1 Introduction

If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat.

— Douglas Adams, British author and satirist (1952-2001)

1.1 History and Appetizer

Aristotle (384 BC – 322 BC), perhaps the most famous pioneer of biological science, described the analysis of growth and development of live in his work *On the Generation of Animals* [Ari]. He opened fertilized chicken eggs at several mature times for observing, when the visible organs were generated. However, Aristotle and many of his successors did empirical research based on macroscopic observations, and the results are often influenced by religious or spiritual beliefs. This held up to the Middle Ages.

During the scientific revolution the research abandoned supernatural argumentation and started to collect facts and involve mathematics. Many important inventions and developments paved the way to modern biology and especially genetics. The invention and enhancements of the microscope enables the view onto living cells. In 1676, the Dutch tradesman Antonie van Leeuwenhoek (1632 - 1723) observed microorganisms for the first time which established the field of *microbiology*.

About 200 years later, in 1865, the Augustinian Gregor Johann Mendel (1822 - 1884) published at two meetings of the Brünn Natural History Society his research results concerning systematical breeding experiments with pea plants, which laid the foundation for the biological field of *genetics*. He suggested the existence of genes, basic units carrying the traits from parents to offspring. His report also contained several mathematical formulas for the laws of heredity [Men66].

Thomas Hunt Morgan (1866 – 1945) was able to prove the existence of genes and that these are situated on inner cellular structures which were called *chromosomes*. In 1933 he was awarded the Nobel Prize in Physiology or Medicine for these results.

Several discoveries as gene mutations and the deoxyribonucleic acid (DNA) led to the *central dogma of molecular biology* articulated by the British molecular biologist Francis Crick (1916 – 2004) in 1958 [Cri58]:

“Once information has got into a protein it can’t get out again.”

This transfer as well as the involved elements will be described in detail in chapter 2.

Finally, the microarray technology was invented in the late eighties of the last century and a gene expression profiling using miniaturized *cDNA microarrays* was presented for the first time in 1995 by Mark Schena, Dari Schalon et al. [SSDB95].

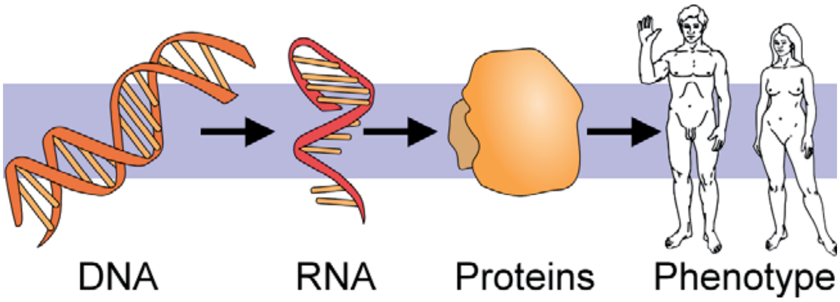

Figure 1.1: Genetic information flow

Although huge steps in biological science were made since Aristotle, this work describes the analysis of experiments which are very similar. However, due to the developments and inventions of the scientists mentioned above as well as many others the analysis could be done using more mathematics. Whereas the chicken mature experiment was mainly based on visible changes of the organism, this work will use subjective visual inspection only in the very first step. Thereafter microarray experiments were set up and evaluated using appropriate mathematical methods.

This skip from biology to mathematics should be used to introduce also some of the mathematicians, whose work were essential for the microarray analysis presented in this thesis, sorted by the usage of their methods during the analysis.

The Briton Sir Ronald Aylmer Fisher (1890-1962) was one of the most famous biologists and statisticians of the 20th century. He especially contributed to statistical design of experiments and analysis.

In 1979, the statistician Bradley Efron (born 1938) published the bootstrap technique for computer-based calculation of estimator accuracies [Efr79]. This method is essential for the time course interpolation of the microarray measurements which was done in this work.

This leads to the Romanian Isaac Jacob Schoenberg (1903-1990) who became famous for the development of interpolating splines [Sch46].

Last but not least George David Birkhoff (1884-1944) who formulated the modern dynamical system, but representing all mathematicians who contributed to systems and control theory which will be needed as final step in the analysis and modeling of the gene interaction network.

1.2 Task and Work flow

This work focuses on experiments made by the Institute of Biotechnology and Drug Research in Kaiserslautern for analyzing the genetic expression time courses during the growth of the fungus Magnaporthe grisea. All steps from experimental design up to the generation of a gene interaction network had to be mathematical well-founded.

However, two bottlenecks hampered the work:

When the project started in 2005 there was exactly one microarray data set published comparing dormant and germinated fungus spores. In fall 2007 the *Magnaporthe grisea Oryza sativa interaction database* (www.mgosdb.org) was set up to allow web-based submission and publishing of microarray data. [WV09]

So the complete experiments and the analysis had to be made from scratch. And even today there are only few suitable data sets freely available. The public repository *Gene Expression Omnibus* of the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/geo/) contains eleven microarray data sets in the beginning of 2010.

Secondly the budget of the project restricted the number of microarrays to be available. While many qualitatively impressing results are presented in literature, the own sight had to be lowered with respect to this boundary. Methods as for example Bonferroni-techniques for handling the statistical significance of gene family-wise test statements had to be neglected since they would result in an increase of the needed sample size and thus the needed microarrays.

Based on these guidelines, the experimental flow was as follows (cf. figure 1.2):

First of all the growth of the fungus was visually inspected for detecting phenotypical changes, which gave rise to the first time points for the measurements. Microarray experiments were made and the resulting data was statistically analyzed. In the time intervals exhibiting the most and highest changes in gene expression levels additional time knots were inserted. A second run of microarrays was used to hybridize all chosen time points of the fungus growth in a balanced and even manner.

The resulting data was normalized and the gene expression levels were statistically estimated. Thereafter the discrete time measurements were interpolated to receive a continuous gene expression time course. These time courses were clustered into large sets of simultaneously expressed genes before they are fitted by a mathematical model.

Anyhow, all these working stages and the mentioned keywords will be explained more precise and step by step in the following chapters.

1.3 Structure of the Thesis

Chapter 2 contains biological and technical basics which are essential for readers without appropriate background knowledge to understand the following analysis methods. A short

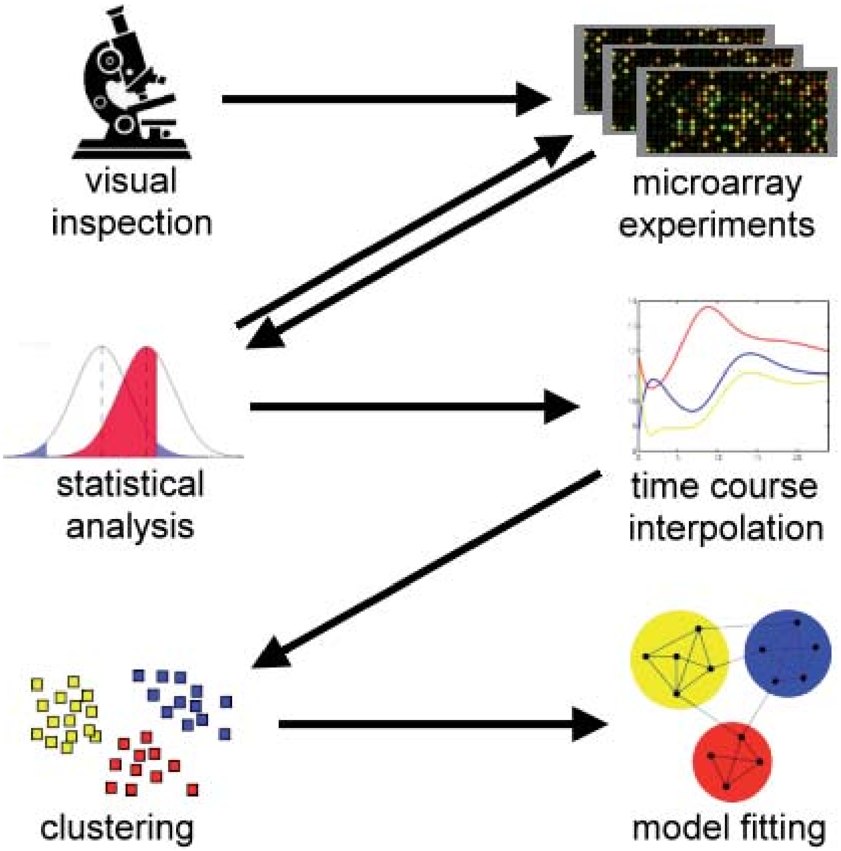

Figure 1.2: Work flow presented in this thesis

introduction into genetics is given and the functionality of microarray chips is presented. Especially sources of variances during the experiments are pointed out. Furthermore, the test organism Magnaporthe grisea, its growth and relevance in agriculture is described.

In chapter 3 all methods used for the design of experiments, the normalization, and data extraction of microarray measurements are presented. Light is shed on the key elements and main influences for experimental design before the finally used design is given. Thereafter several normalization steps for handling different error sources of microarray experiments are shown. The chapter is completed by the statistical analysis of the data using the non-parametric Fisher-Pitman-Test and the calculation of the minimal number of microarrays needed for a specific experiment.

Chapter 4 deals with the interpolation of the microarray measurements during time. Therefore, the interpolation points and its accuracies are calculated. For the latter one an exact variant of the bootstrap method is introduced. Based on these values a smoothing splines are fitted to the data resulting in continuous estimations of the expression time courses of each gene.

Due to the fact, that the calculation of a full-genome interaction network is not possible – Magnaporthe grisea has more than 15000 genes – the genes had to be clustered. This is done in chapter 5. Therefore, several appropriate distance measures of gene time courses as well as three common clustering methods are discussed. The resulting clusters were validated using quality indices. Finally, methods for estimating the overall time course of clusters are presented.

Chapter 6 shows the calculation of the cluster interaction network. Interestingly a linear model fits well to the data while more complex models as recurrent neural networks which take non-linear effects as saturation into account did not yield comparable results or did not converge at all. Neither classical neural network training algorithms as the backpropagation through time [RHW86], [Wer90] nor Bayesian particle filter methods [Hau08] did satisfying jobs fitting non-linear models to the data. Thus this chapter focuses a discrete linear time-invariant state space model, its fitting to the data and the evaluation of the result based on system theoretical properties.

In chapter 7 finally the complete procedure from design of experiments up to the gene interaction model is reviewed and summarized.

Please note also the extensive bibliography with many books and articles containing alternative approaches and possible extensions depending on the available data records.

This work is a mixture of many fields of mathematical application. Unfortunately each field has its own notation, which results in several overlaps in the usage of letters in the different chapters. Nevertheless, the common notation is kept, such that each mathematical chapter (i.e. the chapters 3-6) is written in its own private variable context. At the end of each of these chapters the results are summarized, applied to the given data example. Additionally, the resulting parameters, variables, and formulas which will be used in the following chapters are defined. This notation is kept unique.