1 Motivation and outline

A main field of computational molecular biology covers the exploration of genomic data. One of the basic data sources are gene expression experiments. Here the goal is to target the instantaneous gene activity in a specific cell of a living organism. From the amount of gene activity we can draw a picture of the cell's current state that helps to understand the fundamentals of gene regulation, signal transduction through pathways, and to further explain diseases like cancer. Microarrays provide this snapshot of cell state: a microarray is a device to measure gene expression on a large-scale basis, that is for thousands of genes at the same time. In Chapter 2 we explain how microarray technology works and what preprocessing steps we have to apply to the raw gene expression data to get a sound basis for further analysis. In Chapter 4 we introduce a collection of six microarray data sets from different cancer studies. We illustrate the reviewed methods by applying them on these real-world data sets as well as on simulated data.

Throughout this thesis, we are concerned about the first step in the statistical analysis of gene expression data: the search for *differentially expressed genes*. With each array, we can examine one sample of cells, possibly taken from a single cancer patient. If we then measure the gene expression of a panel of diseased and not-diseased patients, we might observe differences in the expression levels of certain genes between the two groups of patients. The questions to raise are: which genes show differences in gene expression? And, are these differences *significant*? Hence we want to know for each single gene whether it shows significant differences in expression. In statistical terms, each gene i is connected to a null hypothesis, stating that gene i is not differentially expressed between the two patient groups. With tens of thousands of genes measured simultaneously in a microarray experiment, we have to infer on tens of thousands of null hypotheses simultaneously. This situation is known as the *multiple testing problem*. For each hypothesis and gene we get a p-value expressing the significance of the observed difference. A low p-value supports the evidence of the observation. Since biological data is subject

to random fluctuations, a p-value can be small simply by chance. Now the goal is to search for those of the tens of thousands of p-values that provide enough evidence for a significant change in gene expression. We call the resulting set of genes the differentially expressed genes. In multiple testing theory, the search for differentially expressed genes is accomplished by defining a global error rate. A classical and popular error rate is the family-wise error rate, defined as the probability that our set includes at least one gene that is truly not differentially expressed—a false positive finding. Thus we might successively include small p-values into our set as long as the estimated family-wise error rate does not exceed a certain threshold. This probability concept works well with only a few hypotheses under test. With thousands of hypotheses under test, control of the family-wise error rate might be too conservative in the sense that only a few genes pass the search and are called differentially expressed.

A second popular multiple testing concept is termed the false discovery rate. It was recently re-discovered in the light of large-scale inference on microarray data. The false discovery rate is defined as the expected proportion of false positive findings among all positive findings. Similar to the family-wise error rate concept, we search for a set of genes such that the estimated false discovery rate does not exceed a pre-specified threshold. The false discovery rate concept is less conservative that the family-wise error rate concept and thus provides us with a larger set of significant genes. Note that these two gene sets differ with respect to the error rate that was used during the search. In multiple testing theory, we speak of *control* of an error rate if we can guarantee that the estimated value does not exceed the true value. For both concepts rich research exists, which lead to many search procedures providing certain control of the respective error rate. Control of an error rate is important to provide a hard decision rule for dividing the set of genes into those that are significantly differentially expressed and into those that are not.

In this thesis we focus on the *estimation* of error rate values and not on control. We believe that significance analysis of microarray data benefits from improved probability estimates. Thus our major interest is in the individual estimates of single genes and not whether these estimates provide control. We do not present our results in the light of multiple testing theory. For comprehensive reviews on multiple testing issues in presence of microarray data we refer to the paper series of Dudoit *et al.* (2004) and van der Laan *et al.* (2004). To distinguish between the concepts of control and estimation, we use the term *p-value filter* instead of multi-

ple testing procedure. A p-value filter is any procedure that narrows down our set of genes to those providing evidence for differential expression. Our first contribution to an improved significance analysis has its foundations in false discovery rate theory. We provide a review on false discovery rates in Chapter 3, starting with an introduction on p-value filters in Section 3.1 and the definition of the false discovery rate in Section 3.2. A p-value filter based on the false discovery rate has advantages over other filters but has also certain disadvantages. We illustrate the drawbacks with examples in Section 3.3. A variant of the false discovery rate, termed the local false discovery rate, is not affected by this disadvantages. Although the two rates share almost the same name, their underlying concepts are quite opposite to each other. A p-value filter leaves only those genes with p-values below a certain threshold. The local false discovery rate does not draw this hard line of separation. Instead, an estimated local false discovery rate value is assigned to each gene expressing the probability of not being differentially expressed. We motivate these opposite ideas in Section 3.1.

The focus of this thesis is on the estimation of the local false discovery rate. Two chapters, that is Chapters 5 and 6, contain two different concepts to improve the estimation. The first contribution is to improve the estimator itself. There exist several approaches to it, yet we aim for a robust and reliable estimator. In Chapter 5 we propose our iteration-based estimator of the local false discovery rate. The procedure works by dividing the set of p-values into two parts. One part represents differentially expressed genes and the other part represents not-differentially expressed genes, that is the background model. From the p-value distributions of these two parts, we derive estimates for the local false discovery rate. The procedure is introduced in detail in Sections 5.2 to 5.4. We investigate the performance of our procedure on exemplary microarray data sets (Section 5.5) as well as on simulated data (Section 5.6). The procedure shows excellent performance in a comprehensive comparison study with thirteen competing methods.

The second contribution to improved estimates of the local false discovery rate is based on a subtle oddity commonly observed in the significance analysis of microarray data: due to highly correlated data, the computation of p-values is often based on an inadequate background model. We motivate the problem in the introduction section of Chapter 6 and show in Section 6.2 that this disadvantageous behavior is common to many biological data sets. We propose a simple but efficient algorithm to extract a valid representation of the background model.

The procedure is explained in Section 6.3. When basing the significance analysis on the valid background, we observe substantial benefits from the improved local false discovery rate estimates (Section 6.4). Both contributions are novel concepts that help to improve every-day's significance analysis of large-scale microarray data.

In summary, the principle outline of this thesis is the introduction to microarray technology and its significance analysis (Chapter 2), theory and estimators of false discovery rate variants (Chapter 3), introduction and exploration of six exemplary data sets (Chapter 4), our proposed estimator of the local false discovery rate (Chapter 5) and our proposed permutation filtering approach (Chapter 6). We conclude with a discussion of the results in Chapter 7.