

Novel Concepts for the Significance Analysis of Microarray Data

Stefanie Christina Scheid

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

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Reinhard Meister

Tag der Disputation: 27. Oktober 2006

Preface

Acknowledgments. The projects leading to this thesis were carried out at the Computational Diagnostics Group of the Department for Computational Molecular Biology at the Max Planck Institute for Molecular Genetics. First and foremost, I thank *Rainer Spang* for giving me the opportunity to work and learn in the CompDiag Group, and for supervising this thesis.

I cordially thank the CompDiag members *Juby Jacob*, *Stefan Bentink*, *Jochen Jäger*, *Dennis Kostka*, *Claudio Lottaz*, and *Florian Markowitz* for the pleasant working atmosphere in our group and for many fruitful and inspiring scientific (and non-scientific) discussions. I am also grateful for the bunch of bug reports and special requests, which let *twilight* grow to a proper package. I thank all former and current members of the CMB Department for the nice working atmosphere, for scientific discussions and important as well, for many enjoyable after-work activities.

Special thanks go to *Anja von Heydebreck*, *Ulrich Mansmann*, and *Martin Vingron* for monitoring and supportively advising my projects as members of my PhD Committee, and to *Reinhard Meister* for refereeing this PhD thesis. I thank *Per Broberg* for discussions on π_0 estimators and suggestions, which lead to improvements of the software package.

My deepest thanks go to my family and Toralf for their true love, care, and support.

Related publications. The main parts of the thesis are based on three published papers:

1. Scheid S and Spang R (2004). A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE Transactions on Computational Biology and Bioinformatics* **1**(3):98–108.
2. Scheid S and Spang R (2005). twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics* **21**(12):2921–2922.
3. Scheid S and Spang R (2006). Permutation filtering: A novel concept for significance analysis of large-scale genomic data. In: Apostolico A, Guerra C, Istrail S, Pevzner P, and Waterman M (Eds.), *Research in Computational Molecular Biology: 10th Annual International Conference, Proceedings of RECOMB 2006, Venice, Italy, April 2-5, 2006*, Lecture Notes in Computer Science vol. 3909, Springer, Heidelberg, pp. 338–347.

Software. The computational parts of the thesis were done entirely using R 2.1.0, which is freely available on <http://www.r-project.org>. The cited packages are freely available from the R archive or from the Bioconductor project on <http://www.bioconductor.org>. We used our Bioconductor package *twilight* version 1.5.1 for the computational parts about estimation of local false discovery rates and permutation filtering. For some calculations, the original functions were extended to print out internally generated data.

Notation and abbreviations

Notation	Definition
ALL	Acute lymphoblastic leukemia.
cdf	Cumulative distribution function.
cFDR	Conditional false discovery rate.
FDR	False discovery rate.
fd _r	Local false discovery rate.
iid	Independently and identically distributed.
pFDR	Positive false discovery rate.
MDS	Multi-dimensional scaling.
SEP	Successive exclusion procedure.
B	Number of permutations.
\mathcal{C}	Set of permutations of the observed class-label vector c_0 .
c	Class-label vector of length n .
m	Number of genes.
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2 .
n	Number of patients/samples.
\mathbf{P}	P-value matrix of dimension $m \times (B + 1)$.
\mathbf{p}	P-value vector of length m .
π_0	Prior probability of non-differential expression; equals the overall percentage of non-induced genes.
\mathbf{S}	Score matrix of dimension $m \times (B + 1)$.
U_c	Function mapping a vector of class labels c to a vector of p-values \mathbf{p} based on the set of permutations \mathcal{C} : $U_c(c) = \mathbf{p}$.
\mathbf{X}	Gene-expression matrix of dimension $m \times n$.

Contents

Preface	i
Notation and abbreviations	iii
1 Motivation and outline	1
2 Introduction to microarray data	5
2.1 Outline	5
2.2 Microarray data	5
2.3 Preprocessing microarray data	7
2.4 Assessing differential gene expression	10
3 A review on false discovery rates	19
3.1 Introduction and outline	19
3.2 The false discovery rate	21
3.3 Pitfalls of global false discovery rates	26
3.4 A local measure of significance	27
3.5 Estimating the proportion of non-induced genes	33
4 Exemplary data sets	47
4.1 Outline	47
4.2 Six microarray comparisons on cancer	47
4.3 Exploring differential expression	50
4.4 Comparison of pooled and gene-wise p-values	53
5 A novel estimator of the local false discovery rate	59
5.1 Outline	59
5.2 A stochastic downhill search approach	59
5.3 Calibration of the regularization parameter	63

5.4	Fine-tuning	65
5.5	Features and applications	66
5.6	Simulation and results	68
5.7	Implementation and runtime evaluation	81
6	Permutation filtering	87
6.1	Introduction and outline	87
6.2	Artifacts in real data	90
6.3	A stochastic filtering approach	95
6.4	Benefits of filtering	97
7	Summary and discussion	103
	Bibliography	113
A	Zusammenfassung	123
B	Curriculum Vitæ	125