_2) + 2\sigma_e^2, \tag{4.8}$$

$$\text{Coefficient of skewness } s_{1,2}^{(extreme)} = \gamma_1(Z_{1,2}^{E(extreme)}) = \frac{c(d_1 - d_2)}{\left(d_1 + d_2 + \frac{12\sigma_e^2}{\pi^2\beta^2}\right)^{3/2}}, \tag{4.9}$$

$$\text{Coefficient of kurtosis } k_{1,2}^{(extreme)} = \gamma_2(Z_{1,2}^{E(extreme)}) = 3 + \frac{3\pi^2\beta^4(d_1 + d_2)}{20(\mu_2(Z_{1,2}))^2}. \tag{4.10}$$

The constant c is $12\sqrt{6}\zeta(3)/\pi^3 \approx 1.13955\ldots$, whereas $\zeta(.)$ is the ζ-function.

To avoid notational difficulties the moments are referred to as $v_{1,2}$, $s_{1,2}$, and $k_{1,2}$ if the described properties do not depend on the used MED. The moments estimated from distributions of gene expression difference of real data are abbreviated as $\hat{v}_{1,2}$, $\hat{s}_{1,2}$, and $\hat{k}_{1,2}$. The goal is to estimate the model parameters for the M&E-normal and the M&E-extreme model from real data. To do this one assumes that all gene expression differences follow the same evolutionary process described by equation 4.4. Since the genes are independent of each other in the model, *trans*-effects are not described.

4.2.2. Parameter estimation with a χ^2-fit method

One way to estimate parameters is the method of moments (Pearson, 1894) which has been used by Khaitovich *et al.* (2005b) for the M model. To this end, the moment estimates $\hat{v}_{1,2}$, $\hat{s}_{1,2}$, and $\hat{k}_{1,2}$ are inserted into the equations 4.5–4.7 for the M&E-normal model or into the equations 4.8–4.10 for the M&E-extreme model. However, in both cases it is not possible to get a unique solution.

If the MED is a normal distribution, only the equations 4.5 and 4.7 can be used, since the equation for the coefficient of skewness (equation 4.6) is always zero. Since the model has three parameters, a unique solution cannot be obtained with two equations. If the MED is an extreme value distribution, four parameter have to be estimated which is not possible with the three equations 4.8–4.10.

However, it is possible to give a partial solution depending on one of the parameters by transforming the equation system of the moments. For instance, a solution depending on σ_e for the M&E-normal model looks as follows (a solution for the M&E-extreme model can be given analogous):

$$\hat{\sigma_m} = \hat{v}_{1,2}\sqrt{\frac{\hat{k}_{1,2} - 3}{3\hat{v}_{1,2} - 6\sigma_e^2}} \tag{4.11}$$

$$\hat{d} = \frac{3(2\sigma_e^2 - \hat{v}_{1,2})}{(\hat{v}_{1,2})^2(\hat{k}_{1,2} - 3)} \tag{4.12}$$

Thus, if one can estimate σ_e, a solution for the remaining parameters can be obtained by applying equations 4.11 and 4.12. To this end, a χ^2-fit approach is applied which uses model-based computer simulation.

Simulation of gene expression differences: The parameters of the M&E-normal model are summarised by $\Theta^{(normal)} = (\sigma_m, \sigma_e, d)$, and the parameters of the M&E-extreme model are summarised by $\Theta^{(extreme)} = (\beta, \sigma_e, d_1, d_2)$. If the model is not specified, Θ is used instead. A simulated distribution of gene expression differences depending on the model parameters Θ is denoted as $S(\Theta)$. After specifying the parameters, the process is simulated by choosing zero as the initial gene expression value.

For normal distributed MEDs the two samples are simulated (cf. figure 4.1) by adding $M(d/2)$ mutation effects for each of the two lineages to the initial value. Please note that the position of the root has no effect here (cf. chapter 4.2.3 for an explanation that the root position does not matter in the M&E-normal model). For extreme value distributed MEDs the two samples are determined (cf. figure 4.1) by adding independently to the initial value $M(d_1)$ mutation effects for the first sample and $M(d_2)$ mutation effects for the second sample.

Thereupon, the non-mutational effects are simulated for both samples in each case by adding a number drawn from a normal distribution with standard deviation σ_e. After that, the gene expression difference between sample 1 and 2 is calculated. The simulation is iterated up to 10^7 times to obtain a large number of simulated gene expression differences.

Optimisation: Let $O = (o_1, \ldots, o_\eta)$ be the collection of all η gene expression differences between the two samples in an applied data set ($O =$ observation). It is the objective to choose the model parameters Θ that way that a distribution $S(\Theta)$ generated by the parameters Θ matches O best ($S =$ simulation). To this end, a χ^2-fit approach is used. Let $\texttt{min}_O$ be the smallest element of O and let $\texttt{max}_O$ be the largest element of O. According to O, a partition of the real-numbers into 100 disjoint intervals I_i with $i = 1, \ldots, 100$ is introduced. These intervals are

$$I_1 =]-\infty, \texttt{min}_O],\ I_2 =]\texttt{min}_O, \texttt{min}_O + \delta],\ \ldots,\ I_{100} =]\texttt{max}_O, \infty], \tag{4.13}$$

each with interval length

$$\delta = \frac{\mathtt{max}_O - \mathtt{min}_O}{100 - 2}. \tag{4.14}$$

$S(\Theta)$ is binned into the same intervals $I_1, \ldots, I_{100}$. Let p_i be the size of the subset of O which is mapped to interval I_i divided by η, and let $q_i(\Theta)$ be the size of the subset of $S(\Theta)$ which is mapped to interval I_i divided by $|S(\Theta)|$. Then the following χ^2 objective function is minimised:

$$\chi^2(\Theta) = \sum_{i=1}^{100} \frac{(p_i - q_i(\Theta))^2}{q_i(\Theta)} \tag{4.15}$$

However, there might be different assignments of Θ with the same minimal $\chi^2(\Theta)$ value. To solve this problem of local optima, the moments of $Z^E_{1,2}$ are used. For the M&E-normal model an estimate for σ_e is inserted into the equations 4.11 and 4.12 to calculate $\hat{\sigma}_m$ and $\hat{d}$. The moments $\hat{v}_{1,2}$ and $\hat{k}_{1,2}$ required for the equations 4.11 and 4.12 are estimated from O. Afterwards, $\hat{\sigma}_e$ and the calculated $\hat{\sigma}_m$ and $\hat{d}$ are used to generate $S(\Theta)$ by simulations. Thereafter, $\chi^2(\Theta)$ is calculated. Hence, the χ^2-value depends only on σ_e. Thus, it can be described by a one-dimensional function referred to as $\chi^2(\sigma_e)$ (analogous for the M&E-extreme model).

$\chi^2(\sigma_e)$ can be minimised by using the Brent's method (cf. chapter 2.4.3 or Brent (1972)). Before starting the Brent's method, it is necessary to initialise it with three points $x < y < z$ with $\chi^2(y) \leq min\{\chi^2(x), \chi^2(z)\}$ so that y is a best estimate for the minimum (starting condition).

To find these starting values, one chooses $x := 0, y := \sigma_{e_{max}}/2$ and $z := \sigma_{e_{max}}$, whereas $\sigma_{e_{max}} < \sqrt{\hat{v}_{1,2}/2}$ is the largest possible value for σ_e which is defined by the domain of equation 4.11 in case of a normal distributed MED (analogous for an extreme value distributed MED).

If $\chi^2(x) < \chi^2(y) < \chi^2(z)$, one chooses $x := 0, y := y/2, z := z/2$ and checks the starting condition recursively. If $\chi^2(x) > \chi^2(y) > \chi^2(z)$, one chooses $x := y, y := y + (z - y)/2, z := z$ and checks the starting condition recursively. If the starting condition is fulfilled, the Brent's method is started (cf. Brent (1972)).

Within each optimisation step a series of 10^7 simulations is accomplished to regenerate $S(\Theta)$. The method stops if the distance between x and z is smaller than 0.001.

4.2.3. Parameter estimation with a ML method

A ML method provides an alternative to estimate model parameters. To build the likelihood function, the probability mass/density functions of the used random variables must be derived. The number of mutations on the lineage to sample 1 depending on parameter d_1 is described by the discrete Poisson distribution. Its probability mass function is given by

$$p_i = \frac{e^{-d_1} d_1^i}{i!}, \tag{4.16}$$

for $i = 0, 1, \ldots$. The density function of the MED is denoted by $f_X(x)$. The density function of the N-MED is referred to as $f_E(x)$. Let $f_X^{(i)}$ be the i-th convolution of the MED which describes the sum $X_1 + \ldots + X_i$. Thus, the density function for the compound Poisson process $Y_1^E(d_1)$ is

$$f_{Y_1^E}(x) = \left(\sum_{i=0}^{\infty} \frac{e^{-d_1} d_1^i}{i!} f_X^{(i)}(x) \right) * f_E(x), \tag{4.17}$$

whereas $*$ indicates the convolution of mutational effects and non-mutational effect. Please note that $f_X^{(0)} = 1$. Due to the monotone convergence theorem and because the convolution satisfies distributivity, it follows

$$f_{Y_1^E}(x) = \sum_{i=0}^{\infty} \frac{e^{-d_1} d_1^i}{i!} (f_X^{(i)}(x) * f_E(x)) \tag{4.18}$$

(cf. Taylor and Karlin (1998) for details about compound Poisson processes). This equation describes the evolution of gene expression on the branch to sample 1. Unfortunately, the ancestral strength of the gene expression level $Y_1(0)$ is not observable. However, two contemporary samples derived from the same ancestor are considered. The gene expression difference between them is computable and is described by the random variable $Z_{1,2}^E$ (cf. equation 4.4). The density function of $Z_{1,2}^E$ is the convolution of $f_{Y_1^E}(x)$ and $f_{Y_2^E}(-x)$ which is

$$f_{Z_{1,2}^E}(x) = \int_{-\infty}^{\infty} f_{Y_1^E}(y) \cdot f_{Y_2^E}(y - x) dy. \tag{4.19}$$

This equation does not make any assumptions about the MED and the N-MED. However, it might be difficult to obtain an analytical formula for the integral. In the following a M&E-normal model is considered. Thus, the X_i with $1, 2, \ldots$ follow a normal distribution with standard deviation σ_m, and E_1 and E_2 follow a normal distribution with standard deviation σ_e. Since a convolution of two normal distributed random variables with variances σ_1^2 and σ_2^2 is equal to a normal distributed random variable with variance $\sigma_1^2 + \sigma_2^2$, it follows that the i-th convolution of f_X is

$$f_X^{(i)}(x, \sigma_m) = \frac{1}{\sqrt{2\pi i \sigma_m^2}} e^{-x^2/2i\sigma_m^2} \tag{4.20}$$

under the M&E-normal model. The density function of the N-MED is given by the normal distribution

$$f_E(x, \sigma_e) = \frac{1}{\sqrt{2\pi}\sigma_e} e^{-x^2/2\sigma_e^2}. \tag{4.21}$$

Thus, the convolution of $f_X^{(i)}$ and f_E is also a normal distribution with variance $i\sigma_m^2 + \sigma_e^2$ described by

$$f_X^{(i)}(x, \sigma_m) * f_E(x, \sigma_e) = \frac{1}{\sqrt{2\pi}\sqrt{i\sigma_m^2 + \sigma_e^2}} e^{-x^2/2(i\sigma_m^2 + \sigma_e^2)}. \tag{4.22}$$

Since all effects are normal distributed, it is not necessary to calculate a complex numerical solution for the integral to describe $Z_{1,2}^E$ in equation 4.19. It is rather possible to replace the difference of two compound Poisson processes with parameters d_1 and d_2 by one Poisson process with parameter $d = d_1 + d_2$ to describe the same gene expression difference, since it is not possible to estimate the ratio between d_1 and d_2 anyway. To prove this claim two propositions have to be shown: (1) the sum of the expected number of events of two Poisson processes with parameters k and l is equal to the expected number of events of a Poisson process with parameter $k + l$. This is well known (cf. Taylor and Karlin (1998)). Hence, the expected number of events is equal. (2) it has no effect on the gene expression difference whether mutation effects are subtracted or summed up for random variables X_i which follow to the same normal distribution. To prove this it has to be shown that $X^{(i-1)} + X = X^{(i-1)} - X$. The equalisation of both density functions is

$$\int_{-\infty}^{\infty} f_X^{(i-1)}(y) \cdot f_X(x - y) dy = \int_{-\infty}^{\infty} f_X^{(i-1)}(y) \cdot f_X(y - x) dy. \tag{4.23}$$

A value inserted into the density function of a normal distribution is squared. Thus, the term $f_X(x-y) = f_X(y-x)$ holds for the density function of any normal distribution. Therefore, the proposition is proved, and the density function of $Z^E_{1,2}$ for normal distributed mutational and non-mutational effects is

$$f_{Z^E_{1,2}}(x) = \sum_{i=0}^{\infty} \frac{e^{-d} d^i}{i!} (f_X^{(i)}(x) * f_E^{(2)}(x)). \tag{4.24}$$

Please note that $f_E^{(2)}(x)$ is used (2-fold convolution of the N-MED), because non-mutational effects take place at both samples. Equation 4.24 allows a complete ML estimation of the parameters σ_m, σ_e, and d. Assume now that $O = (o_1, \ldots, o_\eta)$ is the collection of all η gene expression differences between the two samples in the data set. Under the assumption that these differences are independently and identically distributed and that all genes follow the same process, the likelihood to observe these values is

$$f_{Z^E_{1,2}}(o_1) \cdot \ldots \cdot f_{Z^E_{1,2}}(o_\eta) = \prod_{i=1}^{\eta} f^E_{Z_{1,2}}(o_i). \tag{4.25}$$

To maximise this likelihood function, the downhill simplex search is applied with a stop value of 0.0001 (cf. chapter 2.4.4 or Nelder and Mead (1965)). If not stated otherwise, 10 repeats of the downhill simplex search are performed for each estimation, and the parameters leading to the smallest log-likelihood constitute the final estimates.

4.2.4. A Bayesian method to detect the number of mutations

When the M&E-normal model is applied to microarray data of two samples, estimates for σ_m, σ_e, and d are obtained. However, it is important to get more information about single genes. It is a major task to decide between genes mutated in their regulatory region during evolution and genes in which only non-mutational effects caused gene expression differences. To this end, a Bayesian method is applied. This method estimates the number of mutations which describes the expression difference of a gene between two samples best. For this purpose it is necessary to calculate the likelihood that a fixed number of mutations cause an observed gene expression difference given $\hat{\sigma}_m$, $\hat{\sigma}_e$, and $\hat{d}$. These parameter values have been estimated before with the ML method (cf.

chapter 4.2.3). Corresponding to equation 4.16 the probability for k mutations is given by

$$P(M(\hat{d}) = k) = \frac{e^{-\hat{d}}\hat{d}^k}{k!} \tag{4.26}$$

which is the prior probability for the Poisson distributed number of mutations $M(d)$. The likelihood for a gene expression difference x given k mutations and the non-mutational effects is calculated by the k-fold convolution of the MED with standard deviation $\hat{\sigma}_m$ convolved with the 2-fold convolution of the N-MED with standard deviation $\hat{\sigma}_e$. This is equal to a normal distribution with standard deviation $\sqrt{\hat{\sigma}_m^2 k + 2\hat{\sigma}_e^2}$. Its density function is, corresponding to equation 4.24, given by

$$h(x|k) = \frac{1}{\sqrt{2\pi}\sqrt{\hat{\sigma}_m^2 k + 2\hat{\sigma}_e^2}} e^{-x^2/(2\hat{\sigma}_m^2 k + 4\hat{\sigma}_e^2)}. \tag{4.27}$$

Combining now equation 4.26 and 4.27, the likelihood to observe x is

$$\sum_{k=0}^{\infty} P(M(\hat{d}) = k) \cdot h(x|k). \tag{4.28}$$

It follows that the posterior density is

$$pp(k|x) = \frac{P(M(\hat{d}) = k) \cdot h(x|k)}{\sum_{k=0}^{\infty} P(M(\hat{d}) = k) \cdot h(x|k)}, \tag{4.29}$$

whereas the denominator is the normalising constant. Let ω be the number of all gene expression difference for a gene j between two samples (which is greater 1 if at least one sample contains more than one individual). For this collection $O_j = (o_{1,j}, \ldots, o_{\omega,j})$ it is the goal to find that value for k so that

$$\prod_{i=1}^{\omega} pp(k|o_{i,j}) \tag{4.30}$$

is maximal. Then k is the number of mutations which has the largest posterior probability.

4.3. Experiments and results

4.3.1. Evaluation of the parameter estimation method

Gene expression differences from synthetic data sets were used to evaluate and compare the estimation methods. The synthetic data sets were generated for different parameter assignments via computer simulation as described in chapter 4.2.2. For each gene one simulation was performed. Depending on the random nature of the stochastic process for each selected parameter assignment $1,000$ data sets were generated each with (1) $1,000$ genes, (2) $10,000$ genes, and (3) $100,000$ genes, respectively. Accordingly, an estimation method was used to estimate the parameters from the synthetic data sets. The resulting estimates were compared with the real parameter values used for generation.

M&E-normal model: At first the M&E-normal model was evaluated. Here, both the χ^2-fit method and the ML method were applied. The test cases are shown in the first column of table 4.1. They have all been chosen that way that one mutation takes place between the two samples in expectation. The standard deviation of non-mutational effects differs between 0.1 and 1. Mean estimates and 95 % confidence limits (in brackets) from all $1,000$ data sets are presented in the remaining columns of table 4.1.

A general observation is that the distance of the mean estimates from the selected parameters is increased if the standard deviation of non-mutational effects σ_e is increased. An increase of σ_e also increases the width of the confidence limits. Deviation from the selected parameters and the confidence limit as well, are also affected by the number of genes. If their number is increased, deviation and confidence limit width are decreased in the majority of cases. One would expect this, since the number of realisations of the stochastic process is increased which result in better estimates (i.e., closer to the selected parameters).

Overall the ML method is superior to the χ^2-fit method. Especially in those cases in which the variance caused by non-mutational effects is equal or greater than the variance of mutational effects the mean d-estimates are extremely overestimated by the χ^2-fit method (up to five times greater than expected, cf. the case $(1, \sqrt{1/2}, 1)$ and the case $(1, 1, 1)$ in table 4.1). In contrast, the ML method seems to be asymptotically unbiased. With the smallest standard daviation of non-mutational effects $\sigma_e = 0.1$ the estimates with the ML method are nearly perfect on the third decimal place, even when

Table 4.1.: Results for synthetic data with the M&E-normal model. In some cases σ_e was set to $\sqrt{1/2} \approx 0.70711$ so that the variance of mutational and non-mutational effects is equal (to 1).

(σ_m, σ_e, d)	χ^2-fit			ML		
#genes	$\hat{\sigma}_m$	$\hat{\sigma}_e$	$\hat{d}$	$\hat{\sigma}_m$	$\hat{\sigma}_e$	$\hat{d}$
$(1, 0.1, 1)$	0.975	0.096	1.098	1.001	0.100	1.000
10^3	(0.764 , 1.257)	(0.067 , 0.130)	(0.636 , 1.687)	(0.920 , 1.090)	(0.087 , 0.114)	(0.872 , 1.120)
$(1, 0.1, 1)$	0.998	0.100	1.011	1.000	0.100	1.001
10^4	(0.921 , 1.091)	(0.090 , 0.111)	(0.842 , 1.170)	(0.971 , 1.028)	(0.095 , 0.105)	(0.959 , 1.042)
$(1, 0.1, 1)$	1.002	0.103	0.994	1.000	0.100	1.000
10^5	(0.973 , 1.035)	(0.096 , 0.106)	(0.933 , 1.052)	(0.988 , 1.013)	(0.098 , 0.102)	(0.981 , 1.019)
$(1, 0.5, 1)$	0.999	0.459	1.390	1.028	0.484	1.110
10^3	(0.583 , 1.587)	(0.100 , 0.689)	(0.261 , 4.047)	(0.726 , 1.457)	(0.227 , 0.673)	(0.296 , 2.440)
$(1, 0.5, 1)$	0.993	0.491	1.063	1.004	0.499	1.015
10^4	(0.844 , 1.163)	(0.398 , 0.567)	(0.638 , 1.636)	(0.866 , 1.159)	(0.413 , 0.574)	(0.646 , 1.524)
$(1, 0.5, 1)$	0.998	0.497	1.016	1.005	0.501	0.997
10^5	(0.935 , 1.071)	(0.431 , 0.538)	(0.804 , 1.274)	(0.920 , 1.098)	(0.448 , 0.548)	(0.762 , 1.273)
$(1, \sqrt{1/2}, 1)$	0.934	0.500	3.458	1.212	0.818	1.331
10^3	(0.378 , 1.983)	(0.000 , 0.905)	(0.101 , 12.43)	(0.794 , 1.990)	(0.360 , 1.119)	(0.128 , 2.860)
$(1, \sqrt{1/2}, 1)$	1.002	0.660	1.372	1.011	0.697	1.097
10^4	(0.672 , 1.498)	(0.304 , 0.865)	(0.225 , 3.914)	(0.797 , 1.317)	(0.550 , 0.825)	(0.377 , 2.087)
$(1, \sqrt{1/2}, 1)$	0.985	0.688	1.123	0.994	0.698	1.073
10^5	(0.802 , 1.142)	(0.484 , 0.770)	(0.630 , 2.405)	(0.854 , 1.167)	(0.607 , 0.777)	(0.596 , 1.720)
$(1, 1, 1)$	0.938	0.543	5.374	1.080	1.018	0.993
10^3	(0.467 , 2.861)	(0.008 , 1.184)	(0.037 , 13.25)	(0.224 , 1.946)	(0.698 , 1.223)	(0.043 , 2.330)
$(1, 1, 1)$	0.749	0.586	5.534	1.064	1.007	1.027
10^4	(0.471 , 1.517)	(0.015 , 1.115)	(0.227 , 12.92)	(0.774 , 1.677)	(0.858 , 1.152)	(0.129 , 2.088)
$(1, 1, 1)$	0.811	0.665	4.230	1.016	0.995	1.088
10^5	(0.547 , 1.346)	(0.000 , 1.102)	(0.314 , 9.854)	(0.819 , 1.355)	(0.888 , 1.102)	(0.326 , 2.064)

using only $1,000$ genes per data set. Actually, this is a more precise result than with the χ^2-fit method with $100,000$ genes. The ranges of the confidence limits are smaller in almost each case if the ML method is applied. Thus, as the conclusion ML should be used, when using the M&E model.

M&E-extreme model: The situation for the M&E-extreme model is different. Thereby, it is not practical to use a ML method, since the convolution of extreme value distributions is not as easy to calculate as the convolution of normal distributions. Since numerical methods and simulations would be necessary, the computational resources would become very large, particularly because the search space has one additional dimension depending on the separation of $\hat{d}_1$ and $\hat{d}_2$. Thus, only results for the χ^2-fit method are presented here. They are shown in the tables 4.2 and 4.3, whereupon table 4.2 contains

Table 4.2.: Results for synthetic data with the M&E-extreme model (clock-like cases with $d_1 = d_2$). The variance of the mutation effect distribution was set to 1 by choosing $\beta = \sqrt{6/\pi^2} \approx 0.77970$. In some cases σ_e was set to $\sqrt{1/2} \approx 0.70711$ so that the variance of mutational and non-mutational effects is equal.

$(\beta, \sigma_e, d_1, d_2)$,#genes	$\hat{\beta}$	$\hat{\sigma}_e$	$\hat{d}_1$	$\hat{d}_2$
$(\beta, 0.1, 0.5, 0, 5), 10^3$	0.759	0.103	0.567	0.564
	(0.552 , 1.104)	(0.070 , 0.200)	(0.154 , 1.062)	(0.145 , 1.075)
$(\beta, 0.1, 0.5, 0.5), 10^4$	0.780	0.102	0.508	0.503
	(0.700 , 0.888)	(0.089 , 0.125)	(0.354 , 0.657)	(0.349 , 0.660)
$(\beta, 0.1, 0.5, 0.5), 10^5$	0.780	0.102	0.499	0.501
	(0.751 , 0.810)	(0.096 , 0.105)	(0.448 , 0.548)	(0.451 , 0.554)
$(\beta, 0.5, 0.5, 0.5), 10^3$	0.736	0.432	0.779	0.789
	(0.445 , 1.191)	(0.135 , 0.630)	(0.107 , 2.245)	(0.090 , 2.138)
$(\beta, 0.5, 0.5, 0.5), 10^4$	0.783	0.498	0.525	0.519
	(0.643 , 0.988)	(0.408 , 0.598)	(0.244 , 0.868)	(0.233 , 0.871)
$(\beta, 0.5, 0.5, 0.5), 10^5$	0.778	0.497	0.506	0.508
	(0.722 , 0.840)	(0.431 , 0.537)	(0.397 , 0.651)	(0.398 , 0.649)
$(\beta, \sqrt{1/2}, 0.5, 0.5), 10^3$	0.699	0.528	1.354	1.421
	(0.347 , 1.299)	(0.002 , 0.849)	(0.037 , 4.818)	(0.042 , 5.021)
$(\beta, \sqrt{1/2}, 0.5, 0.5), 10^4$	0.779	0.689	0.582	0.581
	(0.579 , 1.047)	(0.477 , 0.813)	(0.184 , 1.327)	(0.168 , 1.361)
$(\beta, \sqrt{1/2}, 0.5, 0.5), 10^5$	0.777	0.702	0.522	0.520
	(0.687 , 0.859)	(0.623 , 0.748)	(0.359 , 0.782)	(0.360 , 0.799)
$(\beta, 1.0, 0.5, 0.5), 10^3$	0.571	0.544	5.668	5.671
982 valid	(0.157 , 1.411)	(0.000 , 1.137)	(0.002 , 28.34)	(0.002 , 26.53)
$(\beta, 1.0, 0.5, 0.5), 10^4$	0.727	0.863	1.310	1.313
	(0.402 , 1.183)	(0.029 , 1.118)	(0.103 , 5.161)	(0.106 , 5.317)
$(\beta, 1.0, 0.5, 0.5), 10^5$	0.729	0.891	1.026	1.027
	(0.455 , 0.943)	(0.038 , 1.071)	(0.232 , 4.225)	(0.242 , 4.313)

results on clock-like cases with $d_1 = d_2$ and table 4.3 exemplifies parameter settings which were used to simulate accelerated evolution on one branch. The parameter β was always set to $\sqrt{6/\pi^2} \approx 0.77970$ so that the variance of mutation effects is 1 which is equal to the variance of mutation effects in the test cases used for the M&E-normal model (cf. table 4.1).

Table 4.3.: Results for synthetic data with the M&E-extreme model (asymmetric cases with $d_1 > d_2$). The variance of the mutation effect distribution was set to 1 by choosing $\beta = \sqrt{6/\pi^2} \approx 0.77970$. In some cases σ_e was set to $\sqrt{1/2} \approx 0.70711$ so that the variance of mutational and non-mutational effects is equal.

$(\beta, \sigma_e, d_1, d_2)$,#genes	$\hat{\beta}$	$\hat{\sigma}_e$	$\hat{d}_1$	$\hat{d}_2$
$(\beta, 0.1, 2/3, 1/3), 10^3$	0.757	0.101	0.752	0.389
	(0.546 , 1.085)	(0.070 , 0.159)	(0.290 , 1.302)	(0.046 , 0.841)
$(\beta, 0.1, 2/3, 1/3), 10^4$	0.779	0.102	0.675	0.337
	(0.694 , 0.885)	(0.089 , 0.120)	(0.494 , 0.854)	(0.213 , 0.466)
$(\beta, 0.1, 2/3, 1/3), 10^5$	0.780	0.101	0.666	0.334
	(0.749 , 0.815)	(0.096 , 0.105)	(0.600 , 0.725)	(0.294 , 0.376)
$(\beta, 0.5, 2/3, 1/3), 10^3$	0.739	0.435	1.028	0.565
	(0.429 , 1.271)	(0.160 , 0.635)	(0.182 , 2.648)	(0.014 , 1.968)
$(\beta, 0.5, 2/3, 1/3), 10^4$	0.779	0.496	0.699	0.356
	(0.647 , 0.981)	(0.411 , 0.588)	(0.349 , 1.085)	(0.153 , 0.628)
$(\beta, 0.5, 2/3, 1/3), 10^5$	0.778	0.497	0.676	0.341
	(0.725 , 0.837)	(0.431 , 0.537)	(0.544 , 0.844)	(0.255 , 0.440)
$(\beta, \sqrt{1/2}, 2/3, 1/3), 10^3$	0.674	0.500	1.920	1.136
	(0.318 , 1.242)	(0.002 , 0.842)	(0.121 , 6.780)	(-0.010 , 4.440)
$(\beta, \sqrt{1/2}, 2/3, 1/3), 10^4$	0.766	0.686	0.798	0.416
	(0.560 , 1.007)	(0.479 , 0.809)	(0.286 , 1.865)	(0.097 , 1.098)
$(\beta, \sqrt{1/2}, 2/3, 1/3), 10^5$	0.775	0.699	0.702	0.354
	(0.640 , 0.863)	(0.498 , 0.748)	(0.499 , 1.428)	(0.220 , 0.799)
$(\beta, 1.0, 2/3, 1/3), 10^3$	0.594	0.583	7.173	3.340
983 valid	(0.177 , 1.391)	(0.000 , 1.118)	(0.102 , 27.07)	(-0.097 , 19.24)
$(\beta, 1.0, 2/3, 1/3), 10^4$	0.707	0.847	1.736	1.079
	(0.387 , 1.145)	(0.055 , 1.111)	(0.176 , 6.587)	(0.058 , 4.682)
$(\beta, 1.0, 2/3, 1/3), 10^5$	0.724	0.889	1.315	0.778
	(0.456 , 0.934)	(0.150 , 1.065)	(0.357 , 5.064)	(0.150 , 3.458)

Similar to the M&E-normal model, the deviations in the mean estimates from the selected parameters and the ranges of the confidence limits as well are increased with an increase of the non-mutational component described by its standard deviation σ_e. Again, both are reduced with an increase of the gene number. In the case $\sigma_e = 1.0$ the estimates for d_1 and d_2 are extremely overestimated, actually for the case with $100,000$

genes (about twice as big). A comparison of the tables 4.2 ($d_1 = d_2$) and 4.3 ($d_1 > d_2$) shows that there are no big differences in the magnitude of deviation from the real values or the confidence limits. Please note that in the cases with $\sigma_e = 1.0$ and $1,000$ genes a small number of parameter estimations stopped (18 in table 4.2 $\rightarrow$ 982 valid, 17 in table 4.3 $\rightarrow$ 983 valid), because the algorithm reached the limits of its search space. In these cases, the model could not fit the generated data. This is neglected here, since it occurred very infrequently.

4.3.2. Evaluation of the Bayesian mutation detection methods

The Bayesian mutation detection method described in chapter 4.2.4 was evaluated with synthetic data. To this end, different assignments for the parameters σ_m, σ_e, and d were used for simulation. A major difference to the simulation approach to generate data described in chapter 4.2.2 is that the number of mutations was set to a fixed value k instead of using a Poisson process, for example, for $k = 1$ exactly 1 mutation was simulated in each gene. For each assignment $100,000$ genes were generated. Subsequently, the Bayesian mutation detection method was applied to each gene of each data set and the percentages of the correctly predicted numbers of mutations were calculated. Table 4.4 shows the results.

All differences in the data sets with no mutation ($k = 0$) were caused by non-mutational effects. The chance to detect a mutation in these data sets which would be a false positive is low, but it depends strongly on the parameters. For example, in the case $(0.5, 0.5, 1.0)$ the percentage to detect one mutation is 26.93 %, since thereby it is likely that a mutation occurred, because the expectation for the number of mutations is 1 ($d = 1$). Additionally, the non-mutational component overlays the mutational component which complicates a prediction. The other way round, false negatives occur (genes in which a given mutation is not detected) if the chance for a mutation is low, for example, in the case $(0.5, 0.5, 0.5)$. Thereby, 94.22 % of the genes were predicted as not mutated, although one mutation had occurred. If in the same case two mutations had occurred, still 89.89 % of the genes were predicted as not mutated. However, the percentage of correct predictions increases if the standard deviation of non-mutational effects decreases. Thus, in applications it depends on the parameter estimates $\hat{\sigma}_m$, $\hat{\sigma}_e$, and $\hat{d}$ which are summary statistics for all genes.

Table 4.4.: Results of the evaluation of the Bayesian mutation detection method on synthetic data. The first column shows the real number of mutations in the data. The next three columns show the distribution of percentages, how many mutations have been detected. The corresponding parameter values used by the mutation detection method are presented in the last column.

Number of	Predicted mutation class			Parameter
mutations k	0	1	≥2	(σ_m, σ_e, d)
0	96.20 %	3.80 %	0.00 %	$(0.5, 0.1, 0.5)$
	90.75 %	9.25 %	0.00 %	$(0.5, 0.1, 1.0)$
	97.96 %	2.04 %	0.00 %	$(0.5, 0.5, 0.5)$
	72.93 %	26.93 %	0.14 %	$(0.5, 0.5, 1.0)$
	97.98 %	2.02 %	0.00 %	$(1.0, 0.1, 0.5)$
	95.54 %	4.46 %	0.00 %	$(1.0, 0.1, 1.0)$
	94.50 %	5.50 %	0.00 %	$(1.0, 0.5, 0.5)$
	79.80 %	20.18 %	0.02 %	$(1.0, 0.5, 1.0)$
1	42.93 %	56.32 %	0.75 %	$(0.5, 0.1, 0.5)$
	35.01 %	61.01 %	3.98 %	$(0.5, 0.1, 1.0)$
	94.22 %	5.74 %	0.04 %	$(0.5, 0.5, 0.5)$
	63.43 %	35.63 %	0.94 %	$(0.5, 0.5, 1.0)$
	25.65 %	73.53 %	0.82 %	$(1.0, 0.1, 0.5)$
	22.11 %	73.76 %	4.14 %	$(1.0, 0.1, 1.0)$
	73.47 %	26.16 %	0.37 %	$(1.0, 0.5, 0.5)$
	54.25 %	42.83 %	2.92 %	$(1.0, 0.5, 1.0)$
2	31.64 %	62.93 %	5.43 %	$(0.5, 0.1, 0.5)$
	25.85 %	60.14 %	14.01 %	$(0.5, 0.1, 1.0)$
	89.89 %	9.87 %	0.24 %	$(0.5, 0.5, 0.5)$
	56.54 %	41.01 %	2.54 %	$(0.5, 0.5, 1.0)$
	18.31 %	75.55 %	6.14 %	$(1.0, 0.1, 0.5)$
	15.86 %	69.64 %	14.51 %	$(1.0, 0.1, 1.0)$
	61.46 %	35.86 %	2.68 %	$(1.0, 0.5, 0.5)$
	43.26 %	47.57 %	9.16 %	$(1.0, 0.5, 1.0)$

Table 4.5.: Overview about the primate (Khaitovich *et al.*, 2005a) and mice (Voolstra *et al.*, 2007) data sets.

Data set	Sample 1	Sample 2	#Genes
	#Individuals	#Individuals	
human/chimpanzee brain	6	5	15,526
human/chimpanzee heart	6	5	14,988
human/chimpanzee kidney	6	5	17,865
human/chimpanzee liver	6	5	15,046
human/chimpanzee testis	6	5	21,731
dom/mus brain	6	6	19,406
dom/mus liver/kidney	6	6	19,510
dom/mus testis	6	6	19,348
ssp/cas brain	6	3	19,406
ssp/cas liver/kidney	6	3	19,510
ssp/cas testis	6	3	19,348

4.3.3. Analysis of primate data

The M&E-normal and the M&E-extreme model were both applied to analyse the gene expression differences between human and chimpanzee (regarded as sample 1 and 2, respectively) in five different tissues: Brain, heart, kidney, liver, and testis. The data was collected at the Max-Planck-Institute for Evolutionary Anthropology in Leipzig with Affymetrix U133plus2 arrays and was published in Khaitovich *et al.* (2005a). Table 4.5 gives an overview about the used data sets. In case of the M&E-normal model, the ML parameter estimation method was used, since validation with synthetic data comprises superior results for the ML method than for the the χ^2-fit method. In case of the M&E-extreme model the χ^2-fit method was applied.

Since the samples 1 and 2 both consist of more than one individual, the data was treated as follows: Let sample 1 consists of n_1 individuals, let sample 2 consists of n_2 individuals, and let g be the number of genes. All gene expression differences from the $\eta = n_1 \cdot n_2 \cdot g$ pairwise comparisons between the samples over all genes were regarded as the observation from the data $O = (o_1, \ldots, o_\eta)$. Additionally, when using the M&E-extreme model with the χ^2-fit method, the mean values of $\hat{v}_{1,2}$ and $\hat{k}_{1,2}$ of all $\omega = n_1 \cdot n_2$ pairwise comparisons

Table 4.6.: ML parameter estimates from the real data sets (M&E-normal model). The 95 % confidence intervals from 1,000 bootstraps are shown in brackets.

Data set	$\hat{\sigma}_m$	$\hat{\sigma}_e$	$\hat{d}$
human/chimpanzee brain	0.737	0.205	0.330
	(0.686 , 0.807)	(0.188 , 0.255)	(0.273 , 0.387)
human/chimpanzee heart	0.955	0.322	0.339
	(0.817 , 1.169)	(0.266 , 0.402)	(0.188 , 0.492)
human/chimpanzee kidney	0.749	0.206	0.561
	(0.693 , 0.822)	(0.183 , 0.236)	(0.434 , 0.689)
human/chimpanzee liver	0.878	0.257	0.569
	(0.809 , 0.966)	(0.236 , 0.282)	(0.391 , 0.767)
human/chimpanzee testis	0.711	0.231	0.851
	(0.664 , 0.781)	(0.210 , 0.259)	(0.636 , 1.061)
dom/mus brain	0.515	0.125	0.177
	(0.418 , 0.634)	(0.109 , 0.144)	(0.119 , 0.268)
dom/mus liver/kidney	0.490	0.125	0.164
	(0.440 , 0.554)	(0.111 , 0.140)	(0.123 , 0.209)
dom/mus testis	0.440	0.118	0.233
	(0.374 , 0.555)	(0.099 , 0.143)	(0.120 , 0.329)
ssp/cas brain	0.498	0.165	0.343
	(0.450 , 0.708)	(0.137 , 0.274)	(0.136 , 0.463)
ssp/cas liver/kidney	0.460	0.129	0.319
	(0.426 , 0.524)	(0.114 , 0.153)	(0.180 , 0.457)
ssp/cas testis	0.362	0.097	0.346
	(0.310 , 0.404)	(0.077 , 0.120)	(0.200 , 0.475)

between the samples were calculated. These mean values were used to calculate $\chi^2(\sigma_e)$ (cf. chapter 4.2.2). The simulation of one gene was performed n_1 times for sample 1 and n_2 times for sample 2 accordingly which led to ω gene expression differences. Thus, the dependencies between the individuals were described corresponding to the data analysis. Please note that the overall number of simulations of gene expression differences per optimisation step was 10^7. If this number had been reached, simulations were stoped and the χ^2 was calculated. Additionally, 95 % confidence limits were constructed with bootstrapping (Efron, 1979) over the genes and the individuals for both types of models.

Table 4.7.: Parameters estimates with the χ^2-fit method from the real data sets (M&E-extreme model). The 95 % confidence intervals from 1,000 bootstraps are shown in brackets.

Data set	$\sqrt{\frac{\pi^2\hat{\beta}^2}{6}}$	$\hat{\sigma}_e$	$\hat{d}_1$	$\hat{d}_2$
human/chimpanzee brain	0.720	0.206	0.279	0.045
	(0.640 , 0.835)	(0.177 , 0.249)	(0.175 , 0.373)	(0.004 , 0.095)
human/chimpanzee heart	1.129	0.363	0.130	0.016
	(0.871 , 1.483)	(0.282 , 0.425)	(0.055 , 0.290)	(-0.011 , 0.095)
human/chimpanzee kidney	0.745	0.194	0.330	0.264
	(0.667 , 0.845)	(0.172 , 0.232)	(0.219 , 0.461)	(0.155 , 0.405)
human/chimpanzee liver	0.867	0.241	0.447	0.177
	(0.757 , 0.991)	(0.209 , 0.278)	(0.302 , 0.678)	(0.089 , 0.290)
human/chimpanzee testis	0.695	0.227	0.577	0.324
	(0.605 , 0.791)	(0.185 , 0.271)	(0.376 , 0.853)	(0.186 , 0.531)
dom/mus brain	0.705	0.132	0.074	0.024
	(0.554 , 0.841)	(0.115 , 0.149)	(0.052 , 0.113)	(0.010 , 0.048)
dom/mus liver/kidney	0.531	0.124	0.061	0.089
	(0.450 , 0.636)	(0.109 , 0.139)	(0.026 , 0.104)	(0.051 , 0.139)
dom/mus testis	0.795	0.136	0.015	0.051
	(0.596 , 0.972)	(0.113 , 0.154)	(0.002 , 0.048)	(0.032 , 0.092)
ssp/cas brain	0.557	0.168	0.122	0.158
	(0.441 , 0.823)	(0.131 , 0.271)	(0.011 , 0.195)	(0.051 , 0.261)
ssp/cas liver/kidney	0.567	0.135	0.105	0.109
	(0.467 , 0.661)	(0.115 , 0.168)	(0.048 , 0.167)	(0.052 , 0.192)
ssp/cas testis	0.584	0.114	0.047	0.076
	(0.419 , 0.766)	(0.092 , 0.135)	(0.013 , 0.122)	(0.032 , 0.168)

For each data set 1,000 bootstrap data sets were generated and the parameters were estimated on the resampled data. The results of the analysis with both types of models are presented in table 4.6 and 4.7, respectively.

When using the M&E-normal model, the largest estimate for the standard deviation in mutation effects $\hat{\sigma}_m$ can be observed in heart (0.955), while the smallest is in testis (0.711). The estimate for the standard deviation of non-mutational effects $\hat{\sigma}_e$ is also largest for heart (0.322). For brain and kidney $\hat{\sigma}_e$ is roughly 30% smaller (0.205 and 0.206, respectively). The greatest differences within the tissues can be observed for $\hat{d}$. For testis $\hat{d}$ is 0.851, while for brain $\hat{d}$ is merely 0.330. The result for heart is 0.339, while the d-estimates for kidney and liver are 0.561 and 0.569, respectively.

The results from the M&E-extreme model are in good agreement with the results from the M&E-normal model. Please note that the parameter β has a different meaning than the standard deviation σ_m for the normal distributed MED in the M&E-normal model. In order to make comparisons between the two models easier, the standard deviations $\sqrt{(\pi^2\beta^2)/6}$ of the extreme value distributed MEDs are presented in table 4.7. Again, heart comprises the largest result (1.129), while testis has the smallest estimate for the standard deviation in the mutation effects (0.695). For the non-mutational effects the largest estimate for the standard deviation $sigma_e$ is in heart likewise (0.363), whereas the smallest ones are in brain and kidney as well (0.206 and 0.194). When using the M&E-extreme model, the split of d in d_1 and d_2 is estimated: The results of $\hat{d}_1 + \hat{d}_2$ agree roughly with the results of $\hat{d}$ from the M&E-normal model. One exception is heart tissue. Thereby, the sum $\hat{d}_1+\hat{d}_2 = 0.146$ is considerably smaller than $\hat{d} = 0.339$ estimated with the M&E-normal model. It is noticeable that in all cases $\hat{d}_1 > \hat{d}_2$ which indicates an acceleration on the human lineage (for brain and liver even with non-overlapping confidence limits).

Subsequently, the data was applied to the Bayesian mutation detection method. Thereby, the estimates of the M&E-normal model were used. For each gene j within a data set all $\omega = n_1 \cdot n_2$ pairwise gene expression differences between sample 1 and 2 were regarded as the observation $O_j = (o_{1,j}, \ldots, o_{\omega,j})$. The results are presented in table 4.8. The percentages of genes which mutated in their regulatory regions correlate roughly with the estimates for d which means that the larger $\hat{d}$ the larger is the percentage of mutated genes. The largest number of mutated genes can be observed in testis (40.68 %), while in brain and heart only 13.73 % and 11.58 % genes, respectively, mutated during evolution.

4.3.4. Analysis of mice data

Additionally, the M&E-normal and the M&E-extreme models were applied to different mice data sets, collected with spotted arrays (OligoLibrary by Sigma-Genosys / Compugen spotted on Schott Nexterion Slides H) (Voolstra *et al.*, 2007). An overview of the data sets is given in table 4.5. In two approaches expression profiles of two different subspecies of *Mus musculus* were compared. In the first approach, differences between individuals of *Mus musculus domesticus* (dom, corresponding to sample 1) and *Mus musculus musculus* (mus, corresponding to sample 2) were examined. Free-living individuals of these species were captured in Czech Republic and kept individually under laboratory conditions. Subsequently, individuals of *Mus musculus ssp.* (ssp, regarded as sample 1) and *Mus musculus castaneus* (cas, regarded as sample 2) were analysed. Both subspecies had been kept between two and ten generations under laboratory conditions. The analysis was performed accordingly to the analysis of the primate data. The results for the M&E-normal model from the ML method are shown in table 4.6. The results for the M&E-extreme model estimated with the χ^2-fit method are presented in table 4.7.

When using the M&E-normal model, the estimates for the standard deviation of mutation effects σ_m are slightly greater in the comparison between dom and mus than between ssp and cas. Thereby, in brain $\hat{\sigma}_m$ is the largest (0.515 and 0.498, respectively), while in testis that parameter has the smallest estimates (0.440 and 0.362, respectively). The standard deviation estimate of non-mutational effects $\hat{\sigma}_e$ is greater in dom/mus testis (0.118) than in ssp/cas testis (0.097). For the remaining two tissues brain and liver/kidney the σ_e-estimates are greater in the ssp/cas- (0.165 and 0.129) than in the dom/mus comparison (both 0.125). It is noticeable that the d-estimates are smaller in dom/mus than in ssp/cas for all three tissues. Thereby, also testis has the largest $\hat{d}$ with 0.233 and 0.346, respectively. Liver/kidney shows the smallest d-estimates with 0.164 and 0.319, respectively.

When the M&E-extreme model is applied, the magnitude of the parameter estimates is similar. An exception is that the estimate for the standard deviation of the MED $\sqrt{(\pi^2\beta^2)/6}$ is smaller for liver/kidney in dom/mus than ssp/cas which is different from the other model. Furthermore, it is noticeable that dom/mus testis has the largest estimate for σ_e within the dom/mus tissues, while the same estimate is the second smallest of all mice sets when using the M&E-normal model. However, all estimates are in the same range. Larger differences can be observed for the d-estimates. While for the M&E-normal model testis shows the largest $\hat{d}$ of all tissues, for the M&E-extreme model the sum of $\hat{d}_1$ and $\hat{d}_2$ is the smallest for testis. Within the comparisons of the same subspecies, differences in the ratios of $\hat{d}_1$ and $\hat{d}_2$ can be observed. Dom shows a greater number of mutations than mus in brain (0.074 against 0.024), while mus shows a greater number of mutations than dom in liver/kidney and in testis (0.061 against 0.089 and 0.015 against 0.051, respectively). In the ssp/cas data sets, cas has a slightly greater number of mutations than ssp in all three tissues (brain: 0.158 against 0.122, liver/kidney: 0.109 against 0.105, testis: 0.076 against 0.047). Overall, in some comparisons the sum of $\hat{d}_1 + \hat{d}_2$ is considerably smaller than the d-estimate from the M&E-normal model.

Table 4.8.: Results of the Bayesian mutation detection method for the primate and mice data sets. The second column shows the number of genes in which no mutation was detected. The third column shows genes with at least one mutation.

Data set	No mutation	At least one mutation	%
human/chimpanzee brain	13,395	2,131	13.73
human/chimpanzee heart	13,252	1,736	11.58
human/chimpanzee kidney	12,422	5,443	30.47
human/chimpanzee liver	10,281	4,765	31.67
human/chimpanzee testis	12,890	8,841	40.68
dom/mus brain	17,945	1,461	7.53
dom/mus liver/kidney	18,326	1,184	6.07
dom/mus testis	17,479	1,869	9.66
ssp/cas brain	16,901	2,505	12.91
ssp/cas liver/kidney	16,484	3,026	15.51
ssp/cas testis	15,415	3,933	20.33

Finally, the parameter estimates from the M&E-normal model were used to apply the data to the Bayesian mutation detection method. The results are presented in table 4.8. Like in the primate data sets, the percentages of mutated genes correlates roughly with the d-estimates. For dom/mus the percentages are smaller in all three tissues in comparison to ssp/cas. Within each tissue the percentages for testis are greatest (dom/mus: 9.66 %, ssp/cas: 20.33 %). For dom/mus the smallest ratio of mutated genes can be observed in liver/kidney tissue (6.07 %), while for ssp/cas the smallest ratio can be observed in brain (12.91 %).

4.3.5. Comparison of the data fit of the different models

The M&E-normal and the M&E-extreme model were compared with the M model (Khaitovich *et al.*, 2005b) which has extreme value distributed mutation effects but no non-mutational effects. To this end, the χ^2-score was used as a criterion to measure the goodness of fit to the data. To compare the the models, data sets were applied which were originally used with the M model. These data sets are liver 95, brain 95, liver 133, brain 133 (cf. chapter 3 or Khaitovich *et al.* (2005b) for details). Parameter

Table 4.9.: Comparison between χ^2-scores (cf. Equation 4.13) of real data and simulated data as a measure of goodness (calculated with $1,000$ bins). The results of the M model were taken from Khaitovich *et al.* (2005b) (extreme value distributed mutational effects and no non-mutational effects). The results for the M&E-normal model were estimated with the ML method, while the results for the M&E-extreme model were estimated with the χ^2-fit method.

M model	$\hat{\beta}$		$\hat{d}_1$	$\hat{d}_2$	χ^2-score
Liver 95	0.38		0.83	0.45	0.70482
Brain 95	0.29		0.87	0.47	0.67673
Liver 133	0.44		1.07	0.58	0.54605
Brain 133	0.33		0.83	0.25	0.85037
M&E-normal	$\hat{\sigma}_m$	$\hat{\sigma}_e$	$\hat{d}$		χ^2-score
Liver 95	0.639	0.217	0.483		0.33651
Brain 95	0.507	0.194	0.435		0.30688
Liver 133	0.784	0.273	0.570		0.13484
Brain 133	0.612	0.197	0.345		0.23418
M&E-extreme	$\hat{\beta}$	$\hat{\sigma}_e$	$\hat{d}_1$	$\hat{d}_2$	χ^2-score
Liver 95	0.436	0.189	0.475	0.206	0.27968
Brain 95	0.385	0.179	0.308	0.099	0.28752
Liver 133	0.512	0.234	0.571	0.338	0.08525
Brain 133	0.425	0.192	0.337	0.065	0.24564

estimates for the M model were taken from Khaitovich *et al.* (2005b). Parameters of the M&E-normal model were estimated with the ML method, and parameters of the M&E-extreme model were estimated with the χ^2-fit method. The results are depicted in table 4.9.

Thereupon, one synthetic data sets with 10^7 genes was generated for each of the parameter sets in table 4.9 with the corresponding model. Accordingly, the χ^2-scores between the synthetic data sets and the original data sets were calculated with equation 4.15 (cf. last column of table 4.9). However, $1,000$ bins were used instead of 100. For instance, the distributions of gene expression differences for two of the data sets (liver133 and brain133) and for their corresponding synthetic data sets are illustrated in figure 4.2. Please note that the distributions of synthetic data generated with the M model contain

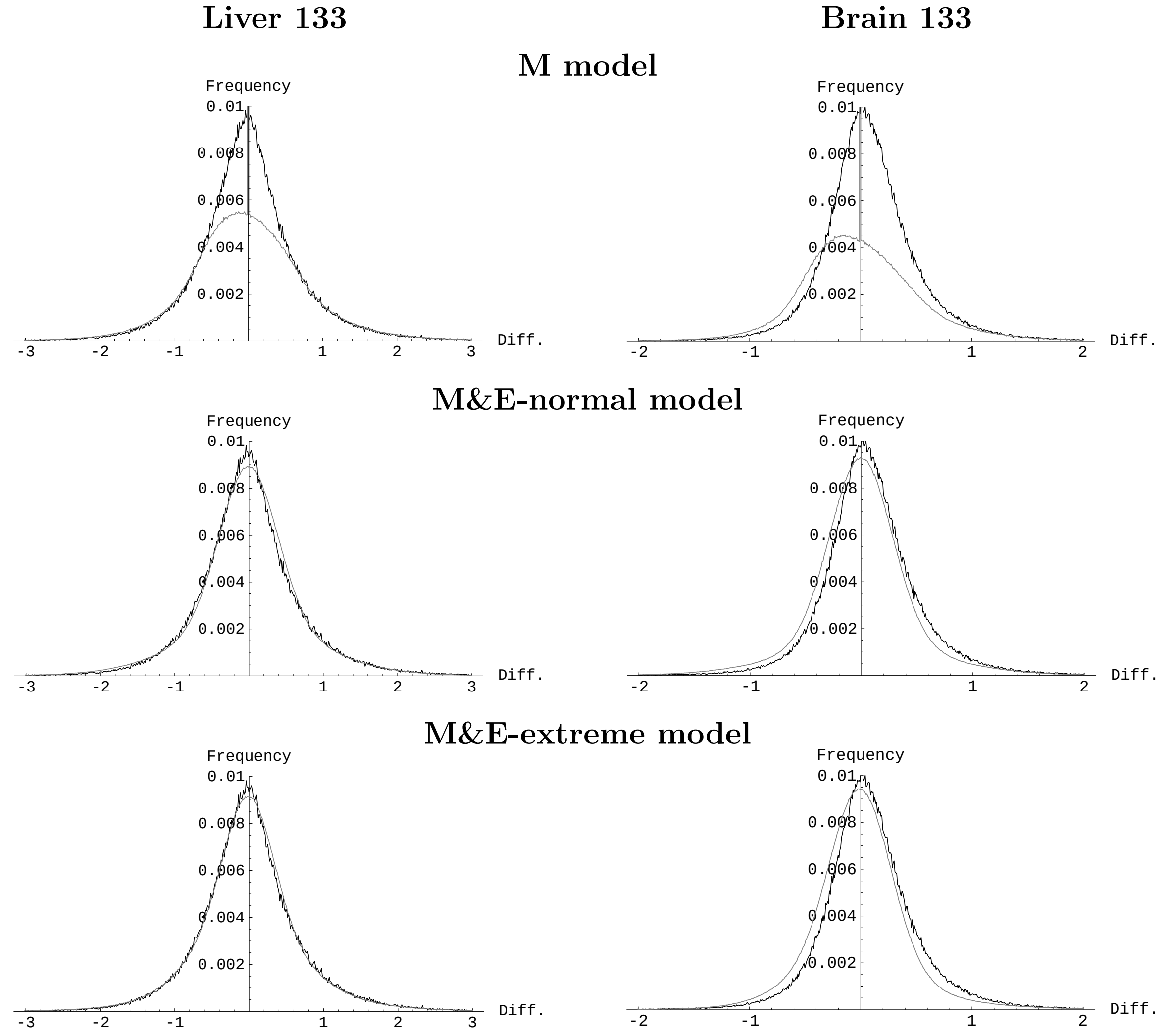


Figure 4.2.: Comparison between distributions of gene expression differences from real data (black) and distributions from simulations based on the corresponding parameter estimates (grey).

a peak at zero which results from those genes in which no mutations occurred during simulation. Therefore, their expression level did not changed (the M model does not contain non-mutational effects). The M&E model has smaller χ^2-scores than the basic model in all data sets. Skewed mutation effects were superior in three of four cases, since the χ^2-scores for the M&E-extreme model are smaller than for the M&E-normal model with one exception in brain 133. However, the difference in the χ^2-scores for brain 133 is small between the two different M&E models. Therefore, the fit of the M&E-normal model is only slightly better.

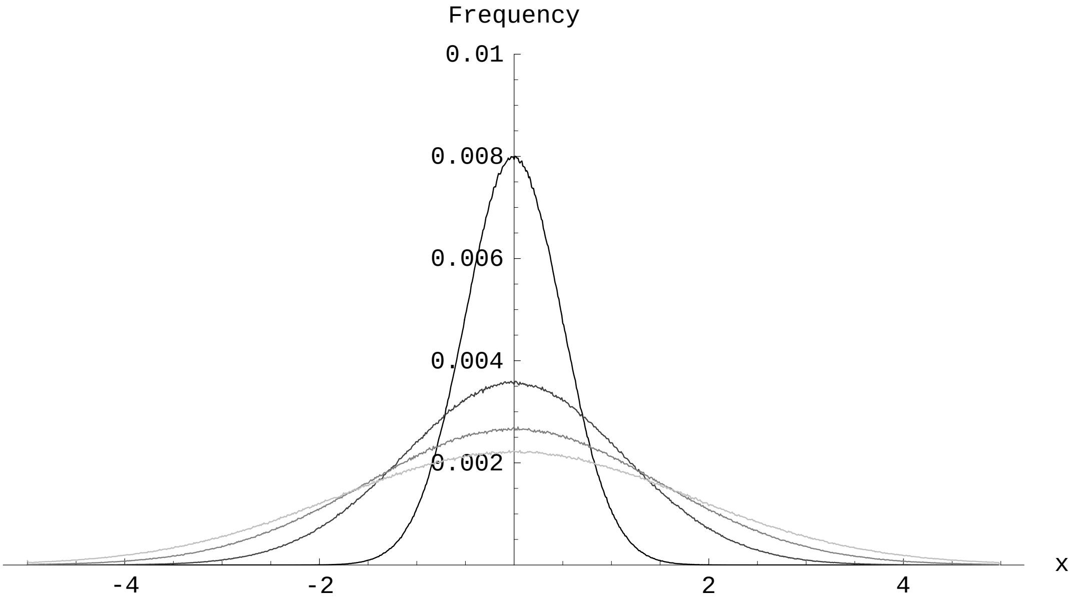


Figure 4.3.: Density functions of the gene expression difference x given a fixed number of mutations. The parameters which have been used are $\sigma_m = 1.0$ and $\sigma_e = 0.5$. The picture show $h(x|0)$ (black), $h(x|1)$ (dark grey), $h(x|2)$ (medium grey), and $h(x|3)$ (light grey).

4.4. Discussion

The data fit of the models extended by non-mutational effects is superior than the data fit of the M model. This was shown exemplarily for a number of data sets. To estimate the model parameters, two methods were presented. Therefrom, the ML method is substantially better on synthetic data sets, particularly for non-mutational effects with a large standard deviation in comparison to the standard deviation of mutation effects (cf. table 4.1). However, at present the ML method is only practical for the M&E-normal model which has a normal distributed MED. Normal distributed random variables have the advantage that their convolution can be calculated easily. In contrast, for the M&E-extreme model the χ^2-fit method is used, since the convolution of extreme value distributed random variables is difficult to calculate. Unfortunately, the use of the χ^2-fit method is problematic for a large standard deviation in non-mutational effects. However, in all applied biological data sets the estimated standard deviation of non-mutational effects is considerably smaller than standard deviation of mutation effects. Since estimates are relatively accurate in this case, one can trust in the results of the M&E-extreme model on the biological data sets (cf. table 4.2 and 4.3). Furthermore, the estimates of the M&E-normal and the M&E-extreme model show similar patterns which is another evidence that the χ^2-fit method works well for the applied data sets.

The better fit to data of the M&E-extreme model in comparison to the M&E-normal model in three of four cases (cf. table 4.9) supports the suggestion by Khaitovich *et al.* (2005b) that positively skewed mutation effects are superior than symmetric ones. Indeed, for brain 133 the M&E-normal model shows a slightly better fit to the data. However, the gene expression difference distribution of brain 133 seems to be slightly shifted to the right side (cf. figure 4.2). Thereby, a particular good fit was not observed with both MEDs (cf. figure 4.2). Thus, it is possible that the slightly superior result with a normal distributed MED might be due to the shift of the distribution.

In addition, a Bayesian method was used to detect mutations. However, the results on synthetic data sets are weak for some parameter assignments, since large percentages of false negative or false positive predictions were obtained. Indeed, this is caused by the nature of the gene expression difference distributions. To demonstrate the difficulties of the method, examples of different density functions of the gene expression difference given a fixed number of mutations are shown. Please note that the result of the mutation detection method also depends on the Poisson distribution with parameter d which describes the probability for the number of mutations (cf. equation 4.26). Figure 4.3 shows densities of gene expression difference distributions $h(x|k)$ with a fixed number of mutations k with $k \in \{0, 1, 2, 3\}$ and $\sigma_m = 1.0$ and $\sigma_e = 0.5$. A large part of the probability masses overlap between different distributions. For example, a gene expression difference between -0.71 and 0.71 does not indicate a mutation, since $h(x|0)$ has the highest frequency in this interval. Consequently, one would expect that the results of the mutation detection method have a large error. Admittedly, the robustness is increased with an increase of the number of individuals in the two samples which increases the number of pairwise gene expression differences. This could explain the correlation between d-estimates and percentages of mutated genes in the different data sets, because the data sets contain up to 36 differences per gene (dom/mus with $6 \cdot 6$ comparisons).

To detect the exact number of mutations is particularly problematic. If there is a gene expression difference between two samples caused by one mutation, the chance is great that this difference is reduced by a second mutation (because the mean of the MED is always zero). However, the chance to see that at least one mutation has occurred, increases with the overall number of mutations, since the variance of the gene expression difference is increased linearly to the number of mutations. Thus, the method is useful if one is interested to decide between genes with fixed expression levels and genes with changes in regulation between two samples.

The results of the analysis on the primate data sets are in good agreement with previous results by Khaitovich *et al.* (2005a). It has been reported that gene expression patterns differ less between human and chimpanzee in brain than in the other four examined tissues. This could be confirmed by the d-estimates of both types of the M&E model and the percentages of mutated genes from the Bayesian mutation detection method as well. According to this, only 13.73 % of genes changed their expression in brain (cf. table 4.8). Further, it was reported that the ratio of divergence between species to diversity within species is greatest in testis in comparison to the other tissues. This can be explained by directional selection during speciation Khaitovich *et al.* (2005a). While the within species diversity is not estimated here, the d-estimate for testis (0.851) is the largest of all tissues applied to the M&E-normal model. For the M&E-extreme model also the sum of $\hat{d}_1$ (0.577) (branch to human) and $\hat{d}_2$ (0.324) (branch to chimpanzee) is the largest in comparison to the other tissues. This is important, since the sum of $\hat{d}_1$ and $\hat{d}_2$ is similar to $\hat{d}$ estimated from the M&E-normal model for all primate sets, except for heart. However, this might be an outlier due to variation in the small data set. Estimates from the M&E-extreme model provide additional results: The ratios between $\hat{d}_1$ and $\hat{d}_2$ can indicate accelerated evolution on the branch to human in comparison to the branch to chimpanzee. Since, this ratio depends on the estimate for the coefficient of skewness $\hat{s}_{1,2}$, it is potentially not very robust. However, faster evolution on the human lineage can be observed for all tissues. The $\hat{d}_1/\hat{d}_2$ ratio which is expected with 1 in the neutral case is especially large for brain ($0.279/0.045 = 6.200$) and for heart ($0.130/0.016 = 8.125$). The result for brain agrees with the suggestion that gene regulation in brain evolved faster in human than in chimpanzee (Khaitovich *et al.*, 2005b). However, it will be necessary to apply larger data sets to the models to reinforce these findings.

The results on mice data show that the compared subspecies of mice (dom vs mus and ssp vs cas) have smaller divergence times than the primate species (< 1 million years). Altogether, the estimates for d and d_1/d_2, respectively, are smaller than in the primates. However, different array technologies and normalisation techniques were used. Thus, the results are difficult to compare. Overall, within the mice data sets the number of mutations are greater in the comparisons between ssp and cas than between dom and mus which indicates a closer connection between dom and mus than between ssp and cas. Unfortunately, it is unknown, when the subspecies split. The parameter estimates $\hat{d}_1+\hat{d}_2$

from the M&E-extreme model and $\hat{d}$ from the M&E-normal model are very different in some data sets which might be a result of outliers in the small data sets. Since ssp and cas were kept under laboratory conditions over several generations, it would be reasonable to observe non-mutational effects with a smaller standard deviation than for dom and mus. Indeed, it seems that actually the standard deviation of non-mutational effects in dom and mus is slightly smaller. Like in the primate sets, a larger sample size would be preferable. For mice it should be easy to keep all subspecies under exactly the same conditions which leads to more accurate results.

4.5. Conclusion

Both types of the M&E model give a better description of data taken from real microarray experiments than the M model, since they do not neglect the large influence of non-mutational effects. With the separation into mutational and non-mutational changes of gene expression a better fit to the data is achieved. Moreover, the additional parameter for the non-mutational effects enables a more detailed data analysis. This improves, for example, comparative studies between different tissues which are affected by a different amount of non-mutational effects. The analysis of primate data provides new evidence for different rates of gene expression evolution in different tissues. Results on mice data indicate a closer connection between *Mus musculus domesticus* and *Mus musculus musculus* than between *Mus musculus ssp* and *Mus musculus castaneus*. Under the assumption of the model one can trust in the results, since the estimated parameter values are in intervals for which simulation studies achieved good results. However, after incorporating the non-mutational effects, further research is necessary to get more precise models. An extension should address *trans*-effects on the gene expression level, since changes in the regulation of a transcription factor for many genes lead to wrong parameter estimates in the current models.

5. A Tajima-type test for gene expression data

A statistical test is introduced in this chapter in order to detect selection effects in gene expression data sampled from natural populations. The test depends on the comparison of two estimators for the population mutation rate. Therefore, it is similar to the Tajima's D test.

5.1. Introduction

A large amount of variability is typically maintained in natural populations. Some variants on sequence level lead to phenotypic differences which can change the fitness of individuals in positive or negative way. These variants are under selection. However, there is also large variability on the transcriptome level. Microarray technology (Baldi and Hatfield, 2002; Speed, 2003) provides insights into this variability among individuals in populations. Gene expression differences within populations have been examined in numerous studies (Cavalieri *et al.*, 2000; Enard *et al.*, 2002; Gibson *et al.*, 2004; Jin *et al.*, 2001; Morley *et al.*, 2004; Nuzhdin *et al.*, 2004; Oleksiak *et al.*, 2005; Schadt *et al.*, 2003; Storey *et al.*, 2007; Wayne *et al.*, 2004), cf. Ranz and Machado (2006) for a review. Since it is assumed that evolution in gene regulation plays an important role in phenotypic evolution (Wilson *et al.*, 1974; King and Wilson, 1975), it is important to determine which gene expression changes lead to variation are under selection.

In several studies two populations are compared to infer the mode of gene expression evolution. Thereby, mainly the ratios of within species variation and between species divergence are estimated and explained by neutral, directional or stabilising selection. Oleksiak *et al.* (2002) supposed adaptation to different life conditions for different teleost fish species which would be a result of directional selection. Rifkin *et al.* (2003) examined different species of *Drosophila* during early metamorphosis. They found evidence for stabilising selection as the main mode of evolution, but they also found genes which indicate directional selection or neutrality. Lemos *et al.* (2005) analysed different data sets of mice, *Drosophila*, and primates with an ANOVA based method (Kerr *et al.*, 2000) and concluded that the majority of genes evolve under stabilising selection. Analysis of different primate species indicate accelerated evolution in human brain in comparison to chimpanzee brain (Enard *et al.*, 2002; Khaitovich *et al.*, 2005b). Further, an overall up-regulation of gene expression in human brain in comparison to chimpanzee brain was observed (Caceres *et al.*, 2003; Hsieh *et al.*, 2003). These result might be due to directional selection (Khaitovich *et al.*, 2005b). Indeed, other effects might lead to the same results, for example, a relaxation of selective constraints or differential hybridisation due to loss of perfect sequence matching. However, results for liver tissue from the same primate species do not show acceleration or up-regulation in human which contradicts differential hybridisation. Khaitovich *et al.* (2004) observed a linear correlation between divergence time and sample difference for different primate species which can be explained by neutrality as the main mode of evolution. However, Gilad *et al.* (2006b) explained this correlation by effects of differential hybridisation depending on human-specific microarrays. They found evidence for stabilising selection in primates obtained with multispecies arrays. In Khaitovich *et al.* (2005a) patterns of gene expression were compared with differences in the corresponding DNA sequences. The results suggest also stabilising selection as an important factor. However, they suppose directional selection in primate testis.

Thus, these interpretations lead to discussions about the mode of evolution in gene regulation, for example, a recent review by Gilad *et al.* (2006a) concludes that stabilising selection is likely to be the dominant mode of gene expression evolution. However, a review by Khaitovich *et al.* (2006) postulates a neutral model as a useful null hypotheses, since the neutral theory by (Kimura, 1983) does not eliminate the role of selection. A large number of mutations might be deleterious if the corresponding alleles are under strong stabilising selection. Indeed, deleterious mutants disappear from the population over the years. Thus, one would not expect large differences in present days individuals. This would explain the relatively small expression differences for many genes. However, Khaitovich *et al.* (2005b) postulated that a small fraction of genes which evolved under directional selection might cause patterns observed in human brain. This would also not contradict the neutral theory, since this theory alludes to the majority of all genes. Thus, here it is assumed that the majority of gene expression differences which are observed can be explained by neutral or nearly neutral evolution. Mutations occur according to a molecular clock so that the variation in a population can give information about selection.

In this chapter within-species variation on gene expression level is analysed to detect selection in natural populations. On DNA level the variation in sequences is conveniently summarised by the population mutation rate $\theta = 2N\mu$, whereas N being the number of (diploid) individuals in the population (or $2N$ haploid individuals instead), and μ being the mutation rate per sequence and per generation. To estimate θ from n individuals sampled from a population, various sequenced based summary statistics have been suggested. Well known are (1) the *average pairwise distance* $\hat{\pi}$ (Tajima, 1983) which depends on the frequency of variants, (2) the *Watterson's estimator* $\hat{\theta}_W$ calculated by the *number of segregating sites* (Watterson, 1975), and (3) the number of alleles (Ewens, 1972) ((2) and (3) are independent of the frequency of a variant in the population) (cf. chapter 2.3.4).

Tajima (1989) suggested a test to detect selection in a sample of n sequences which is based by and large on the normalised difference between $\hat{\pi}$ and $\hat{\theta}_W$. The n sequences in the sample have been randomly drawn from a population. Thereby, a Wright-Fisher population is assumed (cf. chapter 2.3.3 or Hein *et al.* (2005)) which is used as a null model. It assumes discrete and non-overlapping generations, haploid individuals, constant population size, equal fit of all individuals (no selection), no geographical or social structure and no recombination. The genealogy of a Wright-Fisher population can be described by the coalescent process (Hudson, 1991). The test statistic D of the Tajima's D test is given by

$$D = \frac{\hat{\pi} - \hat{\theta}_W}{\sqrt{Var(\hat{\pi} - \hat{\theta}_W)}} \tag{5.1}$$

which is zero in expectation. The distribution of D resembles a beta distribution which is used to calculate confidence limits. If the test is applied to real data, the Wright-Fisher model is falsified if D which is estimated from the data is outside the confidence limits. If D is smaller than the lower limit, it can be explained by directional selection. If so, the estimator from segregating sites $\hat{\theta}_W$ exceeds the frequency dependent mean pairwise distance $\hat{\pi}$ which is not affected much by slightly deleterious mutants in the population. In the case that D is greater than the upper limit, balancing selection is an explanation, because in this case one would assume a small number of different alleles which are fixed in the population. Then the mean pairwise distance $\hat{\pi}$ is greater than $\hat{\theta}_W$ estimated from the number of segregating sites. However, there are other effects like bottlenecks or population subdivisions which can influence D (Hein *et al.*, 2005).

While analysis of population history is well understood if sequences are used, the analysis is still in its infancy for gene expression data. In this study two estimators for θ from gene expression variation in natural populations are presented. A model is used to correct the estimates for noise caused by non-mutational effects. Depending on the corrected estimates, a Tajima-type test to detect selection is suggested. The applicability of the test is illustrated with synthetic data and finally, a biological example is discussed.

5.2. Materials and methods

5.2.1. The evolution model

The Tajima's D test is based on the discrete nature of sequence evolution. Since gene expression is a continuous trait, a gene expression evolution model (Khaitovich *et al.*, 2004) is mapped on the genealogy of individuals in a population. Since non-mutational effects overlay gene expression changes caused by mutations, an extension of the basic model is applied here which takes non-mutational effects into account (cf. chapter 4). It is assumed that the mutation effect distribution (MED) is a normal distribution with standard deviation σ_m, and the non-mutational effect distribution (N-MED) is a normal distribution with standard deviation σ_e (M&E-normal model). The M&E-normal model can be easily superimposed on genealogies relating n individuals from a Wright-Fisher population which is described in algorithm 5.2. For each iteration of the main loop of that algorithm, gene expression values of one gene in each individual are generated together with the underlying genealogy. The output file contains the synthetic microarray data.

5.2.2. Estimators for θ

In the following two approaches are described to estimate the population mutation rate θ from gene expression data for n individuals from a Wright-Fisher population.

Algorithm 5.2: Create synthetic data along genealogies

```
Data: Sample size $n$, number of genes $g$, model parameters $(\sigma_m, \sigma_e, \theta)$
Result: Synthetic data set described by matrix $S$ with $n$ individuals and $g$ genes
Initialise matrix $S$ with the dimensions $n \times g$;
Initialise global counter variables $i$ and $j$;
for $j := 1$ to $g$ do
    Generate a genealogy $G$ of size $n$ with mutation rate $\theta$ by the coalescent process
    (Hudson, 1991);
    Let $i := 1$;
    Start DepthFirstSearch(0) in the root of $G$;
end
```

Procedure DepthFirstSearch(x)

```
Data: Gene expression value $x$
if current root is a leaf then
    Draw the non-mutational effect from the N-MED with standard deviation
    $\sigma_e$, add it to the gene expression value $x$, and store the result in $S(i, j)$;
    $i := i + 1$;
else
    if current root has a left child then
        Let $l$ be the length of the edge to the left child;
        Draw the number of mutations on this edge from a Poisson distribution
        with parameter $l\theta/2$;
        For each mutation draw a mutation effect from the MED with standard
        deviation $\sigma_m$ and let $x'$ be the sum of these mutation effects;
        Start DepthFirstSearch($x + x'$) in the left child of $G$;
    end
    if current root has a right child then
        Let $l$ be the length of the edge to the right child;
        Draw the number of mutations on this edge from a Poisson distribution
        with parameter $l\theta/2$;
        For each mutation draw a mutation effect from the MED with standard
        deviation $\sigma_m$ and let $x'$ be the sum of these mutation effects;
        Start DepthFirstSearch($x + x'$) in the right child of $G$;
    end
end
```

$\hat{\theta}_{\textsf{var}}$ from the variance of gene expression: The variance of the expression level of a gene which evolves according to the gene expression evolution model correlates linearly with time (Khaitovich *et al.*, 2005b). Time is measured here in units of $2N$ generations. Thus, the variance correlates linearly with θ. Since the gene expression evolves along genealogies, the variance of expression is $(\sigma_m^2\theta)/2$. The division by 2 is necessary, since $l\theta/2$ mutations occur on each edge of a genealogy, whereas l is the length of the edge (cf. algorithm 5.2). Additionally, each gene expression value is influenced by non-mutational effects. Hence, the variance of the expression values of a gene is

$$Var(\text{Gene expression}) = \sigma_m^2 \frac{\theta}{2} + \sigma_e^2 \tag{5.2}$$

After transformation the following estimate $\hat{\theta}_{\text{var}}$ is derived from the gene expression variance:

$$\hat{\theta}_{\text{var}} = \frac{2 \cdot (Var(\text{Gene expression}) - \sigma_e^2))}{\sigma_m^2} \tag{5.3}$$

The estimate $\hat{\theta}_{\text{var}}$ is similar to the mean pairwise distance $\hat{\pi}$ estimated from sequences. It is linearly to time except for the correction of non-mutational effects.

$\hat{\theta}_{\textsf{alleles}}$ from the number of alleles: The Tajima's D test uses the number of segregating sites to derive the second estimate $\hat{\theta}_W$. This is not possible here, since microarray data comprises only a single expression value per gene and per individual. However, an infinite alleles model can be assumed (Ewens, 1972), since one can count different gene expression values. Under the assumption of an idealised situation without any non-mutational effects, all differences between gene expression values would result from mutations. Thus, it can be assumed that two individuals carry two different alleles of the analysed gene if their expression level in the gene is different. Under this assumption it is easy to count the number of alleles a. Hence, the estimate $\hat{\theta}_{\text{alleles}}$ can be derived by the following equation (Ewens, 1972):

$$a = \sum_{i=0}^{n-1} \frac{\hat{\theta}_{\text{alleles}}}{\hat{\theta}_{\text{alleles}} + i} \tag{5.4}$$

There is no closed form for equation 5.4, but it can be solved with a root find method (cf. chapter 2.4 or Brent (1972)). Please note that the maximum value for a is restricted

Algorithm 5.3: Count the number of alleles

```
Data: Expression values of a gene from $n$ different individuals, model parameters
      $\sigma_m, \sigma_e$
Result: The number of alleles $a$
Let $a := 0$;
Let $distanceToLastAllele := 0$;
Sort the $n$ individuals according to their expression level in the analysed gene. Let
$x_i$ be the gene expression level in individual $i$ after sorting so that
$x_1 \leq \ldots \leq x_i \leq \cdots \leq x_n$;
for $i := 1$ to $n - 1$ do
    $distanceToLastValue := x_{i+1} - x_i$;
    $distanceToLastAllele := distanceToLastAllele + distanceToLastValue$;
    if $distanceToLastValue > 2\sigma_e$ OR $distanceToLastAllele > 2\sigma_m$ then
        $a := a + 1$;
        $distanceToLastAllele := 0$;
    end
end
```

to $n - 1$, because the solution for $a = n$ is infinite. This has no effect for large sample sizes, since it is unlikely to count as much alleles as individuals except for cases in which θ is extremely large. Unfortunately, real data is not ideal which is caused by non-mutational effects. However, the estimates for the standard deviations of mutational and non-mutational effects can be used to develop a heuristic implemented in algorithm 5.3. The algorithm counts a new allele if the distance between two gene expression values is greater than twice of the standard deviation of non-mutational effects. A new allele is also counted if the difference of the expression value of the current individual and the expression level of the first individual of the current allele is greater than twice of the standard deviation of mutation effects. The algorithm was tested on synthetic data and percentages of correct predicted number of alleles of about 80 % to 90 % were observed. However, the percentages depend on the parameters and the sample size (not shown here).

5.2.3. The Tajima-type test

Following Tajima's D-test (Tajima, 1989) a neutrality test based on the following statistic is suggested:

$$\Delta = \hat{\theta}_{\text{var}} - \hat{\theta}_{\text{alleles}} \tag{5.5}$$

Unfortunately, the complex nature of the model does not allow an analytical treatment of the test statistic to estimate the confidence limits. Hence, algorithm 5.2 is used to create genealogies which depend on σ_m, σ_e and θ estimated from the data under the assumption that the majority of gene expression levels evolved neutral. The resulting distribution of Δ-values describes the assumed neutral case and is used to calculate the confidence limits. According to the Tajima's D-test, genes from the real data which have a Δ-value smaller than the lower limit might be under directional selection, whereas genes with a Δ-value greater than the upper limit might be under balancing selection. However, as mentioned before, other effects might influence the Δ-value of a gene.

5.3. Experiments and results

5.3.1. Analysing the ratios of the two θ-estimators

In order to show the ratios between the two estimators for θ, simulations with different sample sizes n and mutation rates θ were carried out. Four cases were tested: (a) $n = 10, \theta = 0.1$, (b) $n = 100, \theta = 0.1$, (c) $n = 10, \theta = 1$, (d) $n = 100, \theta = 1$. Please note that the choice of σ_m and σ_e is nonrelevant, since both estimators for θ use these parameters to correct the estimates again. For each of the cases (a)–(d) $10,000$ genealogies were created, and on each genealogy evolution of gene expression was simulated $1,000$ times. The mean variances and allele-numbers of the $1,000$ simulations were used to determine the estimates $\hat{\theta}_{\text{var}}$ and $\hat{\theta}_{\text{alleles}}$. Subsequently, scatter plots were used to illustrate the estimates $\hat{\theta}_{\text{var}}$ and $\hat{\theta}_{\text{alleles}}$ for the $10,000$ genealogies (cf. figure 5.1). The θ_{var}-estimates were mapped to the x-axis, while the θ_{alleles}-estimates were mapped to the y-axis.

With an increase of the mutation rate, a corresponding linear increase of the θ-estimates can be observed (cf. the scaling of the axis in figure 5.1). With an increase of the

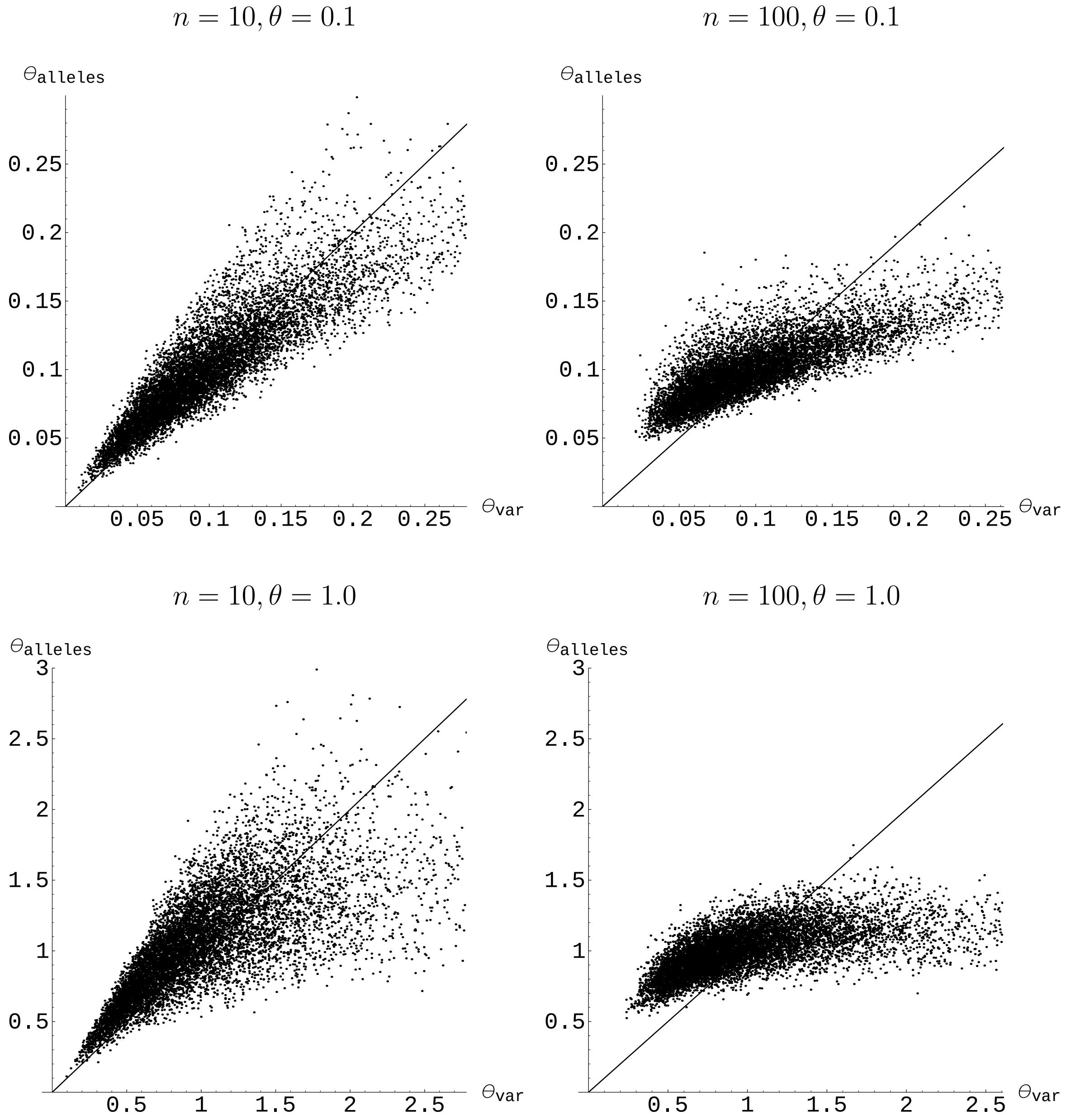


Figure 5.1.: $\theta_{\text{var}}/\theta_{\text{alleles}}$-plots from 10,000 genealogies for four cases. Each point results from the mean values of estimates from 1,000 simulations per genealogy.

sample size n, the variance in the estimates is reduced, since the scatter plots are more compact for $n = 100$ than for $n = 10$. This decrease of the variance is stronger in the θ_{alleles}-estimates than in the θ_{var}-estimates which results in a curve in the scatter plots (cf. cases with $n = 100$ in figure 5.1).

Table 5.1.: Mean values and empirical 95 % confidence intervals of the Δ statistics from simulations on $100,000$ Wright-Fisher genealogies.

n	$\theta = 0.1$	$\theta = 0.5$	$\theta = 1.0$
5	-0.048	-0.280	-0.559
	(-0.685 , 0.393)	(-2.195 , 2.043)	(-6.354 , 3.909)
10	-0.021	-0.102	-0.219
	(-0.429 , 0.550)	(-1.653 , 2.152)	(-2.905 , 3.836)
50	-0.008	-0.026	-0.051
	(-0.404 , 0.651)	(-1.030 , 2.212)	(-1.717 , 3.797)
100	-0.005	-0.027	-0.029
	(-0.381 , 0.686)	(-0.925 , 2.171)	(-1.554 , 3.795)

5.3.2. Analysing the distribution of Δ-values

After comparing the distributions of the two different estimators for θ, the distribution of their difference $\Delta = \theta_{\text{var}} - \theta_{\text{alleles}}$ was considered. Thereby, a larger number of different parameter values were used to determine the distributions of Δ in a larger number of cases. Table 5.1 displays parameter settings for the mutation rates ($\theta = 0.1, 0.5, 1.0$) and sample sizes ($n = 5, 10, 50, 100$). Algorithm 5.2 was used to generate the corresponding distributions of Δ-values. Mean values and 95 % confidence intervals are presented in table 5.1. Here, the results base on simulated $100,000$ genealogies. The range of the confidence intervals correlates positively with θ. The mean estimates for θ are underestimated in all cases, since the expectation $E[\Delta]$ is zero. However, if the sample size n grows, Δ tends to zero. Beside, the ranges of the confidence intervals are decreased and shifted if n is increased. However, the differences of means values and confidence limits between the cases with $n = 50$ and $n = 100$ are marginal in comparison to the differences between $n = 5$ and $n = 10$. Thus, one would not await large changes in the results if n is set to a value substantially greater than 100.

Four of the distributions described by their mean values and confidence limits in table 5.1 were illustrated in figure 5.2. As already discussed, the distributions are more wide in case of a larger θ-value. In addition, the distributions are affected by the sample size n which can be observed especially for the cases with $\theta = 1.0$. Thereby, the distribution for $n = 10$ shows several peaks. In contrast, the peaks disappear nearly completely for

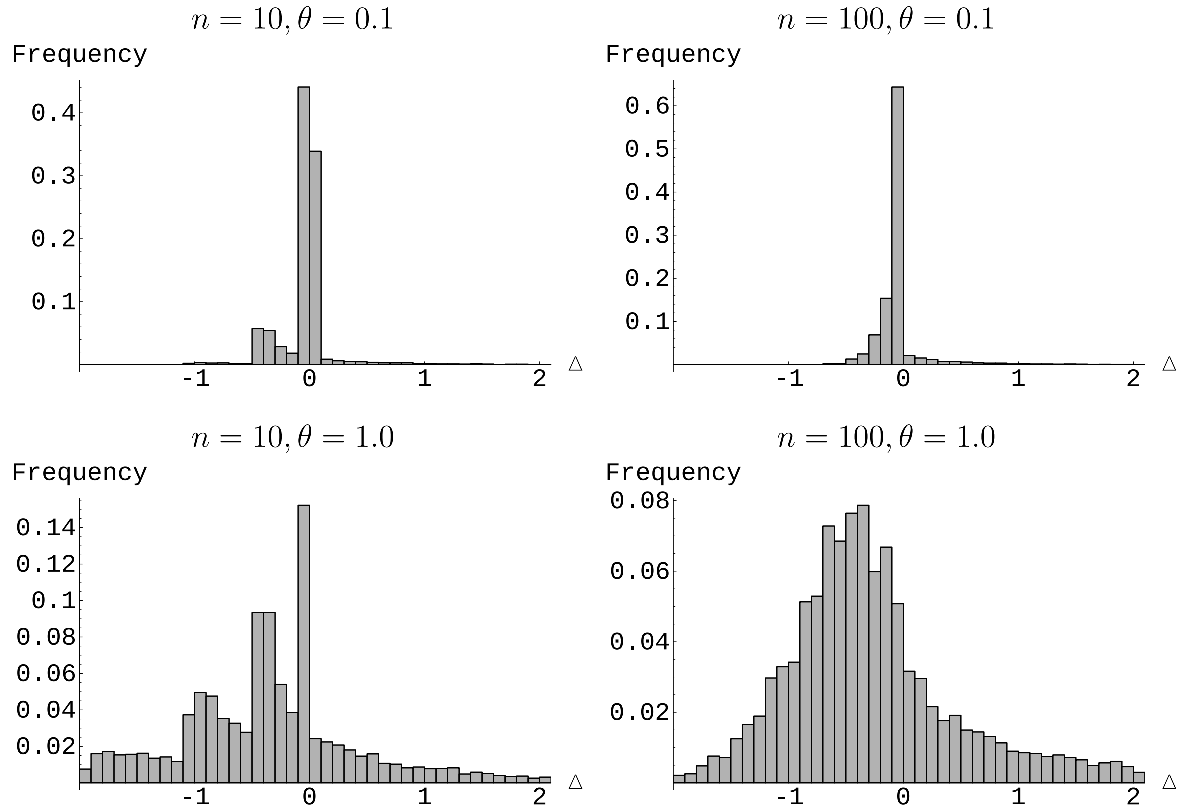


Figure 5.2.: Δ-distributions for four different cases.

Table 5.2.: Results from the analysis with model with mutational and non-mutational effects.

Data set	#Individuals	#Genes	$\hat{\sigma}_m$	$\hat{\sigma}_e$	$\hat{\theta}$
Brain	6	15,104	0.519	0.167	0.221
Heart	6	14,582	0.665	0.234	0.264
Kidney	6	17,387	0.594	0.198	0.492
Liver	6	14,668	0.622	0.195	0.426
Testis	6	21,164	0.367	0.133	0.249

$n = 100$ (cf. figure 5.2).

Table 5.3.: Results of the Tajima-type test. The second and the third column represent the 95 % confidence limits, Δ_{Mean} is the mean Δ-values of the real data, and the last two columns show the number of genes falsifying the null hypothesis at the lower (#Genes <) and the upper limit (#Genes >), respectively.

Data set	Lower limit	Upper limit	Δ_{Mean}	#Genes <	#Genes >
Brain	-1.359	1.215	0.024	39 (0.25 %)	214 (1.38 %)
Heart	-1.531	1.423	0.038	42 (0.28 %)	185 (1.23 %)
Kidney	-2.224	2.226	0.117	60 (0.34 %)	277 (1.55 %)
Liver	-1.780	1.990	0.106	94 (0.62 %)	272 (1.81 %)
Testis	-1.479	1.371	0.087	15 (0.07 %)	359 (1.65 %)

5.3.3. Human data

The Tajima-type test was applied to expression profiles from five different tissues of human: Brain, heart, kidney, liver, and testis (Khaitovich *et al.*, 2005a). The data sets were derived from six individuals. Sex-related genes had been filtered out before the analysis to avoid sex-bias. Subsequently, the ML method (cf. chapter 4.2.3) was used to estimate $\hat{\sigma}_m$, $\hat{\sigma}_e$, and $\hat{\theta}$ for all $n(n-1)/2$ pairs of individuals in a sample. The mean values of the respective $n(n-1)/2$ estimates were calculated and used as the final results. They are shown in table 5.2. The smallest estimate for the standard deviation in mutation effects σ_m is shown for testis (0.367), whereas this estimate varies for the other four tissues between 0.519 and 0.665. Testis also shows the smallest standard deviation in non-mutational effects σ_e (0.133). The second smallest estimate for σ_e is in brain (0.167), whereas the largest estimate for σ_e is in heart (0.234). The estimates for the population mutation rate θ are similar in brain (0.221), heart (0.264), and testis (0.249), while they are larger in kidney (0.492) and liver (0.426).

Subsequently, the parameter estimates were used to perform the Tajima-type test. The mutation rates $\hat{\theta}_{\text{var}}$ and $\hat{\theta}_{\text{alleles}}$ were estimated for each gene in the five data sets and the distributions of Δ-values were calculated. Furthermore, distributions were generated under neutrality with algorithm 5.2 depending on the corresponding parameter estimates in order to obtained the lower and the upper 95 % confidence limits. The confidence limits and the number of genes located outside of these limits are presented in table 5.3 for each of the data sets. Figure 5.3 displays the corresponding distributions of Δ-values.

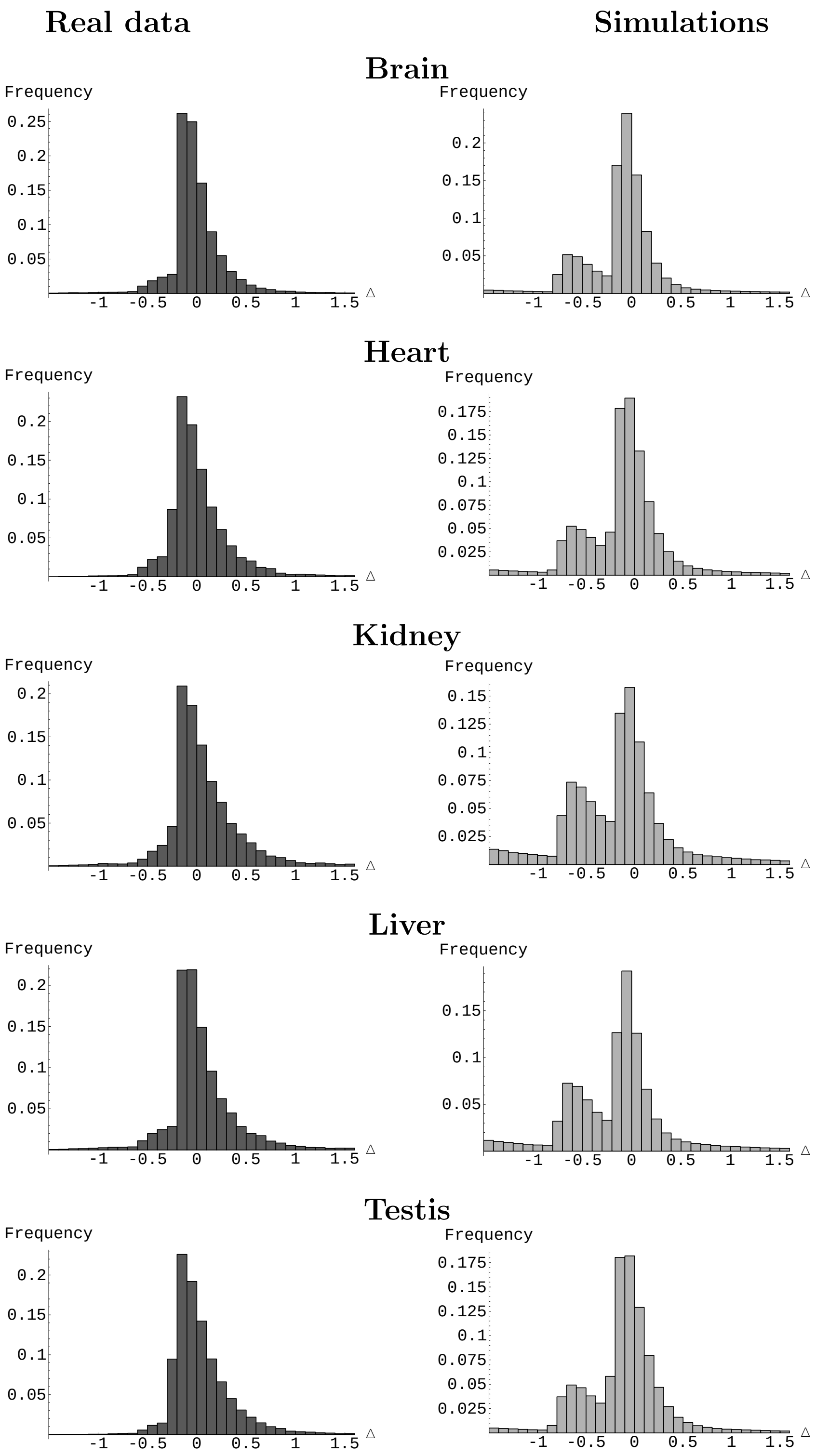


Figure 5.3.: Δ-distributions for the different tissues for the real data and simulations using the estimated parameters from the real data.

The range of the confidence limits correlates positively with θ. Thus, the largest range can be observed in kidney $(-2.224, 2.226)$, while the smallest is in brain $(-1.359, 1.215)$. In all five data sets the number of genes whose Δ-estimates are greater than the upper confidence limit exceeds the number of genes whose Δ-estimates are smaller than the lower confidence limit by far. However, the number of genes which are significantly smaller than the lower limit or greater than the upper limit is smaller than one would expect in all five cases: Based on a 95 % confidence interval, one would await about 5 % of genes in each tissue in which the neutrality hypothesis is rejected. All distributions of Δ-values from real data show a peak around $\Delta = 0$ (cf. figure 5.3). However, their mean values are slightly greater than zero (cf. the column Δ_{Mean} in table 5.3). This differs from the mean of the Δ-values describing the neutral case which is smaller than zero (cf. test cases in table 5.1).

5.4. Discussion

For DNA sequences different estimators are widely known to evaluate the population mutation rate θ from samples of natural populations. Tajima (1989) used the difference of two estimators to decide whether observed sequences evolved neutrally or not. Here, the estimators were first adapted to the transcriptome level, to analyse the variability in gene expression within populations. Thereby, it is assumed that variability in gene expression depends on mutations in regulatory regions which change the level of transcript abundance (Khaitovich *et al.*, 2005b). Considered as a mathematical problem, the major difference between sequences and gene expression levels is that differences between sequences are discrete, and differences between gene expressions levels are continuous. Therefore, the discrete mutation model which describes mutations on genealogies of sequences was replaced by a continuous gene expression evolution model.

The first estimator for the population mutation rate under the continuous model is $\hat{\theta}_{\text{var}}$. This estimator is based on the linear correlation of gene expression variance and time. It can be considered as an adaptation of the average pairwise distance $\hat{\pi}$ used for sequences. Both, $\hat{\theta}_{\text{var}}$ and $\hat{\pi}$, grow linearly in time under a neutral model and both depend on the frequency of the variants in the population. A relative small number of sequences from a new allele in a population would not change $\hat{\pi}$ much. Likewise, a few individuals with a different level of gene expression would not change the variance of gene expression much.

The second estimator $\hat{\theta}_{\text{alleles}}$ used here is not frequency dependent which is related to the estimator taken from the number of alleles (Ewens, 1972). On the level of sequences and on the level of gene expression as well, already one mutated individual changes the estimate in the same way as many other mutants of the same type. This is equal to the estimator from the number of segregating sites used by Tajima (1989).

Thus, the test statistic $\hat{\theta}_{\text{var}} - \hat{\theta}_{\text{alleles}}$ has the same meaning as the Tajima's D test statistic. Assume a population in which the appearance of different alleles of a gene is advantageous. In such a situation of balancing selection the estimator $\hat{\theta}_{\text{var}}$ is assumed to be greater than $\hat{\theta}_{\text{alleles}}$. Under directional selection a single allele is preferred so that the allele frequencies shift in one direction over the time. Thus, $\hat{\theta}_{\text{alleles}}$ should exceed $\hat{\theta}_{\text{var}}$. Indeed, corresponding to the Tajima's D test, other effects like bottlenecks in the population size or linkages of neutral sites to selective sites can influence the distribution of Δ.

A problem is that non-mutational effects exacerbate the counting of different alleles in gene expression data. Thus, here the M&E-normal model was applied which considers non-mutational effects (cf. chapter 4). Admittedly, the estimate of the standard deviation of non-mutational effects $\hat{\sigma}_e$ is a summary over all genes. A less simple consideration would be beneficial, since non-mutational impact differs between different genes. However, if one takes this into account, a larger sample size would be necessary to obtain faithful results. Relating to the data currently available, this would not be practical. An advantage of the M&E model is its additivity. Thus, the variance of non-mutational effects can be subtracted from the overall variance of gene expression.

In chapter 5.3 simulation studies were performed to examine the distributions of the two θ-estimators and their difference Δ. It was shown that Δ is smaller than zero for small values of the population size n. This deviation is reduced if n is increased. Furthermore, the asymmetric distribution of Δ shows conspicuous peaks for small n (cf. the case with $n = 10$ and $\theta = 1.0$ in figure 5.2). For larger n the peaks do not appear in the distributions. A reason for these peaks is that the method of moments to estimate $\hat{\theta}_{\text{alleles}}$ is biased under an infinite alleles model (Joyce, 1995). This bias depends on the discrete nature of the number of alleles. If n is increased, the number of possible states to describe the diversity by a is increased. Then the resulting estimate $\hat{\theta}_{\text{alleles}}$ is more accurate which results in a Δ-value closer to the expectation of zero. By the reason that the number of possible states to describe diversity is increased, the peaks in the distribution of Δ disappear (cf. the case with $n = 100$ and $\theta = 1.0$ in figure 5.2).

In this study, simulations were performed for various choices of the parameter values. In the test by Tajima (1989) the beta distribution was used to approximate the distribution of D. Depending on the speed of present time computers, it is unproblematic to simulate the distribution for the neutral case for each situation. The advantage of this approach is that the discussed problem of bias resulting from small n is included.

All simulations were performed to examine the neutral case. It would be favourable to observe the distribution of Δ-values under a non-neutral model to see the differences between the neutral case and a case with directional or balancing selection. However, this causes difficulties so that further research is necessary. The main problem is that under selection not only the genealogies differ from the neutral case but also the level of gene expression itself. With the current model it is not possible to describe both. Further, one has to decide how strong influences of selection are. However, even without regarding the non-neutral case explicitly, the simulations from the neutral case can be used to decide which genes from real data do not correspond to the expectations of the neutral theory.

An analysis of microarray data from different human tissues found only a very small number of genes which deviate significantly from a neutral model. Its number is below the expectation for the 5 % confidence limits. Furthermore, the number of genes which are greater than the upper limit exceeds the number of genes which are smaller than the lower limit by far. Since this picture is consistent in all five data sets, one suspects that a general problem exists. One can argue that the small number of individuals (6 in each set) is problematic for this kind of analysis, for example, because of the biased estimator $\hat{\theta}_{\text{alleles}}$. Thus, in further studies one should use data sets with a larger sample size. Additionally, a comparison between the genes of different tissues which rejected the neutral model might lead to interesting results. Because of the small n, this kind of analysis was abandoned here.

However, the results on the real data show some interesting aspects, even if the sample size is small. The parameter estimates for the M&E-normal model signify a small influence of non-mutational effects on the level of gene expression in testis in comparison to the other tissues. Also for brain, these influences are relatively small. One can argue that these two tissues are more insulated from environmental effects than heart, kidney, and liver. Gene expression in heart is assumed to change according to the physical situation of an individual, while gene expression in kidney and liver is affected by numerous metabolic processes. The standard deviation estimate of mutation effects σ_m

is comparatively small in testis. Together with the relatively small θ-estimate (0.249), it reflects observations from other studies. In these studies small changes in testis within primates species (Khaitovich *et al.*, 2005a) and within mice subspecies (Voolstra *et al.*, 2007) in comparison to other tissues were observed. The θ-estimates differ between 0.221 (brain) and 0.492 (kidney). This indicates a relatively large number of gene expression differences caused by mutations among the individuals, since it describes the expected number of mutations per gene between two randomly chosen individuals from the population. This large number of differences agrees with results by Storey *et al.* (2007). They obtained in a different human data set that even 83 % of all genes show differential gene expression among individuals which evinces large variation.

Although, the sample size of the data sets is very small for a population study, the results of the Tajima-type test are in good agreement with other recent studies: The gene expression of the majority of genes in the human data evolved according to a neutral model as suggested by Khaitovich *et al.* (2004). However, the results can be explained by stabilising selection as well. Genes whose regulation is under stabilising selection will not change their expression much. Thus, large variation will not be obtained from the data sets. In this case one would expect a Δ-value close to zero. Thus, also stabilising selection might be an important factor which has been considered, for example, by Rifkin *et al.* (2003) and Lemos *et al.* (2005) (cf. Gilad *et al.* (2006b) for a review).

5.5. Conclusion

In this chapter the Tajima's D test was adapted to gene expression data by changing the underlying evolutionary model. The simulation studies show that it is possible to estimate the population mutation rate θ from gene expression data with two different estimators. The difference of the resulting estimates is zero in expectation under a neutral model. For human data the majority of genes does not contradict the neutral theory. Unfortunately, the gene expression data sets currently available are too small for a meaningful population analysis. However, one can suppose that the data sets grow in the next years and more detailed analyses become possible. In that time complex stochastic models become more important. Hence, the way to combine the coalescent theory and gene expression evolution models might be a starting point for further research, for example, to adapt other important tests used for DNA sequences like the Hudson-Kreitman-Aguade test (Hudson *et al.*, 1987).

6. Using gene expression evolution models for medical applications

In this chapter the process of carcinogenesis is considered as an evolutionary process. Under this assumptions medical data is analysed with a gene expression evolution model in order to detect genes which are involved in the disease.

6.1. Introduction

This thesis is focused mainly on the development of models for evolution of gene expression which is regarded as a result of sequence evolution in regulatory sequences. The observed time periods of evolution between near related species are in the range of a few million years. In contrast, in this chapter relatively small time periods are considered: Medical data sets from different types of malignant diseases are analysed. First, it is discussed why gene expression evolution models can be applied to that type of data, since the biological process of carcinogenesis is apparently different from species evolution. Subsequently, the experimental design and the analysis is explained.

Tumours and stem cell diseases originate from a formerly normal cell. By mutations in specific regions which cannot be repaired gene products are changed in a fashion that the transformation to a malignant cell occurs. These mutations affect proliferation and apoptosis which results in an uncontrolled cell division. Evidence exist that mutations which take place in regulatory regions can lead to that process by changing the expression level. For instance, it is known that specific mutations permit the transcription of genes which induce the transformation to a malignant cell. These genes are called *oncogenes*. However, it is supposed that in most cases a number of mutations must occur, before malignant growing starts. Later, mutations accumulate faster during progress of the disease, since cell repair functions break down, and the disease becomes more aggressive.

Hence, the gene expression profile of a malignant cell depends strongly on the phase of the disease. Because of this progression starting from a first mutation, the process can be regarded as an evolutionary process. A more detailed introduction into the biology of human cancers is given by Schulz (2005).

Altogether, it is beyond all questions that the gene expression profiles of tumour tissues or malignant stem cells are different from their normal counterparts. Furthermore, the expression profiles of different kinds of cancers, even in the same tissue type, are variable. Thus, in many studies microarray experiments were used to classify cancer types by their specific expression profiles (e.g., by Golub *et al.* (1999); Ramaswamy *et al.* (2001); Driscoll *et al.* (2003)). The goal is an accurate diagnosis which is necessary for the best possible treatment. To this end, a big number of supervised machine learning methods and unsupervised clustering algorithms have been developed and applied. Before using one of these methods, the large number of genes typically obtained with microarrays has often been reduced to those ones with a significant gene expression difference between the examined types of cancer (cf. chapter 2.2.3). Thereby, in many cases a large ratio of genes show a significant gene expression difference, although it is assumed that only a relatively small number of mutations has occurred on sequence level. It is suggested that the majority of differences results from variations in transcription factors.

In this chapter a model based analysis is performed on a lung cancer and a leukaemia data set. First, the model parameters are estimated. Subsequently, the Bayesian mutation detection method is used to find out those genes which are involved in the disease. The method is compared with the significance analysis of microarrays (SAM) (Tusher *et al.*, 2001) which is a common approach to analyse microarray data of different classes to find out significantly expressed genes.

6.2. Materials and methods

6.2.1. The model

The gene expression evolution model with non-mutational effects (M&E model) is applied to microarray data of two different classes. It is noted that the expression of each gene evolves independently of each other in this model. This simplification might be too strong when using medical data, since the majority of gene expression changes might be no direct result of mutations but of interactions between genes. Thus, a mutation is interpreted here as a variation in the expression level of a gene which is a result of the carcinogenesis, but not necessarily a result of a mutation in the regulatory sequence of that gene. Mutation effects are assumed to follow a normal distribution with standard deviation σ_m. There are other expression level changes resulting from the environment and different pathways not affected by the disease. They are summarised by a random variable describing non-mutational effects. This variable follows a normal distribution with standard deviation σ_e. Thus, the model parameters can be estimated by a maximum-likelihood (ML) method (cf. chapter 4.2.3). The parameter estimates are used to perform the Bayesian mutation detection method to select genes which changed their expression level by a mutation (cf. chapter 4.2.4). These genes are referred to as 'mutated genes' here even if their regulatory sequences are unchanged. They considered as candidates to cause the transformation to a cancer cell.

Microarray data sets which are applied to the model can contain either two different classes of cancer which evolved from the same type of normal cells or they can describe a time series, for example, a disease which evolved from normal tissue.

6.2.2. Significance analysis of microarrays (SAM)

Beside analysis with the gene expression evolution model, SAM (Tusher *et al.*, 2001) is used to find out genes which are significantly different expressed between two classes of expression profiles. SAM assigns a score to each gene. This score depends on the expression change of that gene between the classes relative to their standard deviations. For genes with scores greater than a selected threshold, a number of permutations are used to estimate the false discovery rate (FDR) which is the percentage of genes significant by chance. Each permutation permutes the class labels which signifies the affiliation of an expression profile to its class.

6.2.3. Data sets

Two data sets each with data from two classes are analysed here. The data sets have been collected at the University Hospital Düsseldorf with Affymetrix HG Focus microarrays which measure expression levels of $8,746$ genes (and some housekeeping genes). The first set contains expression profiles from two types of lung cancer: *Adeno carcinomas*

Table 6.1.: Medical data sets which have been applied to SAM and the Bayesian mutation detection method to detect genes with differences in expression between both classes.

Data set	Array	Class 1	Class 2
Lung cancer	Affy HG Focus	Adeno carcinoma	Squamous cell carcinoma
	8746 genes	10 profiles (A1–A10)	10 profiles (P1–P10)
Leukaemia	Affy HG Focus	Normal bone marrow	Chronic myeloid leukaemia
	8746 genes	8 profiles (N1–N8)	9 profiles (C1–C9)

Table 6.2.: ML parameter estimates from the M&E-normal model for the lung cancer and leukaemia data sets.

Data set	$\hat{\sigma}_m$	$\hat{\sigma}_e$	$\hat{d}$
Lung cancer	0.956	0.203	0.560
Leukaemia	0.430	0.114	0.516

(class 1 of the lung cancer set) and *squamous cell carcinomas* (class 2 of the lung cancer set) (Rohr *et al.*, 2005). The second set includes expression profiles sampled from CD34+ hematopoietic stem and progenitor cells from bone marrow of healthy volunteers (class 1 of the leukaemia set) and CD34+ cells from bone marrow of patients with *chronic myeloid leukaemia* (CML) in chronic phase (class 2 of the leukaemia set) (Diaz-Blanco *et al.*, 2007). Table 6.1 gives an overview on the data. In the first set the two classes descended from former normal undifferentiated lung tissue. The situation in the second set corresponds to a time series in which normal cells transform to malignant cells.

6.3. Experiments and results

The ML parameter estimation method was used to estimate the model parameter values $\hat{\sigma}_m$, $\hat{\sigma}_e$, and $\hat{d}$ for both data sets (cf. chapter 4.2.3). The data sets were applied in the same way as the primate and mice data sets (cf. chapter 4.3.3). The expression profiles of class 1 were regarded as the sample 1 and the expression profiles of class 2 were regarded as the sample 2, respectively (cf. table 6.1). For both data sets 100 independent runs

Table 6.3.: Results of the Bayesian mutation detection method. The rows show the number of genes which changed their expression by the same number of mutations. The values in the S2N columns are mean values of all signal-to-noise values from all genes with the corresponding number of mutations.

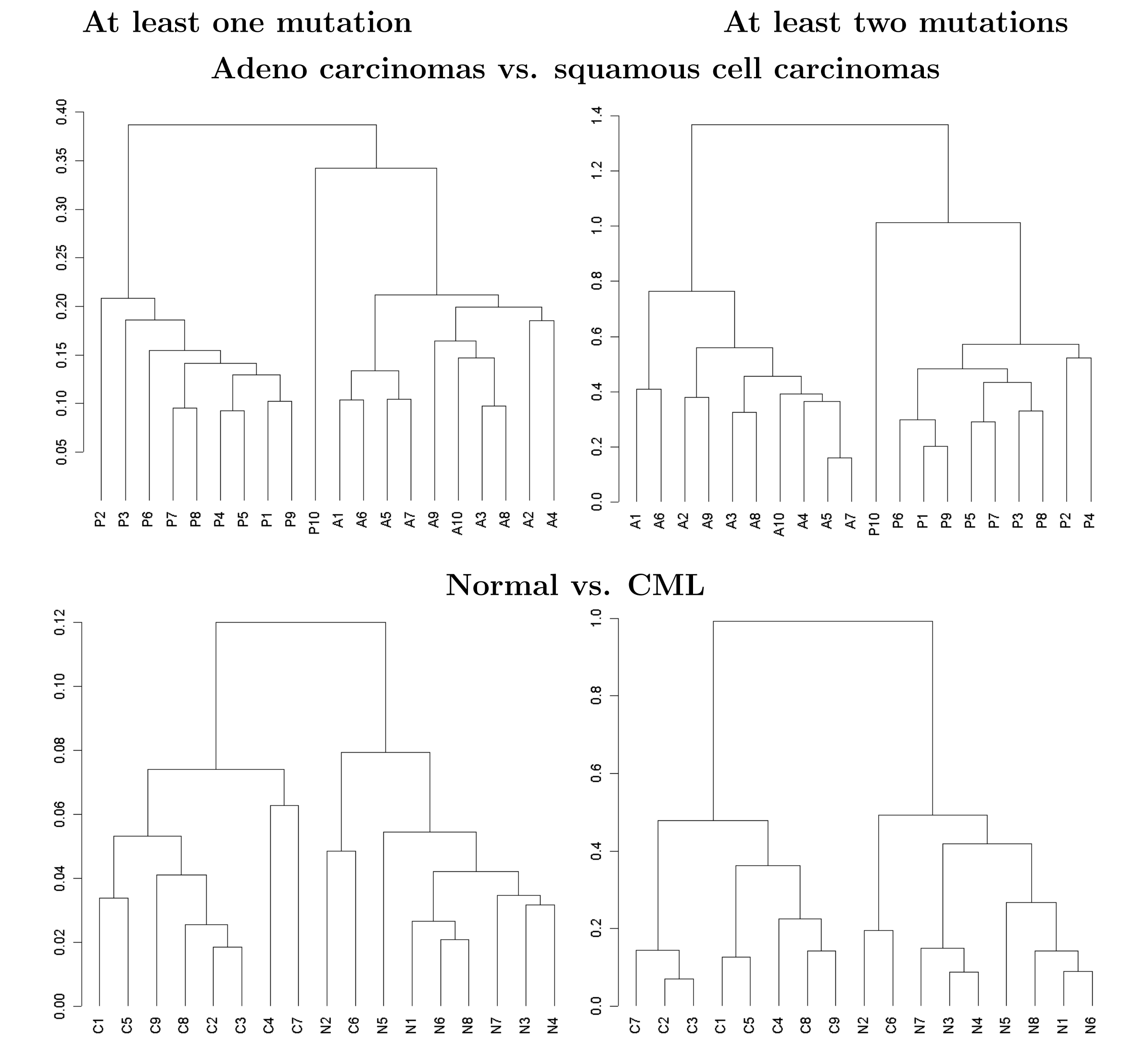
#Mutations	Lung cancer		Leukaemia	
	#Genes	Mean value of |S2N|-scores	#Genes	Mean value of |S2N|-scores
0	4,754	0.254	6,010	0.437
1	3,917	0.403	2,665	0.849
2	72	0.793	64	1.046
3	3	1.444	6	0.899
4	0	-	1	1.606

of the ML method were performed and the corresponding results with the smallest log-likelihood were considered as best estimates (cf. table 6.2). The parameters σ_m and σ_e are about twice as large in the lung cancer set than the corresponding estimates in the leukaemia set. The estimates for parameter d are similar in both sets.

After parameter estimation the Bayesian mutation detection method was used to estimate how many mutations occurred in each gene. The results are depicted in table 6.3. In the lung cancer set the Bayesian mutation detection method found $3,992$ genes in which the expression level is influenced by at least one mutation (45.65 %). In the leukemia set only $2,736$ genes (31.28 %) were found in which at least one mutation affects the expression level. The number of genes in the categories for exactly two or more mutations is very small in comparison to the overall number of $8,746$ genes measured with the microarrays. However, it has been shown that partitions of the exact number of mutations are weak (cf. chapter 4.3.2).

Additionally, for each category the mean values of the absolute values of signal-to-noise ratios (S2N) over all genes in that category were calculated. S2N is a measure which reflects the gene expression difference between two classes relative to the standard deviations within the classes. The measure has been used, for example, by Golub *et al.* (1999) or Ramaswamy *et al.* (2001). It is calculated for a single gene by the equation

$$\text{S2N(Gene)} = \frac{\mu_1(\text{Gene}) - \mu_2(\text{Gene})}{\sigma_1(\text{Gene}) + \sigma_2(\text{Gene})}, \tag{6.1}$$


At least one mutation

At least two mutations

Adeno carcinomas vs. squamous cell carcinomas

Normal vs. CML

Figure 6.1.: Dendrograms resulted from hierarchical clustering of the lung cancer and leukaemia data.

whereas μ_1 and μ_2 are the mean gene expression levels in the classes 1 and 2 and σ_1 and σ_2 are the corresponding standard deviations. In the results in table 6.3 the mean values of the |S2N|-scores over all genes in the same category correlates positively with the number of mutations for each gene. The only exception is in the leukaemia set. The 64 genes with two mutations have a greater mean value of |S2N|-scores than the genes with three mutations. Indeed, there are only six genes with three mutations, whereby the corresponding mean value of the six |S2N|-scores might be instable against outliers. The mean values of |S2N|-scores for zero and one mutation are twice as large in the leukaemia set than in the lung cancer set.

Table 6.4.: Comparison of SAM and the Bayesian mutation detection method. The second column shows the number of genes found by SAM. The last two columns show the number of genes which have been found exclusively by SAM and the number of genes which have been found also with the Bayesian mutation detection method.

Data set	#Significant $q \leq 5\,\%$	Fold Change	#Exclusive genes	#Overlapping genes
Lung cancer	389	2.0	216	173
Leukaemia	315	1.5	151	164

The genes which were detected as mutated in the regulatory region by the Bayesian mutation detection method were used to perform a hierarchical clustering to order the expression profiles in each data sets in a tree-like fashion referred to as *dendrogram*. For both data sets two different analyses were performed. In the first one genes with at least one detected mutation were used, whereas in the second one genes with at least two detected mutations were selected for clustering. Although estimation of the exact number of mutations is imprecise, genes in which two mutations have been detected have a higher chance that at least one mutation occurred than genes in which just one mutation has been detected (cf. chapter 4.3.2). The dendrograms resulting from hierarchical clustering are illustrated in figure 6.1. A good clustering into adeno carcinomas and squamous cell carcinomas was achieved in both analyses. However, the squamous cell carcinoma expression profile P10 clusters closer to the adeno carcinoma profiles than to the other squamous cell carcinoma profiles. In the leukaemia set also a good clustering was obtained with one exception in both cases. C6 is an outlier, since it clusters with the normal CD34+ cells. In both data sets the branches are considerably longer in the analysis with at least two mutations.

The two data sets were also applied to SAM. SAM was used in the "two-class unpaired mode" and $1,000$ permutations were generated (cf. Tusher *et al.* (2001) for details). Genes which were considered as significant for a FDR of $5\,\%$ were selected if their fold change was greater than 2.0 in the lung cancer set (or smaller than 0.5) and greater than 1.5 in the leukaemia set (or smaller than 0.666), respectively. The fold change is the ratio of gene expression mean values of both classes $\mu_1(Gene)/\mu_2(Gene)$. The different fold change cut-off values for the data sets were chosen in order to get a comparable number of genes. Altogether, the gene expression differences in the leukaemia set are smaller so that a fold change of 2 would have decreased the number of genes too strong. With the

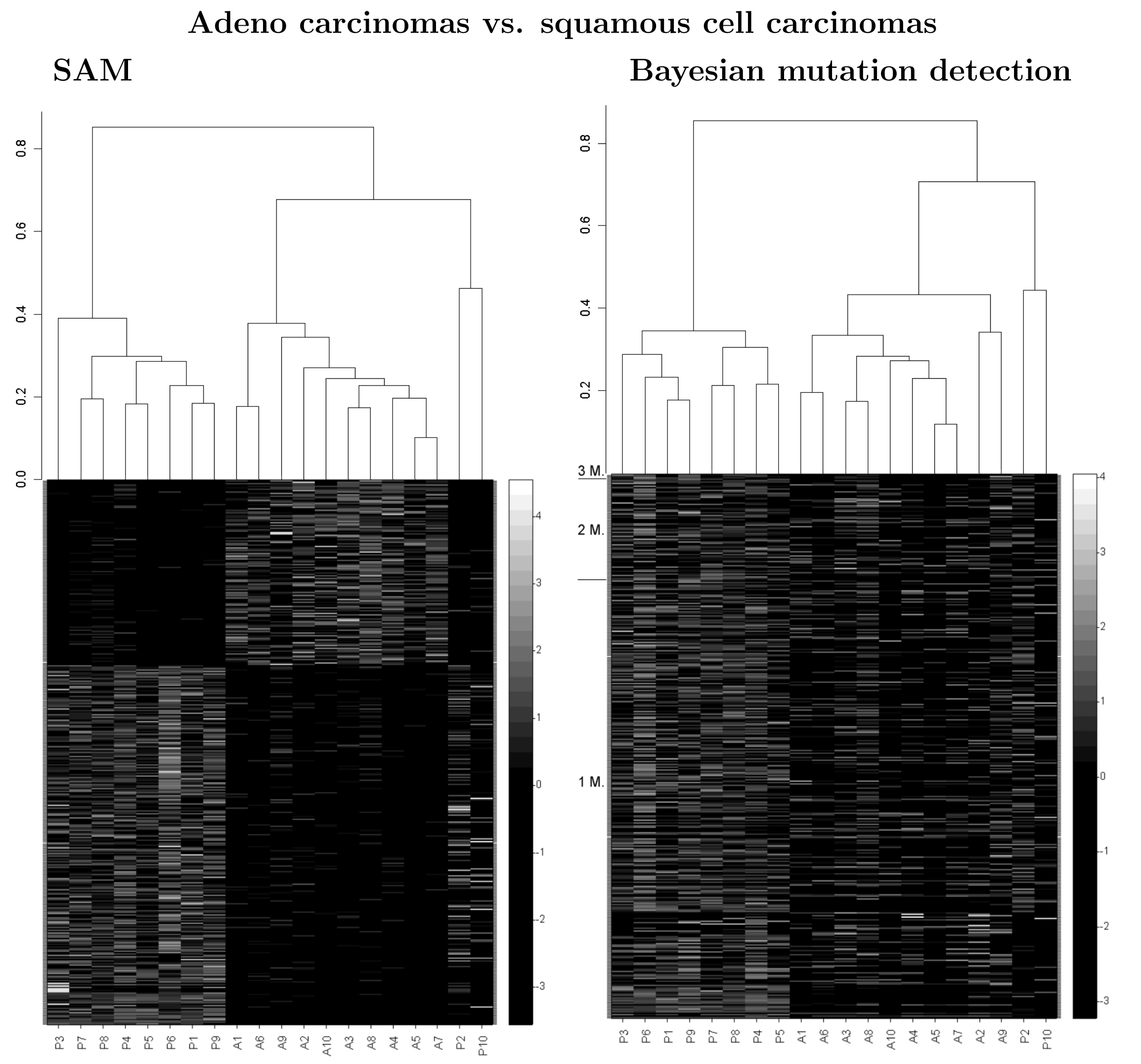
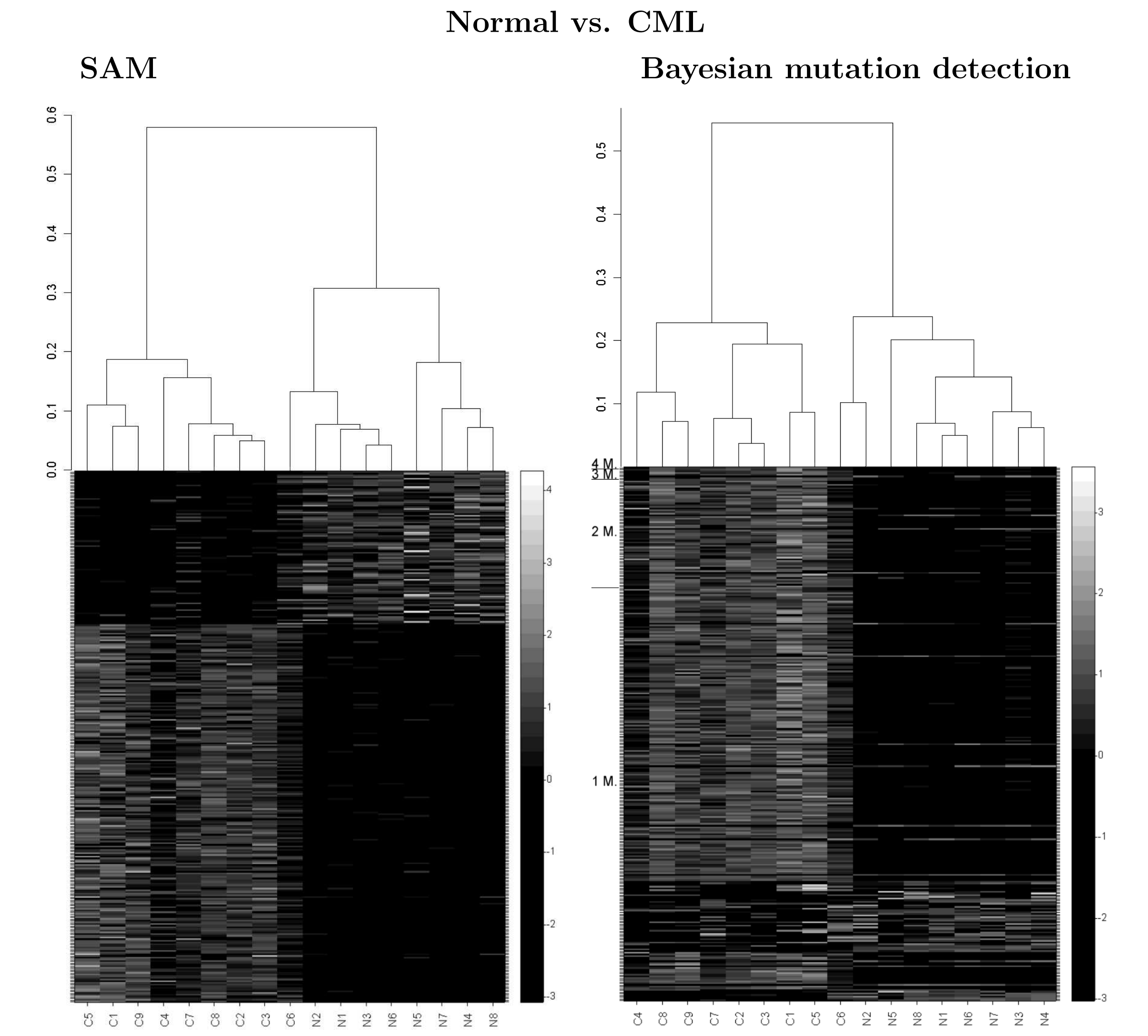


Figure 6.2.: Cluster plots with dendrograms of 389 genes of adeno carcinoma (A1–A10) and squamous cell carcinoma (P1–P10) microarray data. On the left side the genes have been selected by SAM (Fold change = 2, q-value ≤ 5 %). On the right side the ML method have been used to select mutated genes (the number of mutations is shown left of the heat map).

restriction of the fold change 389 significant genes were found in the lung cancer set and 315 ones were found in the leukaemia set. The results are presented in table 6.4.

The genes which were selected by SAM were used to perform hierarchical clustering. The dendrograms are shown in figure 6.2 for the lung cancer set and in figure 6.3 for the leukaemia set. Heat maps of the gene expression values are also shown in the same figures. Before plotting, all selected genes were sorted according to their fold change


Normal vs. CML

SAM

Bayesian mutation detection

Figure 6.3.: Cluster plots with dendrograms of 315 genes from microarray data of CD34+ hematopoietic stem and progenitor cells from bone marrow of healthy volunteers (N1–N8) and patients with CML (C1–C9). On the left side the genes have been selected by SAM (Fold change = 1.5, q-value $\leq 5\,\%$). On the right side the ML method have been used to select mutated genes (the number of mutations is shown left of the heat map).

in a decreasing order. The expression profiles P2 and P10 in the lung cancer set were clustered closer to the adeno carcinoma profiles than to their own class of squamous cell carcinomas. Altogether, a correct clustering was found in both data sets except for C6 in the leukaemia set.

Finally, the results from the Bayesian mutation detection method were compared with the results by SAM. With SAM 389 genes were detected in the lung cancer set and 315 genes were found in the leukaemia set. With the Bayesian mutation detection method each gene was mapped to the number of mutations with the largest posterior probability. All these genes were ranked in a decreasing order according to the number of mutations. In a second step the genes with the same number of mutations were sorted in a decreasing order according to their posterior probability. From the resulting list the first 389 genes from the lung cancer set and the first 315 genes from the leukaemia set were selected for hierarchical clustering. Thus, the results are comparable with the SAM approach, since an equal number of best genes from both methods are compared and used for clustering. The number of genes selected by both methods is 173 for lung cancer and 164 for leukaemia. The number of genes selected exclusively by one of the methods is 216 for lung cancer and 151 for leukaemia (cf. table 6.4). Dendrograms and heat maps resulting from clustering for genes selected with the Bayesian mutation detection method are shown in figure 6.2 for the lung cancer set and in figure 6.3 for the leukaemia set. The genes are sorted in the previous described manner. The dendrograms look very similar to those resulting from genes selected by SAM. Again, the expression profiles P2 and P10 were clustered closer to the other class than to their own class, while C6 was clustered wrong.

6.4. Discussion

It is shown that it is possible to apply the M&E-normal model to gene expression profiles of malignant diseases. It is assumed that the process of carcinogenesis is an evolutionary process on the level of gene expression in which gene expression changes accumulate over time as a result of a transformation of the cells and the progress of the disease. Two data sets are analysed. The results of parameter estimation show that the standard deviation of mutations σ_m and of environmental effects σ_e are smaller in leukaemia than in lung cancer. This indicates smaller gene expression differences, since the expected number of mutations d is similar in both sets. However, smaller gene expression differences between the two classes within the data sets are confirmed by SAM. Thereby, it was necessary to set the fold change to 1.5 instead of 2.0 to get a similar number of significant genes for the leukaemia and the lung cancer set. Also the branch lengths in the dendrograms are longer in the lung cancer set than in the leukaemia set which indicates larger gene expression differences, too.

The calculation of the mean values of the |S2N|-scores show that genes with a larger number of mutations discriminate better between the two classes according the common measure S2N. Indeed, there are overlapping regions (e.g., some genes with two mutations have a smaller |S2N|-score than other genes with one mutation) so that the |S2N|-score is no single criterion which can be used to decide how many mutations happened.

The genes which were detected as mutated were subject to a hierarchical clustering. Perfect clustering was yielded expect for a few outliers. Interestingly, the clustering results for the lung cancer set on genes with at least one ($3,992$ genes) and with at least two mutations (75 genes) (cf. figure 6.1) are better than for a cluster with 389 genes (cf. figure 6.2). The genes with at least two mutations have strong expression differences between the two classes, while there might be genes with weak differences in the group of genes with exact one mutation. This would increase the noise and is an explanation for the fact that both P2 and P10 are clustered beyond the other squamous cell carcinoma profiles. In contrast, the very large number of $3,992$ genes which contains only genes with at least one mutations might have more statistical power than the 389 genes so that outliers have no large impact. Indeed, these results depend on the used data set. They are not a general condition.

If one compares the gene lists of the Bayesian mutation detection method and SAM, a large number of genes is observed which was found exclusively by one of the methods. These genes might be important. Other methods might also deliver important results. Thus, if data is analysed, one should use different methods in combination in order to infer the cause or the state of a disease.

6.5. Conclusion

As a conclusion, the use of the ML parameter estimation method and the Bayesian mutation detection method is a possibility and alternative if medical data of different groups but the same origin are analysed. This fact makes the gene expression evolution models more valuable and enables more possibilities to analyse medical data, even on the base of a solid stochastic model.

7. Summary

Numerous recent studies deal with the examination of differences between gene expression profiles of nearly related species (cf. Ranz and Machado (2006) for a review). The evolutionary process leading to the differences is of special interest. In most studies a positive correlation of time distance between species and gene expression divergence between species was observed. Khaitovich *et al.* (2004) showed for primate species that gene expression differences accumulate linearly with time which can be explained by a neutral model according to the neutral theory by Kimura (1983). Thus, Khaitovich *et al.* (2005b) suggested a neutral model for evolution of gene expression, referred to as the M model here. The M model describes mutations in the regulatory region of a gene which alter the expression level of that gene. The mutations are Poisson distributed. The mutation effects on the level of gene expression are described by a continuous distribution, referred to as mutation effect distribution (MED). Thus, the occurrence and the effect of mutations are combined to a compound Poisson process. This thesis deals with extensions of the M model. After motivating the topic (Chapter 1) and an introduction into the biological and mathematical background (Chapter 2), more complex variants of the M model are described. Furthermore, new applications for data analysis are suggested. Finally, a medical application is given.

A model with gamma-distributed mutation effects (Chapter 3): The M-gamma model is introduced which uses a gamma distributed MED. This distribution is more flexible than previous used ones. Therefore, a better fit to analysed data seems plausible. Furthermore, the shape of the mutation effects can be analysed in more detail. In order to estimate the model parameters an optimisation method is used. After validating this method, biological gene expression data sets from different species are analysed. For primates the results show an acceleration in gene expression evolution on the human lineage in brain in comparison to the chimpanzee lineage. Furthermore, the results indicate that a positively skewed MED is better than other MEDs.

A model with mutational and non-mutational effects (Chapter 4): The M&E model is introduced which incorporates all kinds of non-mutational effects including, for example, environmental effects, the cell cycle, epistatic effects, and measurement errors. In order to estimate the model parameters, a χ^2-fit method and a maximum-likelihood method are presented. Furthermore, a Bayesian method is suggested to detect those genes mutated in their regulatory region. It is shown that the M&E model fits real data taken from microarray experiments better than the M model. Additionally, a more detailed analysis of data is enabled, since it is possible to estimate the impact of non-mutational effects. The results on primate data sets indicate differences in the number of mutations between different tissues, for instance testis show the largest and brain the smallest number of mutations between human and chimpanzee. Results on mice data indicate a closer connection between *Mus musculus domesticus* and *Mus musculus musculus* than between *Mus musculus ssp* and *Mus musculus castaneus*.

A Tajima-type test for gene expression data (Chapter 5): According to the neutral theory it is assumed that the majority of gene expression changes depending on mutations in regulatory regions are neutral. However, the level of expression of some genes might evolve under selection. In this chapter a statistical test related to the Tajima's D test is suggested which can be applied to gene expression data sampled from a natural population. A Wright-Fisher model is assumed so that the genealogy of a sample from the population can be described by the coalescent process. In order to create a test statistic, the M&E model is applied to genealogies. Two estimators for the population mutation rate are suggested. The difference of these estimators is assumed to be zero under a neutral model and can be used to decide which genes reject the neutral model. The test is evaluated with synthetic data. Results on real data taken from a human population indicate that the majority of genes do not reject the neutral model.

Using gene expression evolution models for medical applications (Chapter 6): In this chapter the carcinogenesis of a former normal cell to a malignant cancer cell is regarded as an evolutionary process. Thus, the M&E-normal model is applied, since it is known that expression profiles differ between normal cells and cancer cells. Parameter estimates from real data are used to detect those genes which changed their expression as a result of the carcinogenesis. The method is applied to two types of lung cancer and to normal stem cells and stem cells taken from patients with chronic myeloid leukaemia. The resulting genes are used to generate clusters which are almost perfect.

8. Zusammenfassung

Verschiedene Studien beschäftigen sich mit der Untersuchung von Genexpressionsunterschieden zwischen nahe verwandten Arten (siehe Review von Ranz and Machado (2006)). Ein besonderes Interesse gilt dabei dem Evolutionsprozess, der zu diesen Unterschieden führt. In vielen Fällen wurde eine positive Korrelation zwischen der Zeit, die seit der Aufspaltung der betrachteten Arten vergangen war, und dem Genexpressionsunterschied festgestellt. Bei verschiedenen Primatenarten beobachteten Khaitovich *et al.* (2004) sogar eine lineares Verhältnis zwischen diesen Größen. Dies spricht, gemäß der neutralen Theorie von Kimura (1983), für ein neutrales Evolutionsmodell. Daher entwickelten Khaitovich *et al.* (2005b) ein erstes neutrales Modell zur Beschreibung der Evolution von Genexpression, hier als M Model bezeichnet. Dieses Modell beschreibt Mutationen in der regulativen Sequenz eines Gens, wobei jede Mutation eine Veränderung in der Expressionsstärke zur Folge hat. Das Auftreten von Mutationen folgt einer Poisson-Verteilung. Jede Änderung in der Expressionsstärke wird durch eine kontinuierliche Zufallsverteilung beschrieben, die sogenannte Mutationseffektverteilung (MED). Diese Dissertation schließt an das M Modell an. Nach einer Einleitung (Kapitel 1) und einer Einführung in den biologischen und mathematischen Hintergrund (Kapitel 2) werden Erweiterungen des M Modells behandelt und Anwendungen diskutiert.

M-gamma Modell (Kapitel 3): In diesem Kapitel wird eine komplexere MED eingeführt. Im M Modell wurden zuvor nur 1-parametrige Verteilungen genutzt. Die hier vorgestellte Gammaverteilung ist durch zwei Parameter definiert und somit flexibeler, wodurch eine detailliertere Datenanalyse möglich ist. Zur Schätzung der Modellparameter wird ein Optimierungsverfahren benutzt. Dieses wird durch Simulation mit künstlichen Daten evaluiert und auf echte Daten angewendet. Dabei zeigt sich, dass Mutationseffekte durch schiefe Verteilungen besser beschrieben werden können als durch symmetrische. Außerdem läßt sich bei Menschen eine im Vergleich zu Schimpansen beschleunigte Evolution im Gehirn festgestellt.

M&E Modell (Kapitel 4): In diesem Kapitel wird das M&E Modell als Erweiterung des M Modells vorgestellt. Dieses Modell beschreibt auch solche Effekte, welche die Expressionsstärke eines Gens verändern, aber nicht auf Mutationen, sondern auf Umwelteinflüssen, dem Zellzyklus, Messfehlern und ähnlichen Effekten beruhen. Zum Schätzen der Modellparameter werden eine χ^2- und eine Maximum-Likelihood-Methode vorgestellt. Zusätzlich wird eine einfache bayesianische Methode zum Auffinden der Gene mit Mutationen in regulativen Regionen diskutiert. Es wird gezeigt, dass das M&E Modell echte Daten besser beschreibt als das M Modell. Außerdem ermöglicht es eine detailliertere Datenanalyse. Eine Anwendung auf Daten von Menschen und Schimpansen zeigt eine unterschiedlich schnelle Evolution zwischen verschiedenen Geweben, z.B. gibt es in Hodengewebe mehr Mutationen in regulativen Regionen als im Gehirn. Analysen von Mäusen zeigen eine engere evolutionäre Distanz zwischen *Mus musculus domesticus* und *Mus musculus musculus* als zwischen *Mus musculus ssp* und *Mus musculus castaneus.*

Tajima's D Test für Genexpressionsdaten (Kapitel 5): In diesem Kapitel werden evolutionäre Veränderungen der Expressionsstärke innerhalb von Arten betrachtet und auf Einfluss von Selektion getestet. Dazu wird eine Adaption des Tajima's D Tests vorgenommen. Es wird angenommen, dass sich Populationen gemäß eines Wright-Fisher Modells verhalten und sich ihre Genealogie somit durch einen Coalescent-Prozess beschreiben läßt. Um Varianz in Genexpression zu beschreiben wird das M&E Modell angewendet. Es werden zwei Schätzer für die Populationsmutationsrate präsentiert, deren Differenz gemäß des Tajima's D Tests im neutralen Fall Null ergibt. Signifikante Abweichungen deuten dagegen auf gerichtete oder balancierende Selektion hin. Eine Analyse mit echten Expressionsdaten von Menschen zeigt, dass sich die überwiegende Mehrheit der Gene neutral verhalten.

Medizinische Anwendung (Kapitel 6): In diesem Kapitel wird Entstehung und Verlauf einer Krebserkrankung als ein evolutionärer Prozess betrachtet, der ebenfalls die Expressionsstärke von Genen beeinflusst. Unter dieser Annahme werden medizinische Expressionsdaten von zwei Lungenkrebsarten sowie von normalen Stammzellen und solchen von Patienten mit chronischer myelotischer Leukämie analysiert. Mit Hilfe der bayesianischen Methode aus Kapitel 4 werden Gene herausgefiltert, bei denen Mutationen Expressionsänderungen hervorgerufen haben. Diese Gene werden zur Erstellung von Dendrogrammen verwendet, die eine fast perfekte Aufteilung der Datensätze entsprechend ihres Phänotyps anzeigen.

A. Software packages

All models and applications described in this thesis were implemented into two software packages called EMOGEE and EMOGEE Tools, respectively. Both programs use a configuration file to set up the program functions.

EMOGEE contains the M, the M-gamma, the M&E-normal and the M&E-extreme model. It is possible to apply these models to experimental data sets to estimate the model parameters with the presented optimisation methods. Furthermore, it is possible to generate data.

EMOGEE Tools implements the Bayesian mutation detection method and the Tajima-type test. Before using one of these applications it is necessary to estimate the model parameters of the corresponding data with EMOGEE. The respective results have to be feeded into the configuration file of EMOGEE Tools.

The C++ source code of both packages is available on the homepage

`http://www.cibiv.at/software/emogee`

Detailed manuals for the programs are stored as `.pdf`-files on the same homepage.

B. Abbreviations

A	– Adenine
ANOVA	– Analysis of variance
cDNA	– Complementary deoxyribonucleic acid
cRNA	– Complementary ribonucleic acid
C	– Cytosine
cas	– Mice subspecies *Mus musculus castaneus*
G	– Guanine
DNA	– Deoxyribonucleic acid
dom	– Mice subspecies *Mus musculus domesticus*
EMOGEE	– Estimator for models of gene expression evolution
FDR	– False discovery rate
M model	– The basic gene expression evolution model by Khaitovich *et al.* (2005b)
M-gamma model	– The gene expression evolution model with gamma distributed mutation effects
M&E model	– The general gene expression evolution model with non-mutational effects, but without specifying the mutation effect distribution
M&E-normal model	– The gene expression evolution model with normal distributed mutation effects and normal distributed non-mutational effects
M&E-extreme model	– The gene expression evolution model with extreme value distributed mutation effects and normal distributed non-mutational effects
MED	– Mutation effect distribution

ML	– Maximum-likelihood
MM	– Mismatch
MRCA	– Most recent common ancestor
mRNA	– Messenger ribonucleic acid
mus	– Mice subspecies *Mus musculus musculus*
N-MED	– Non-mutational effect distribution
PM	– Perfect match
RMA	– Robust multichip average
RNA	– Ribonucleic acid
S2N	– Signal to noise ratio
SAM	– Significance analysis for microarrays
spretus	– Mice species *Mus spretus*
ssp	– Mice subspecies *Mus musculus ssp*
T	– Thymine
tRNA	– Transfer ribonucleic acid
U	– Uracil
VSN	– Variance stabilisation and calibration for microarray data

Bibliography

Affymetrix (2004) *GeneChip Expression Analysis.* Technical Manual.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) *Molecular Biology of the Cell, Fourth Edition.* Garland Science, New York.

Bäck, T. and Schwefel, H.-P. (1993) An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, **1**, 1–23.

Baldi, P. and Hatfield, G.-W. (2002) *DNA Microarrays and Gene Expression.* Cambridge University Press, Cambridge, U.K.

Banzhaf, W., Nordin, P., Keller, R. and Francone, F. (1998) *Genetic Programming - An Introduction.* Morgan Kaufmann, San Francisco.

Bickel, P.-J. and Doksum, K.-A. (2001) *Mathematical Statistics Vol. 1, Second Edition.* Prentice Hall, New Jersey.

Bolstad, B.-M., Irizarry, R.-A., Astrand, M. and Speed, T.-P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

Brent, R.-P. (1972) *Algorithms for minimization without derivatives.* Prentice-Hall.

Bulmer, M. (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics*, **129**, 897–907.

Caceres, M., Lachuer, J., Zapala, M.-A., Redmond, J.-C., Kudo, L., Geschwind, D.-H., Lockhart, D.-J., Preuss, T.-M. and Barlow, C. (2003) Elevated gene expression levels distinguish human from non-human primate brains. *PNAS*, **100**, 13030–13035.

Campbell, N.-A. and Reece, J.-B. (2005) *Biology, Seventh Edition.* Benjamin Cummings, San Francisco.

Cavalieri, D., Townsend, J.-P. and Hartl, D.-L. (2000) Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by dna microarray analysis. *PNAS*, **97**, 12369–12374.

Christianini, N. and Shawe-Taylor, J. (2000) *Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, Cambridge.

Chu, T.-M., Weir, B. and Wolfinger, R. (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences*, **176**, 35–51.

Cover, T. and Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**, 21–27.

Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection.* John Murray, London, U.K.

Diaz-Blanco, E., Bruns, I., Neumann, F., Fischer, J.-C., Graef, T., Rosskopf, M., Brors, B., Pechtel, S., Bork, S., Koch, A., Baer, A., Rohr, U.-P., Kobbe, G., von Haeseler, A., Gattermann, N., Haas, R. and Kronenwett, R. (2007) Molecular signature of CD34+ hematopoietic stem and progenitor cells of patients with CML in chronic phase. *Leukemia*, **21**, 494–504.

Driscoll, J.-A., Worzel, B. and MacLean, D. (2003) Classification of gene expression data with genetic programming. In Riolo, R. L. (ed.), *Genetic Programming: Theory and Practise*, Kluwer.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. American Statistical Association*, **97**, 77–87.

Efron, B. (1979) Bootstrap method: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G., Bontrop, R. and Pääbo, S. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.

Ewens, W.-J. (1972) The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, **3**, 87–112.

Ewens, W.-J. and Grant, G.-R. (2001) *Statistical Methods in Bioinformatics.* Springer, New York, Berlin, Heidelberg.

Feller, W. (1957) *An Introduction to Probability Theory and its Applications.* John Wiley and Sons, New York.

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

Fisher, R.-A. (1922) On the dominance rate. *Proc. Roy. Soc. Edinburgh*, **42**, 321–341.

Fisher, R.-A. (1930) *The Genetic Theory of Natural Selection, 1st edn.* Clarendon Press.

Fu, Y.-X. and Li, W.-H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S. and Wayne, M. (2004) Extensive sex-specific non-additivity of gene expression in *Drosophila melanogaster. Genetics*, **167**, 1791–1799.

Gilad, Y., Oshlack, A. and Rifkin, S.-A. (2006a) Natural selection on gene expression. *TRENDS in Genetics*, **22**, 456–461.

Gilad, Y., Oshlack, A., Smyth, G.-K., Speed, T.-P. and White, K.-P. (2006b) Expression profiling in primate reveals a rapid evolution of human transcription factors. *Nature*, **440**, 242–245.

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

Golub, T.-R., Slonim, D.-K., Tamayo, P. and Huard, C. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Griffiths, A.-J.-F., Miller, J.-H., Suzuki, D.-T., Lewontin, R.-C. and Gelbart, W.-M. (2002) *An Introduction to Genetic Analysis.* W.–H. Freeman and Company, New York.

Gu, X. (2004) Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics*, **167**, 531–542.

Haldane, J.-B.-S. (1932) *The Causes of Evolution.* Harpner and Row, New York.

Hardy, G.-H. (1908) Mendelian proportions in a mixed population. *Science*, **28**, 49–50.

Harris, H. (1966) Enzyme polymorphism in man. *Proc. Roy. Soc. London, Ser. B*, **164**, 298–310.

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

Hein, J., Schierup, M.-H. and Wiuf, C. (2005) *Gene Genealogies, Variation and Evolution.* Oxford University Press, Oxford, UK.

Hsieh, W.-P., Chu, T.-M., Wolfinger, R.-D. and Gibson, G. (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, **165**, 747–757.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Hudson, R.-R. (1991) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–49.

Hudson, R.-R., Kreitman, M. and Aguado, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Irizarry, R.-A., Bolstad, B.-M., Collin, F., Cope, L.-M., Hobbs, B. and Speed, T.-P. (2003a) Summaries of affymetrix genechip probe level data. *Nucleic Acid Research*, **31**, 1–8.

Irizarry, R.-A., Hobbs, B., Collin, F., Beazer-Barclay, Y.-D., Antonellis, K.-J., Scherf, U. and Speed, T.-P. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Jin, W., Riley, R.-M., Wolfinger, R.-D., White, K.-P., Passador-Gurgel, G. and Gibson, G. (2001) The contribution of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nature Genetics*, **29**, 389–395.

Johnson, S.-C. (1967) Hierarchical clustering schemes. *Psychometrika*, **2**, 241–254.

Joyce, P. (1995) Robustness of the ewens sampling formula. *J. Appl. Prob.*, **32**, 602–622.

Jukes, T.-H. and Cantor, C.-R. (1969) Evolution of protein molecules. In Munroe, H.-N. (ed.), *Mammalian Protein Metabolism*, pages 21–132, Academic Press.

Kerr, M., Martin, M. and Churchill, G. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Khaitovich, P., Enard, W., Lachmann, M. and Pääbo, S. (2006) Evolution of primate gene expression. *Nature Reviews Genetics*, **7**, 693–702.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M. and Pääbo, S. (2005a) Parallel patterns of evolution in the genomes and transcriptomes of human and chimpanzees. *Science*, **309**, 1850–1854.

Khaitovich, P., Pääbo, S. and Weiss, G. (2005b) Towards a neutral evolutionary model of gene expression. *Genetics*, **170**, 929–939.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. and Pääbo, S. (2004) A neutral model of transcriptome evolution. *PLoS Biology*, **2**, 682–689.

Kimura, M. (1968a) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Kimura, M. (1968b) Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res.*, **11**, 247–269.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.

Kimura, M. (1983) *The neutral theory.* Cambridge University Press, Cambridge, UK.

King, M.-C. and Wilson, A.-C. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.

Kingman, J.-F.-C. (1982) The coalescent. *Stoch. Process. Appl.*, **13**, 235–248.

Knudsen, S. (2002) *Analysis of DNA Microarray Data.* John Wiley and Sons, New York.

Kohonen, T. (2001) *Self-Organizing Maps. 3rd ed.* Springer, New York.

Koza, J. (1992) *Genetic Programming.* MIT Press, Cambridge, MA.

Lande, R. (1976) Natural selection and random genetic drift in phenotypic evolution. *Evolution*, **30**, 314–334.

Lemos, B., Meiklejohn, C.-D., Caceres, M. and Hartl, D.-L. (2005) Rates of divergence in gene expression profiles of primates, mice, and flies: Stabilizing selection and variability among functional categories. *Evolution*, **59**, 126–137.

Lewontin, R.-C. and Hubby, J.-L. (1966) A molecular approach to the study of genic heterozygosity in natural populations, ii. amount of variation and degree of heterozygosity in natural populations of drosophila pseudoobscura. *Genetics*, **54**, 595–609.

Li, L., Darden, T.-A., Weinberg, C.-R., Levine, A.-J. and Pedersen, L.-G. (2001) Gene assessment and sample classification for gene expression data using a genetic algorithm/k−nearest neighbor method. *Combinatorial Chemistry & High Throughput Screening*, **4**, 727–739.

Lynch, M. and Hill, W.-G. (1986) Phenotypic evolution by neutral mutation. *Evolution*, **40**, 915–935.

McDonald, J.-H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila. Nature*, **351**, 652–654.

McQueen, J. (1967) Some methods for classification and analysis of multivariate observations. *5th Berkeley Symposium on mathematics, Statistics and Probability*, **1**, 281–298.

Morley, M., Molony, C.-M., Weber, T.-M., Devlin, J.-L., Ewens, K.-G., Spielman, R.-S. and Cheung, V.-G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Nelder, J. and Mead, R. (1965) A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Nuzhdin, S.-V., Wayne, M.-L., Harmon, K.-L. and McIntyre, L.-M. (2004) Common pattern of evolution of gene expression level and protein sequence in drosophila. *Molecular Biology and Evolution*, **21**, 1308–1317.

Oleksiak, M., Churchill, G. and Crawford, D. (2002) Variation in gene expression within and among natural populations. *Nature Genetics*, **32**, 261–6.

Oleksiak, M.-F., Roach, J.-L. and Crawford, D.-L. (2005) Natural variation in cardic metabolism and gene expression in fundulus heteroclitus. *Nature Genetics*, **37**, 67–72.

Pearson, K. (1894) Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, **185**, 71–110.

Press, W.-H., Teukolsky, S.-A., Vetterling, W.-T. and Flannery, B.-P. (1992) *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK.

Ramaswamy, S., Tamayo, P., Rifkin, R. and Mukherjee, S. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS*, **26**, 15149–15154.

Ranz, J.-M. and Machado, C.-A. (2006) Uncovering evlutionary patterns of gene expression using microarrays. *Trends in Ecology and Evolution*, **21**, 29–37.

Rifkin, S., Kim, J. and White, K. (2003) Evolution of gene expression in the drosophila melanogaster subgroup. *Nat. Genetics*, **33**, 138–144.

Rohr, U.-P., Rohrbeck, A., Geddert, H., Kliszewski, S., Rosskopf, M., von Haeseler, A., Schwalen, A., Steidl, U., Fenk, R., Haas, R. and Kronenwett, R. (2005) Primary human lung cancer cells of different histological subtypes can be distinguished by specific gene expression profiles. *Onkologie 2005*, **28(suppl 3)**, 127.

Schadt, E., Monks, S., Drake, T., Lusis, A., Che, N., Colinayo, V., Ruff, T., Milligan, S., Lamb, J., Cavet, G., Linsley, P., Mao, M., Stoughton, R. and Friend, S. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.

Schena, M. (2003) *Microarray analysis*. John Wiley and Sons, Hoboken, New Jersey, USA.

Schulz, W.-A. (2005) *Molecular Biology of Human Cancers*. Springer, Dordrecht, The Netherlands.

She, J.-X., Bonhomme, F., Boursot, P., Thaler, L. and Catzeflis, F. (1990) Molecular phylogenies in the genus *Mus*: Comparative analysis of electrophoretic scndna hybridization, and mtdna rflp data. *Biol. J. Linn Soc.*, **41**, 83–103.

Speed, T. (2003) *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall/CRC, Boca Raton, London, New York, Washington D.C.

Storey, J.-D., Madeoy, J., Strout, J.-L., Wurfel, M., Ronald, J. and Akey, J.-M. (2007) Gene-expression variation within and among humand populations. *The American Journal of Human Genetics*, **80**, 502–509.

Tajima, F. (1983) Evolutionary relationship of dna sequences in finite populations. *Genetics*, **105**, 437–460.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, **123**, 585–595.

Taylor, H.-M. and Karlin, S. (1998) *An Introduction To Stochastic Modeling.* Academic Press, San Diego, California.

Tusher, V.-G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 5116–5121.

Uzzel, T. and Corbin, K.-W. (1971) Fitting discrete probability distributions to evolutionary events. *Science*, **172**, 1089–1096.

Voolstra, C., Tautz, D., Farbrother, P., Eichinger, L. and Harr, B. (2007) Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. *Genome Research*, **17**, 42–49.

Watterson, E.-A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256–276.

Wayne, M.-L., Pan, Y.-J., Nuzhdin, S.-V. and McIntyre, L.-M. (2004) Additivity and trans-acting effects on gene expression in male *Drosophila simulans. Genetics*, **168**, 1413–1420.

Weinberg, W. (1908) Über den Nachweis der Vererbung beim Menschen. *Jahresh. Verein f. Vaterl. Naturk. Württem*, **64**, 368–382.

Weldon, W.-F.-R. (1901) A first study of natural selection in clausilia laminata. *Biometrika*, **1**, 109–124.

Whittaker, E.-T. and Robinson, G. (1967) *The Calculus of Observations: A Treatise on Numerical Mathematics.* New York, Dover.

Wilson, A.-C., Maxson, L.-R. and Sarich, V.-M. (1974) Two types of molecular evolution. evidence from studies of interspecific hybridization. *PNAS*, **71**, 2843–2847.

Wright, S. (1931) Evolution in mendelian populations. *Genetics*, **16**, 97–159.

Wright, S. (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. VI. Intern. congr. Genet.*, **1**, 356–366.

Zuckerkandl, E. and Pauling, L. (1965) Evolutionary divergence and convergence in proteins. In Bryson, V. and Vogel, H.-J. (eds.), *Evolving Genes and Proteins*, pages 97–166, New York, Academic Press.