5 Summary and Discussion

Microarray technologies allow the screening of thousands of genes simultaneously. This opens up new possibilities for diagnostics. In microarray based clinical studies the gene expression data can be used to derive models for diagnosis and prognosis. These models are complementary to traditional methods of laboratory medicine and provide additional information (Raetz and Moos, 2004). For example, gene expression profiling was already successfully applied to further refine diagnosis by improved prediction of treatment benefits (Holleman *et al.*, 2004). Even further distinction of diseases in subclasses with different clinical disease progression was possible (Sørlie *et al.*, 2001).

In this thesis, I have discussed microarray based gene expression studies in the fields of breast cancer, leukemia, lymphoma, and other diseases. However, most of the thousands of screened genes are not related to the disease at all. Typically, only 5-50 genes are sufficient for disease classification (Li, 2005). These genes can be put on a small, custom diagnostic microarray.

The goal of this thesis was to develop a framework for the design of small diagnostic microarrays based on large scale clinical microarray studies. I have proposed to utilize custom diagnostic microarrays, screening only a few diagnostically relevant genes. I have shown that this provides both, a more cost efficient screening method as well as a reliable basis for the development of effective diagnostic classifiers.

My first contribution to the field of diagnostic microarray development is a novel two step design for deriving a diagnostic signature from a whole genome clinical gene expression study. In a first phase the marker panel is determined from a whole genome microarray study with few patients (phase-1). I suggest to use this marker panel instead of the whole genome microarray in screening a larger patient pool. In the second phase the marker genes are fixed but the diagnostic classifier is further fine tuned (phase-2).

In chapter 2, I furthermore introduced a novel evaluation procedure to determine the loss in classification accuracy depending on the number of patients in phase-1 and the size of the marker panel. Increasing the sample size in general also increases the performance of a classifier but with linearly increasing production and handling costs compared to a sub-linear decreasing gain in performance. Thus weighing sample size versus classification accuracy is an important criteria. I showed on five published clinical microarray datasets that in phase-1 as little as 10 patients per arm are sufficient to identify a marker panel of 100 genes that compromises the final performance of the diagnostic classification only marginally. In general, there is an inverse relationship between the number of samples in

phase-1 and the size of the marker panel. Using more samples in phase-1 facilitates the identification of a more reliable set of markers.

My results demonstrate that early marker panel determination is a feasible design for cost efficient clinical studies based on gene expression levels. Since only few genes in phase-2 need to be examined, it is possible to utilize small custom diagnostic microarrays. Furthermore, when there are only few genes to screen it is possible to switch to other technologies like qRT-PCR, in-situ hybridization, or protein panels (Büssow *et al.*, 2001). These technologies may be closer to the clinical phenotype (protein panel) or more precise (qRT-PCR) (Mocellin *et al.*, 2003; Reimers, 2005).

In chapter 3, I addressed the problem of finding a diagnostic signature with good classification performance. Gene selection strategies are typically used to identify the relevant gene set from all genes on a microarray. Gene selection facilitates higher accuracy, faster computation, and by producing a small signature also provides an opportunity to design cheap and efficient diagnostic biomarker panels. I showed that highly correlated genes are selected by ranking genes according to a test-statistic and then choosing the top genes. However, for classification purposes it is better to have distinct but still highly diagnostically informative genes. Therefore, I proposed a novel prefiltering methods that selects relevant but non redundant genes by applying either a correlation based filter or clustering the genes and then choosing from each gene cluster representatives. The methodology is general and can be applied to any univariate gene selection method. I have shown exemplary for five different statistics that prefiltering provides higher accuracy at relatively low computational costs.

In chapter 4, I have finally pointed out severe problems when standard microarray normalization methods are applied to diagnostic microarrays. Nevertheless, a thorough normalization is essential to be able to compare results between diagnostic microarray and achieve reproducible results. Therefore, I have proposed two strategies that work well for diagnostic microarray normalization. I have shown that they can be successfully applied to both, simulated as well as real data experiments. Both strategies select genes for normalization without any additional experimental costs and little computational effort. The first strategy aims at finding genes that are not influenced by the disease type and can therefore be used for normalization. The second strategy aims at finding a balanced gene selection where all genes, even the normalization genes, are informative and thus help to provide a better classification accuracy. This provides a useful tool for accurate normalization. Additionally it allows to use more informative genes one the diagnostic microarray that can be used for both, normalization and classification.

The potential benefits of diagnostic microarrays in oncological studies are huge. They include enriched response rates, reduced sample size, and shorter trial length (Burczynski *et al.*, 2005). There is a substantial economical and social burden associated with unnecessary extensive healthcare and therapeutic treatment that does not provide substantial benefit. Diagnostic microarrays provide an opportunity to enhance the efficacy of therapeutics in a given subpopulation or a subgroup of patients with severe side effects

(Burczynski *et al.*, 2005). Recently, Trastuzumab (Genentech, CA, USA), a drug for the treatment of breast cancer patients, has been approved for therapy on the basis of diagnosis through the expression levels of HER-2 (Harries and Smith, 2002) and others will follow soon (Park *et al.*, 2004).

The challenges in the future will be the proper validation of signatures, and the refinement of the technological diagnosis tools. There already have been efforts to launch a prospective clinical trial using gene expression profiling in DLBCL (TRANSBIG, Tuma (2004)). Ultimately, this leads from research diagnostic tools to everyday patients care and in the end lead to the development of diagnostic microarrays for personalized medicine (Ayers *et al.*, 2004; Slonim, 2001). In the future, gene expression profiling will help to customize therapy for individual patients with unique features of their disease by improved accuracy of diagnosis and prognosis (Raetz and Moos, 2004).

In summary, I have presented new methods for improving biomarker selection, suggested a novel two phase clinical trial for diagnostic microarray development, and proposed an algorithm for assessing the classification performance of diagnostic microarrays with regard to sample size and the number of genes probed. Finally, I developed a novel method for diagnostic microarray normalization that does offer substantial improvement compared to standard microarray normalization techniques. I conclude that the proposed methods offer more insight into the development of diagnostic biomarker panels and will provide valuable tools for future research in diagnostics.