Deriving small diagnostic biomarker panels from genome wide, clinical microarray studies

Jochen Jäger

2006

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Begutachtet von
Prof. Dr. Martin Vingron
Prof. Dr. Patricia Ruiz

Datum der Disputation: 19. Juli 2006

For curiosity

Contents

Acknowledgments **ix**

Preface **xi**

Motivation xi

Thesis structure xii

1 Introduction **1**

1.1 Molecular genetics 1

1.2 Measuring gene expression 4

1.3 Gene expression analysis in clinical studies 9

1.4 Machine learning 13

2 Early marker panel determination (EMPD) **23**

2.1 A subsampling approach to evaluate the effect of EMPD 24

2.2 EMPD results for four gene expression studies 27

2.3 Relation between sample size and number of screened genes 29

2.4 Marker panel variability 33

2.5 Discussion 33

3 Gene list filtering for improved classification **35**

3.1 Reducing redundancy 36

3.2 Correlation based filtering 36

3.3 Clustering based filtering 37

3.4 Applications 38

3.5 Discussion 42

4 Selecting normalization genes for small diagnostic microarrays **45**

4.1 Problems of standard normalization methods for diagnostic chips 46

4.2 Selection of normalization genes 48

4.3 Selection of a balanced signatures 49

4.4 Normalization of small diagnostic microarrays 50

4.5 Results on simulated data 51

4.6 Results on a leukemia study 52

4.7 Discussion 55

5 Summary and Discussion **57**

Bibliography **61**

A Applications:
Expression Analysis Projects **79**
A.1 Cardiomyopathy project 80
A.2 Melanoma project 85

B Zusammenfassung **91**

C Curriculum vitae **93**

List of Figures

1.1 Central dogma of molecular biology 2
1.2 Published microarray articles per year 5
1.3 Comparison of cDNA and oligonucleotide microarrays 7
1.4 Published microarray articles by disease 10
1.5 Support Vector Machine 14
1.6 Nested cross validation 17
1.7 Fuzzy clustering of Notterman data 20

2.1 Pseudo code EMPD 26
2.2 Classification accuracy of EMPD for different sample sizes n_0 28
2.3 Classification accuracy of EMPD for different sample sizes n_0 for four studies 28
2.4 Classification accuracy of EMPD for different marker panel sizes p_0 29
2.5 Classification accuracy of EMPD for different marker panel sizes p_0 for four studies 30
2.6 Classification accuracy trade-off for n_0 versus p_0 30
2.7 Classification accuracy trade-off for n_0 versus p_0 for four studies 31
2.8 Overlap of marker panels for different n_0 33

3.1 Comparison of fuzzy and k-means clustering results 40
3.2 Comparison of different test statistics for prefiltering Alon 41
3.3 AUC for Alon dataset 41
3.4 AUC for Notterman dataset 42
3.5 AUC for Golub dataset 43

4.1 Normalization effect on diagnostic chips 47
4.2 Pseudo code normalization 50
4.3 Characteristics of simulated data 52
4.4 Recovery of original effect using different normalization methods 53
4.5 Loss of effect for different normalization methods 53
4.6 Classification accuracy using different normalization methods 55

A.1 Observed versus expected test statistic plot for cardiomyopathies 83
A.2 EMPD curve for cardiomyopathies 83
A.3 Diagnostic microarray chip normalization methods for the cardiomyopathy dataset 84

A.4 SVM in silico panel for melanoma classification 88
A.5 EMPD curve for melanoma 89
A.6 Diagnostic microarray chip normalization for the melanoma dataset . . . 90

List of Tables

2.1 Gene expression datasets 27
2.2 Classification accuracies with EMPD 32
2.3 Sample size requirements for EMPD 32

3.1 Adenoma expression values 36
3.2 Correlation of Adenoma expression values 37

Abbreviations and Notation

Abbreviations in alphabetical order

A	adenine
ABC	activated B-cell like DLBCL
ALM	acrolentiginous melanoma
AML	acute myelogenous leukemia
aRNA	amplified antisense RNA; also referred to as cRNA
AUC	area under the curve in ROC analysis
BRCA1	breast cancer 1
BRCA2	breast cancer 2
C	cytosine
cDNA	complementary DNA; DNA synthesized from mRNA by RT
CV	cross validation
cRNA	complementary RNA; also referred to as aRNA
Cy3	Cyanine 3-dNTP; fluorescently labeled DNA
Cy5	Cyanine 5-dNTP; fluorescently labeled DNA
DCM	dilated cardiomyopathy
DLBCL	diffuse large B-cell lymphoma
DNA	deoxyribonucleic acid
EMPD	early marker panel determination
ER+	estrogen receptor positive
ER-	estrogen receptor negative
EST	expressed sequence tag
FDR	false discovery rate
G	guanine
GCB	germinal center B-cell like DLBCL
HCM	hypertrophic cardiomyopathy
HNOCM	hypertrophic non-obstructive cardiomyopathy
HOCM	hypertrophic obstructive cardiomyopathy
HTX	heart transplant
HUGO	human genome

ICM	ischemic cardiomyopathy
kNN	k nearest neighbor
LOOCV	leave-one-out cross validation
MLL	mixed-lineage leukemia
mRNA	messenger RNA
NM	nodular melanoma
PCR	polymerase chain reaction
PMK	pulmonary myocarditis
qRT-PCR	quantitative reverse transcriptase polymerase chain reaction
RMA	robust multi-array analysis
RNA	ribonucleic acid
ROC	receiver operator curve
RSD	rheumatic systemic disease
RT	reverse transcription
RT-PCR	reverse transcription-polymerase chain reaction
SRBC	small round blue cell
SSM	superficial spreading melanoma
SVM	support vector machine
`T`	thymine
tRNA	transfer RNA
`U`	uracil
VCM	viral cardiomyopathy
VSN	variance stabilizing normalization

Typesetting and Layout This document was created using VIM (Vi IMproved 6.3) with the LaTeX-Suite extension and LaTeX(pdfeTeX3.141592-1.21a) for the layout and typesetting. If not otherwise noted, the figures were directly exported from Matlab or R as encapsulated postscript (eps) files. Illustrations from the National Human Genome Research Institute (NHGRI) were freely available from the Talking Glossary of Genetics.

Acknowledgments

This work was carried out at the *Computational Molecular Biology department* of the *Max Planck Institute for Molecular Genetics* in Berlin. In fall 2002, I joined the *Computational Diagnostics group* of Rainer Spang there. First, I want to thank my supervisor Rainer Spang for his guidance, his expertise, and his support during my PhD. I also want to thank all present and former colleagues for the great working atmosphere, the helpful discussions, and the fun activities besides work. Special thanks go to my officemate Claudio Lottaz for his devotion to discuss all kind of brain-teasers with me.

The first two years, I was jointly working in the *Molecular analysis of heart failure research group* of Patricia Ruiz and the *Causes and mechanisms of cardiomyopathies research group* at the medical faculty of the university of Heidelberg. I especially thank Patricia Ruiz, Boris Ivandic, Dieter Weichenhan, and Thilo Storm for their support, scientific discussions, and insights into molecular disease mechanisms during that time.

Many thanks also go to my PhD committee members Patricia Ruiz, Martin Vingron, and Jörg Schultz for their counseling and advice. Stefanie Scheid, Florian Markowetz, and Stefan Bentink read early drafts of the thesis and helped me to improve it by their comments. Thanks a lot.

The idea for the gene selection filtering, introduced in chapter 3, arose during the last months of my two year stay at the University of Washington. It has been a very enjoyable experience abroad and I would like to thank all friends and colleagues there. Special thanks go to Larry Ruzzo, Rimli Sengupta, and Martin Tompa for their support, advice, and fruitful discussions.

Lastly, I am grateful for my old friends with whom I stayed in contact during my PhD as well as the new friends that I met during my time in Berlin. They provided the environment for a great experience in Berlin and the necessary distraction from the inevitable phases of frustration during research. Above all, I would like to thank my parents for their understanding and relentless support.

Jochen Jäger Berlin, 13. Juni 2006

Preface

Motivation

A reliable and precise diagnosis of a disease is essential to make suitable therapy decisions. This holds especially for cancer, which is the second frequent cause of death in the western world (Hoyert *et al.*, 2005). However, established diagnosis strategies are limited in distinguishing between morphologically similar but molecularly different tumors (Schmidt and Begley, 2003). On the other hand, these molecular differences are crucial in predicting the response to therapy and ultimately the outcome for the patient (Schmidt and Begley, 2003).

One way to analyze molecular differences is gene expression profiling. Here, the expression level of genes in different cells are measured and compared. Currently, it is possible to measure thousands of genes in parallel with a high-throughput method called microarrays. Looking at gene expression levels allows a detailed insight into so far invisible changes of the metabolism. This leads to a more complete understanding of the underlying mechanisms of the disease and a more reliable diagnosis (Roepman *et al.*, 2005; Ciro *et al.*, 2003). Therefore, Barrett (2005) predicts considerable implications for medicine. In the end, gene expression profiling might even revise the definition of diseases (Alizadeh *et al.*, 2000).

So far, mostly whole genome microarrays, measuring all genes of the genome, are used. Synthesizing the necessary PCR primers for such a large number of genes increases production costs drastically (Fernandes and Skiena, 2002). However, usually the measurements of 5-100 genes are adequate to build a classifier that distinguishes one disease subtype from another (Li and Yang, 2002). Therefore, for diagnosis it is not necessary to screen gene expression on a whole genome basis but instead customized microarrays with considerably less genes can be used. This eases handling, production, and data analysis.

Throughout this thesis, I refer to diagnostic microarrays as small custom microarrays, holding only few genes. Whole genome microarrays are referred to as genomewide gene expression microarrays, holding tens of thousands of genes.

Thesis Structure

In this thesis, I discuss several problems related to the design of small diagnostic microarrays. Currently, whole genome microarrays are frequently used in clinical trials that aim for diagnostics. Instead of using whole genome microarrays for all patients I propose to screen only a small fraction of the patients with them. This serves the purpose of finding disease relevant genes for diagnosis. Then, I suggest to switch to small diagnostic microarrays carrying these genes. The diagnostic microarrays are now used to screen a larger patient pool. Here, the goal is to fine tune a gene signature that provides accurate diagnosis. In detail, I address the following three questions that arise during the development of a diagnostic microarray:

1. **Accuracy loss of a diagnostic microarray** – What is the loss in classification accuracy when a diagnostic microarray is determined in the early onset of a clinical whole genome microarray study?
   In chapter 2, I present a novel, two-phase design for predictive clinical gene expression studies: early marker panel determination (EMPD). In phase-1, genome-wide microarrays are only used for a small number of individual patient samples. From this phase-1 data a panel of marker genes is derived. The marker genes are used for the design of a custom, diagnostic microarray. In phase-2, whole genome microarray are exchanged by this diagnostic microarray. Then, only the expression of the genes on this diagnostic microarray are measured for a large group of patients. From this data a predictive classification model is learned. Phase-2 does not require the use of whole genome microarrays, thus making EMPD a cost efficient alternative for current trials. Currently, a whole genome Affymetrix array (HGU 133 Plus 2.0) retails for US$975, whereas a custom express array from the same company costs 375 US$ (Affymetrix retail price sheet Jan 2006). The expected performance loss of EMPD is compared to designs that use genome-wide microarrays for all patients. I also examine the trade-off between the number of patients included in phase-1 and the number of marker genes required in phase-2. By analysis of five published datasets, I find that in these studies already 16 patients per group would have been sufficient to determine a suitable marker panel of 10 genes, and that this early decision compromises the final performance only marginally.

2. **Gene selection** – Which genes should be included in a diagnostic signature?
   In chapter 3, I derive a method for improving univariate gene selection techniques for diagnosis of diseases using microarray data. Genes of interest are typically selected by ranking genes according to a test score and then choosing the top genes. I show that using highly discriminative genes that are less correlated amongst each other instead of just choosing the top ranking genes achieves better classification accuracy. I propose three different pre-filter methods to retrieve groups of genes that have a similar gene expression profile. Two are based on clustering and one is based on correlation. For these groups, I apply a score to finally select genes of interest.

I show that the filtered set of genes can be used to significantly improve existing classifiers.

3. **Normalization** – How can a diagnostic microarray be normalized?
   In chapter 4, I show that applying standard microarray normalization strategies to diagnostic microarrays results in decreased classification accuracy. The reason for this is that normalization of gene expression microarrays carrying thousands of genes has strong assumptions: either that some genes are constantly expressed or that the average of all genes is not altered by the disease conditions. This does not hold for diagnostic microarrays carrying exclusively discriminative genes. I point out the differences of normalization between whole genome and diagnostic microarrays and suggest two normalization strategies especially designed for diagnostic microarrays. The first is a data driven selection of additional normalization genes. The second does not need additional genes. Instead it is based on finding a balanced diagnostic signature. I compare both methods to standard normalization protocols known from whole genome microarrays. The use of the latter leads to a loss of diagnostic prediction accuracy, while the two normalization strategies designed for diagnostic microarrays achieve better results.

In the introductory chapter 1, I highlight the potential use of microarray profiling for diagnostics. First, I review the underlying principles of gene expression profiling by providing a basic introduction into molecular genetics and technologies for measuring gene expression. Then, current results of clinical gene expression studies from various diseases are reported. Since I derive diagnostic disease classifiers from microarray data, I shortly outline machine learning approaches, especially classification and clustering. Finally, I introduce evaluation strategies for assessing the performance on future samples. The thesis closes with a summary and an outlook. In the appendix, I briefly report on five gene expression studies I analyzed during the last 4 years. Two of the studies, namely the cardiomyopathy and the melanoma project, are discussed in more detail.