Thesis summary and discussion

In this thesis I presented computational methodology for exploring and communicating molecular characteristics of disease. My primary contributions are the *documentation by value strategy* and putting forward the concept of *differentially coexpressed groups of genes*. Both are novel approaches, the first leading to significantly more consistent (molecular) diagnoses and the latter complementing established exploratory analysis tools. Data underlying the analyses were always taken from clinical gene expression studies and generated by oligonucleotide microarrays. Each array provides a reproducible image of a tissue sample at molecular resolution and reflects the mRNA abundance of genes. The tissue samples we consider are connected to certain instances of diseases (e.g. cancer types), and reflections of disease types in the data provide a characterization of disease on a molecular level.

In the first chapter I provide a review of methodology currently used to *predict* disease type on the basis of microarray data. The underlying assumption is that the type of disease is reflected in expression data, and the methods in use try to find the traces. The paradigm is that of statistical learning: The measured data and the corresponding disease labels are understood as random samples of a disease population, governed by a common distribution function. All the approaches I discuss use these example data to infer a *classification rule*, i.e. instructions how to decode disease type from future measurements. In case such a method works reliably, it must make use of structure present in all data and characteristic for the type of disease it predicts. Therefore a classification rule is also called *molecular signature*. In that sense, the first chapter is on state of the art characterization of disease. There is no shortage of sophisticated algorithms to infer molecular signatures; in all of them regularization plays a crucial role. Note though, that virtually all regularizing methods are "generic" in the sense that they do not reflect the specific nature of the task. We hypothesize that implementing regularization additionally based on prior *biological* knowledge might improve methodology. Currently, such information is rather utilized to sanity-check molecular signatures inferred by more generic methods.

The second chapter deals with the communication of molecular signatures between scientists. As microarray data is not measured on an absolute scale, a scale is estimated from the study data itself. Molecular signatures derived from such *preprocessed* examples depend on the estimated scale. I introduce methodology for communicating, and therefore *documenting*, molecular signatures. The key point is: Not only the molecular

signature has to be documented, but also the scale inferred by preprocessing the study data. Along these lines I provide an implementation of a scale-preserving documentation that complements two popular preprocessing schemes. This is termed *documentation by value*, and an R-language add-on package is available (see Appendix). Eight clinical data sets are used to demonstrate that "proper" documentation of signatures significantly reduces the ambiguity of diagnoses as compared to standard methods. The point of including preprocessing information in signature documentation has, to the best of my knowledge, not been addressed before. The method I provide enables honest assessment of the performance of a signature by the means of evaluation on patients not enrolled in the signature-deriving study. Such testing is crucial to show that molecular signatures bear information complementing existing markers of disease. This, in turn, is a prerequisite for impact on clinical practice. But questions remain, for instance about the choice of preprocessing scheme. This is difficult to answer in general and comparisons are generally done in terms of accuracy and bias (e.g. [16, 75, 102]) in the spirit of calibrating a delicate measuring device. With the methodology provided it is possible to include a *notion of robustness* in comparisons of preprocessing schemes. I do so by comparing the stability of diagnoses between three preprocessing schemes in a resampling experiment (Chapter 2). Nevertheless a thorough comparison comprising more schemes and especially a concept for relating results to other quality measures would be of great value.

In the third chapter I provide an algorithm to explore molecular characteristics of disease. The underlying assumption is that different disease types arise due to differences in the regulation of genes. The `dcoex` algorithm is designed to find groups of genes that are under a common regulatory control in one type of disease, but this control is lost in another type. As coregulation cannot be measured on microarrays, the algorithm looks for *groups of differentially coexpressed genes*. That is, `dcoex` reveals groups of genes coexpressed in one type of disease, whereas the coexpression is lost in another disease type also part of the comparison. We were the first to suggest such an analysis strategy [85], but the approach has been taken up by other researchers, e.g. in [108]. In an application to childhood leukemia, I find a group of genes associated with the ubiquitin-proteasome pathway to be deregulated in children bearing the philadelphia chromosome. The results are shown to be robust and comply to biological reason (Chapter 3). As with all methods for exploratory data analysis, `dcoex` cannot be expected to yield one unique result that is either correct or incorrect. It is rather like looking at data from a new angle and adding a piece to the puzzle of molecular reflections of disease types. Nevertheless, ruling out that results are chance artifacts in indispensable. Unfortunately the permutation procedure employed is slow, and an analytical description of the score distribution under a suitable null hypothesis is greatly desirable. Also, in the context of generating biological hypotheses, *integration of additional information* seems promising. Utilizing annotations regarding the binding of known regulatory elements to promoter regions of genes comes to mind. Such data is available from in silico sequence-based predictions, or as results of ChIP-chip experiments. Both, integration

in the search algorithm as well as utilization for result-filtering are conceivable.

Overall, I introduced two conceptually new methods for the analysis of microarray data. The documentation by value strategy enables the exchange of unambiguous molecular signatures in the scientific community; this facilitates the way of molecular classification into clinical practice. The concept of differential coexpression advances exploratory methodology available for microarray analysis; this improves hypothesizing about disease mechanisms. While I discussed everything in a clinical-diagnostics setting, the approaches generalize to all fields of research employing microarrays as measuring devices. Concerning the reliable diagnosis of disease, algorithmic methodology is well developed and further improvements might as well be achieved by careful study design and advances in measuring technology. In the case of exploring molecular characteristics of disease things look different. While careful planning of experiments also plays a key role, this more comprising area provides conceptually challenging tasks that will benefit from further methodological improvements.

Bibliography

[1] Affymetrix web-site (www.affymetrix.com).

[2] N. Abe, B. Zadrozny, and J. Langford. Outlier detection by active learning. In *to appear in The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[3] Julian Adams. The proteasome: structure, function, and role in the cell. *Cancer Treat Rev*, 29 Suppl 1:3–9, May 2003.

[4] G. Alexe, S. Alexe, D. E. Axelrod, P. L. Hammer, and D. Weissmann. Logical analysis of diffuse large b-cell lymphomas. *Artif Intell Med*, 34(3):235–267, Jul 2005.

[5] Dominic J Allocco, Isaac S Kohane, and Atul J Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5:18, Feb 2004.

[6] C Ambroise and GJ McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, 99(10):6562–6566, Arpil 2002.

[7] Fabrizio Angiulli, Stefano Basta, and Clara Pizzuti. Distance-based detection and prediction of outliers. *IEEE Transactions on Knowledge and Data Engineering*, 18(2):145–160, 2006.

[8] Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41–47, Jan 2002.

[9] Daniel Barbara, Carlotta Domeniconi, and James Rogers. Detecting outliers using transduction and statistical significance testing. In *to appear in The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[10] David G Beer, Sharon LR Kardia, Chiang-Ching Huang, Thomas J Giordano, Albert M Levin, David E Misek, Lin Lin, Guoan Chen, Tarek G Gharib, Dafydd G Thomas, Michelle L Lizyness, Rork Kuick, Satoru Hayasaka, Jeremy MG Taylor, Mark D Iannettoni, Mark B Orringer, and Samir Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, 8(8):816–24, Aug 2002.

[11] Y Bengio and Y Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *J. Mach. Learn. Res.*, 5(Sep):1089–1105, 2004.

[12] A Bhattacharjee, WG Richards, J Staunton, C Li, S Monti, P Vasa, C Ladd, J Beheshti, R Bueno, M Gillette, M Loda, G Weber, EJ Mark, ES Lander, W Wong, BE Johnson, TR Golub, DJ Sugarbaker, and M Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, 98(24):13790–5, Nov 2001.

[13] Elia Biganzoli, Nicola Lama, Federico Ambrogi, Laura Antolini, and Patrizia Boracchi. Prediction of cancer outcome with microarrays. *Lancet*, 365(9472):1683; author reply 1684–5, 2005.

[14] Andrea H Bild, Guang Yao, Jeffrey T Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M Lancaster, Andrew Berchuck, John A Olson, Jeffrey R Marks, Holly K Dressman, Mike West, and Joseph R Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353–7, Jan 2006.

[15] Trond Bø and Inge Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biol*, 3(4):RESEARCH0017, 2002.

[16] BM Bolstad, RA Irizarry, M Astrand, and TP Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93, Jan 2003.

[17] Endre Boros and Peter L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 123(1-3):155–225, 2002.

[18] UM Braga-Neto and ER Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–80, Feb 2004.

[19] L Breiman. Bagging predictors. *Machine Learning*, 24(2):123 – 140, 1996.

[20] L Breiman. Random forests. *Machine Learning*, 45(1):5 – 32, 2001.

[21] L Breiman, J Friedman, R Olshen, and C Stone. *Classification and regression trees*. Wadsworth, 1984.

[22] MP Brown, WN Grundy, D Lin, N Cristianini, CW Sugnet, TS Furey, M Ares, and D Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–7, Jan 2000.

[23] Ching-Ter Chang. On the polynomial mixed 0-1 fractional programming problems. *European Journal of Operational Research*, 131(1):224–227, 2001.

[24] Y Cheng and GM Church. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, 8:93–103, 2000.

[25] MH Cheok, W Yang, CH Pui, JR Downing, C Cheng, CW Naeve, MV Relling, and WE Evans. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat Genet*, 34(1):85–90, May 2003.

[26] Sabina Chiaretti, Xiaochun Li, Robert Gentleman, Antonella Vitale, Marco Vignetti, Franco Mandelli, Jerome Ritz, and Robin Foa. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–2778, Apr 2004.

[27] R Chiarle, LM Budel, J Skolnik, G Frizzera, M Chilosi, A Corato, G Pizzolo, J Magidson, A Montagnoli, M Pagano, B Maes, C De Wolf-Peeters, and G Inghirami. Increased proteasome degradation of cyclin-dependent kinase inhibitor p27 is associated with a decreased overall survival in mantle cell lymphoma. *Blood*, 95(2):619–26, Jan 2000.

[28] JH Cho, D Lee, JH Park, and IB Lee. Gene selection and classification from microarray data using kernel machine. *FEBS Lett*, 571(1-3):93–8, Jul 2004.

[29] J. Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20:37–46, 1960.

[30] Leslie M Cope, Rafael A Irizarry, Harris A Jaffee, Zhijin Wu, and Terence P Speed. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20(3):323–31, Feb 2004.

[31] Marcel Dettling. BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–93, Dec 2004.

[32] Marcel Dettling and Peter Bühlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, 19(9):1061–9, Jun 2003.

[33] Marcel Dettling and Peter Bühlmann. Supervised clustering of genes. *Genome Biol*, 3(12):RESEARCH0069, 2002.

[34] L. Devroye, L. Gyorfi, and G Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, New York, 1996.

[35] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression dat. In *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB'03)*, 2003.

[36] RO Duda, PE Hart, and DG Stork. *Pattern classification*. Wiley, New York, 2001.

[37] S Dudoit, J Fridlyand, and TP Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, 97(457):77–87, 2002.

[38] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl 1:S105–S110, 2002.

[39] Blythe P Durbin and David M Rocke. Variance-stabilizing transformations for two-color microarrays. *Bioinformatics*, 20(5):660–7, Mar 2004.

[40] Ramón Díaz-Uriarte and Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006.

[41] Patrik Edén, Cecilia Ritz, Carsten Rose, Mårten Fernö, and Carsten Peterson. "good old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer*, 40(12):1837–1841, Aug 2004.

[42] B Efron and R Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

[43] PH Eilers, JC van Houwelingen, and JM Boer. Classification of microarray data with penalized logistic regression. In *Proceedings of SPIE*, volume 4266, 2000.

[44] L Ein-Dor, I Kela, G Getz, D Givol, and E Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–8, Jan 2005.

[45] MB Eisen, PT Spellman, PD Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Aced. Sci.*, 95(25):14863–8, dec 1998.

[46] Theodoros Evgeniou, Tomaso Poggio, Massimiliano Pontil, and Alessandro Verri. Regularization and statistical learning theory for data analysis. *Comput. Stat. Data Anal.*, 38(4):421–432, 2002.

[47] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

[48] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[49] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–374, 2000.

[50] Naoya Fujita, Saori Sato, Kazuhiro Katayama, and Takashi Tsuruo. Akt-dependent phosphorylation of p27Kip1 promotes binding to 14-3-3 and cytoplasmic localization. *J Biol Chem*, 277(32):28706–13, Aug 2002.

[51] TS Furey, N Cristianini, N Duffy, DW Bednarski, M Schummer, and D Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–14, Oct 2000.

[52] Laurent Gautier, Rafael Irizarry, Leslie Cope, and Ben Bolstad. Bioconductor vignette: Description of affy.

[53] S Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

[54] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang, and Jianhua Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[55] TR Golub, DK Slonim, P Tamayo, C Huard, M Gaasenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, MA Caligiuri, CD Bloomfield, and ES Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, Oct 1999.

[56] Gavin J Gordon, Roderick V Jensen, Li-Li Hsiao, Steven R Gullans, Joshua E Blumenstock, William G Richards, Michael T Jaklitsch, David J Sugarbaker, and Raphael Bueno. Using gene expression ratios to predict outcome among patients with mesothelioma. *J Natl Cancer Inst*, 95(8):598–605, Apr 2003.

[57] Erik C Gunther, David J Stone, Robert W Gerwien, Patricia Bento, and Melvyn P Heyes. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc Natl Acad Sci U S A*, 100(16):9608–13, Aug 2003.

[58] I Guyon and A Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Reasearch*, 3(mar):1157 – 1182, 2003.

[59] I Guyon, J Weston, S Barnhill, and V Vapnik. Gene selection for cancer classification. *Machine Learning*, 46(1-3):389–422, 2002.

[60] MA Hall and LA Smith. Feature subset selection: a correlation based filter approach. In N Kasabov and et al., editors, *Proc Fourth International Conference on Neural Information Processing and Intelligent Information Systems*, pages 855–858, 1997.

[61] Bettina Harr and Christian Schlötterer. Comparison of algorithms for the analysis of affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Res*, 34(2):e8, 2006.

[62] T Hastie, R Tibshirani, and J Friedman. *The elements of statistical learning*. Springer, New York, 2001.

[63] Anne-Mette K Hein, Sylvia Richardson, Helen C Causton, Graeme K Ambler, and Peter J Green. Bgx: a fully bayesian integrated approach to the analysis of affymetrix genechip data. *Biostatistics*, 6(3):349–373, Jul 2005.

[64] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[65] E Huang, SH Cheng, H Dressman, J Pittman, MH Tsou, CF Horng, A Bild, ES Iversen, M Liao, CM Chen, M West, JR Nevins, and AT Huang. Gene expression predictors of breast cancer outcomes. *Lancet*, 361(9369):1590–6, May 2003.

[66] Xiaohong Huang and Wei Pan. Linear regression and two-class classification with gene expression data. *Bioinformatics*, 19(16):2072–8, Nov 2003.

[67] W Huber, A von Heydebreck, H Sültmann, A Poustka, and M Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1):Art 3, 2003.

[68] Wolfgang Huber, Anja von Heydebreck, Holger S?ltmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

[69] Michael Hummel, Stefan Bentink, Hilmar Berger, Wolfram Klapper, Swen Wessendorf, Thomas F E Barth, Heinz-Wolfram Bernd, Sergio B Cogliatti, Judith Dierlamm, Alfred C Feller, Martin-Leo Hansmann, Eugenia Haralambieva, Lana Harder, Dirk Hasenclever, Michael Kühn, Dido Lenze, Peter Lichter, Jose Ignacio Martin-Subero, Peter Möller, Hans-Konrad Müller-Hermelink, German Ott, Reza M Parwaresch, Christiane Pott, Andreas Rosenwald, Maciej Rosolowski, Carsten Schwaenen, Benjamin Stürzenhofecker, Monika Szczepanowski, Heiko Trautmann, Hans-Heinrich Wacker, Rainer Spang, Markus Loeffler, Lorenz Trümper, Harald Stein, Reiner Siebert, and Molecular Mechanisms in Malignant Lymphomas Network Project of the Deutsche Krebshilfe. A biologic definition of burkitt's lymphoma from transcriptional and genomic profiling. *N Engl J Med*, 354(23):2419–2430, Jun 2006.

[70] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[71] Masato Inoue, Shin-Ichi Nishimura, Gen Hori, Hiroyuki Nakahara, Michiko Saito, Yoshihiro Yoshihara, and Shun-Ichi Amari. Improved parameter estimation for variance-stabilizing transformation of gene-expression microarray data. *J Bioinform Comput Biol*, 2(4):669–679, Dec 2004.

[72] R Irizarry, B Hobbes, F Collin, and T Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249?262, 2003.

[73] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed4. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31, 2003.

[74] Rafael A Irizarry, Daniel Warren, Forrest Spencer, Irene F Kim, Shyam Biswal, Bryan C Frank, Edward Gabrielson, Joe G N Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C Hilmer, Eric Hoffman, Anne E Jedlicka, Ernest Kawasaki, Francisco Martínez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye, and Wayne Yu. Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2(5):345–50, May 2005.

[75] Rafael A Irizarry, Zhijin Wu, and Harris A Jaffee. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, Jan 2006.

[76] J. Jaeger, R. Sengupta, and WL. Ruzzo. Improved gene selection for classification of microarrays. In *Pacific Symposium on Biocomputing 2003*, pages 53–64. Pac Symp Biocomput, World Scientific, 2003.

[77] Jochen Jaeger, Dieter Weichenhan, Boris Ivandic, and Rainer Spa ng. Early diagnostic marker panel determination for microarray based clini cal studies. *Statistical Applications in Genomics and Molecular Biology, Berkeley electronic press*, page in press, 2005.

[78] George H. John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *International Conference on Machine Learning*, pages 121–129, 1994. Journal version in AIJ, available at http://citeseer.nj.nec.com/13663.html.

[79] M. I. Jordan. Why the logistic function? a tutorial discussion on probabilities and neural networks. Computational cognitive science report 9503, MIT, August 1995.

[80] L Kaderali and A Schliep. Selecting signature oligonucleotides to identify organisms using dna arrays. *Bioinformatics*, 18(10):1340–9, Oct 2002.

[81] Kazuhiro Katayama, Naoya Fujita, and Takashi Tsuruo. Akt/protein kinase B-dependent phosphorylation and inactivation of WEE1Hu promote cell cycle progression at G2/M transition. *Mol Cell Biol*, 25(13):5725–37, Jul 2005.

[82] Thomas B Kepler, Lynn Crosby, and Kevin T Morgan. Normalization and analysis of dna microarray data by self-consistency and local regression. *Genome Biol*, 3(7):RESEARCH0037, Jun 2002.

[83] Yongdai Kim and Jinseog Kim. Gradient lasso for feature selection. In *ICML '04: Twenty-first international conference on Machine learning*, New York, NY, USA, 2004. ACM Press.

[84] R Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1145, 1995.

[85] Dennis Kostka and Rainer Spang. Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20 Suppl 1:I194–I199, Aug 2004.

[86] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

[87] JW Lee, JB Lee, M Park, and SH Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, april 2005.

[88] Cheng Li and Wing Hung Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98:31?36, 2001.

[89] B. W. Lindgren. *Statistical Theory*. Chapman & Hall, 1993.

[90] Carl N. Yoshizawa Loic and Le Marchand. Re: Misinterpretation and misuse of the kappa statistic. *Am. J. Epidemiol.*, 128(5):1179–1180, 1988.

[91] M. Maclure and W. C. Willett. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol*, 126(2):161–169, Aug 1987.

[92] Malcolm Maclure and Walter C. Willett. THE AUTHORS REPLY. *Am. J. Epidemiol.*, 128(5):1180–1181, 1988.

[93] SC Madeira and A L Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1), Jan-Mar 2004.

[94] Michael Z. Man, Greg Dyson, Kjell Johnson, and Birong Liao. Evaluating methods for classifying expression data. *Journal of Biopharmaceutical Statistics*, 14(4):1065 – 1084, 2004.

[95] P Masdehors, H Merle-Béral, H Magdelénat, and J Delic. Ubiquitin-proteasome system and increased sensitivity of B-CLL lymphocytes to apoptotic death activation. *Leuk Lymphoma*, 38(5-6):499–504, Aug 2000.

[96] P Masdehors, H Merle-Béral, K Maloum, S Omura, H Magdelénat, and J Delic. Deregulation of the ubiquitin system and p53 proteolysis modify the apoptotic response in B-CLL lymphocytes. *Blood*, 96(1):269–74, Jul 2000.

[97] P Masdehors, S Omura, H Merle-B?ral, F Mentz, JM Cosset, J Dumont, H Magdel?nat, and J Delic. Increased sensitivity of CLL-derived lymphocytes to apoptotic death activation by the proteasome-specific inhibitor lactacystin. *Br J Haematol*, 105(3):752–7, Jun 1999.

[98] GJ McLachlan, KA Do, and C Ambroise. *Analyzing Microarray Gene Expression Data*. Wiley, 2004.

[99] Stefan Michiels, Serge Koscielny, and Catherine Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–92, 2005.

[100] Frank F Millenaar, John Okyere, Sean T May, Martijn van Zanten, Laurentius A C J Voesenek, and Anton J M Peeters. How to decide? different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics*, 7:137, 2006.

[101] S. A. Murphy and A. W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

[102] Felix Naef, Nicholas D Socci, and Marcelo Magnasco. A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics*, 19(2):178–184, Jan 2003.

[103] EE Ntzani and JPA Ioannidis. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet*, 362(9394):1439–44, Nov 2003.

[104] Catherine L Nutt, D. R. Mani, Rebecca A Betensky, Pablo Tamayo, J. Gregory Cairncross, Christine Ladd, Ute Pohl, Christian Hartmann, Margaret E McLaughlin, Tracy T Batchelor, Peter M Black, Andreas von Deimling, Scott L Pomeroy, Todd R Golub, and David N Louis. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, 63(7):1602–1607, Apr 2003.

[105] M Osborne, R. Presnell, B Turlach, and A Berwin. On the lasso and its dual. *Journal of Computational & Graphical Statistics*, 2000.

[106] PJ. Park, M. Pagano, and M. Bonetti. A nonparametric scoring algorithm for identifying informative genes from micr oarray data. *Pac Symp Biocomput*, pages 52–63, 2001.

[107] SL Pomeroy, P Tamayo, M Gaasenbeek, LM Sturla, M Angelo, ME McLaughlin, JY Kim, LC Goumnerova, PM Black, C Lau, JC Allen, D Zagzag, JM Olson, T Curran, C Wetmore, JA Biegel, T Poggio, S Mukherjee, R Rifkin, A Califano, G Stolovitzky, DN Louis, JP Mesirov, ES Lander, and TR Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–42, 2002.

[108] C. Prieto, M. J. Rivas, J. M. Sánchez, J. López-Fidalgo, and J. De Las Rivas. Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. *Bioinformatics*, 22(9):1103–1110, May 2006.

[109] S Rahmann. Rapid large-scale oligonucleotide selection for microarrays. *Proc IEEE Comput Soc Bioinform Conf*, 1:54–63, 2002.

[110] A Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research*, 3(Mar):1357–1370, 2003.

[111] S Ramaswamy, P Tamayo, R Rifkin, S Mukherjee, CH Yeang, M Angelo, C Ladd, M Reich, E Latulippe, JP Mesirov, T Poggio, W Gerald, M Loda, ES Lander, and TR Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, 98(26):15149–54, Dec 2001.

[112] Sridhar Ramaswamy and Charles M Perou. DNA microarrays in breast cancer: the promise of personalised medicine. *Lancet*, 361(9369):1576–7, May 2003.

[113] Sridhar Ramaswamy, Ken N Ross, Eric S Lander, and Todd R Golub. A molecular signature of metastasis in primary solid tumors. *Nat Genet*, 33(1):49–54, Jan 2003.

[114] DF Ransohoff. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer*, 4(4):309–14, Apr 2004.

[115] BD Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996.

[116] David M Rocke and Blythe Durbin. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, 19(8):966–972, May 2003.

[117] DM Rocke and B Durbin. A model for measurement error for gene expression arrays. *J Comput Biol*, 8(6):557–69, 2001.

[118] Mary E Ross, Xiaodong Zhou, Guangchun Song, Sheila A Shurtleff, Kevin Girtman, W Kent Williams, Hsi-Che Liu, Rami Mahfouz, Susana C Raimondi, Noel Lenny, Anami Patel, and James R Downing. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102(8):2951–2959, 2003.

[119] V Roth. The generalized lasso. *IEEE Transactions on Neural Networks*, 15(1):16–28, 2004.

[120] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[121] Peter J. Rousseeuw and Katrien Driessen. Computing lts regression for large data sets. *Data Min. Knowl. Discov.*, 12(1):29–45, 2006.

[122] M Ruschhaupt, W Huber, A Poustka, and U Mansmann. A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

[123] B Schoelkopf and AJ Smola. *Learning with kernels*. MIT Press, Cambride (MA), 2001.

[124] Hans Schwerdtfeger. *Introduction to linear algebra and the theory of matrices*. Groningen, P. Noordhoff, 1950.

[125] Ruty Shai, Tao Shi, Thomas J Kremen, Steve Horvath, Linda M Liau, Timothy F Cloughesy, Paul S Mischel, and Stanley F Nelson. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene*, 22(31):4918–4923, Jul 2003.

[126] SK Shevade and SS Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–53, Nov 2003.

[127] Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo CT Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S Pinkus, Tane S Ray, Margaret A Koval, Kim W Last, Andrew Norton, T Andrew Lister, Jill Mesirov, Donna S Neuberg, Eric S Lander, Jon C Aster, and Todd R Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[128] R Simon. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br J Cancer*, 89(9):1599–604, Nov 2003.

[129] R Simon, MD Radmacher, K Dobbin, and LM McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, 95(1):14–8, Jan 2003.

[130] Richard Simon. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol*, 23(29):7332–41, Oct 2005.

[131] Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, Eric S Lander, Massimo Loda, Philip W Kantoff, Todd R Golub, and William R Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, Mar 2002.

[132] Therese Sorlie, Robert Tibshirani, Joel Parker, Trevor Hastie, J. S. Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, Janos Demeter, Charles M Perou, Per E Lønning, Patrick O Brown, Anne-Lise Børresen-Dale, and David Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100(14):8418–8423, Jul 2003.

[133] R Spang, C Blanchette, H Zuzan, JR Marks, J Nevins, and M West. Prediction and ucertainty in the analysis of gene expression profiles. *In Silico Biology*, 2(3):369–381, 2002.

[134] T Speed, editor. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall/CRC, Florida, USA, 2003.

[135] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

[136] Aik Choon Tan and David Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics*, 2(3 Suppl):S75–83, 2003.

[137] Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng, and George C Tseng. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, Jul 2006.

[138] R Tibshirani, T Hastie, B Narasimhan, and G Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 99(10):6567–72, May 2002.

[139] R Tibshirani, T Hastie, B Narasimhan, and G Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statist. Sci.*, 18(1):104–117, 2003.

[140] RJ Tibshirani and B Efron. Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology*, 1(1):1–18, 2002.

[141] Robert Tibshirani. Immune signatures in follicular lymphoma. *N Engl J Med*, 352(14):1496–7; author reply 1496–7, Apr 2005.

[142] AN Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet. Math. Dokl.*, 4:1035–1038, 1963.

[143] Jung-Fa Tsai. Global optimization of nonlinear fractional programming problems in engineering design. *Engineering Optimization*, 37(4):399–409, 2005.

[144] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[145] Heather Turner, Trevor Bailey, and Wojtek Krzanowski. Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics & Data Analysis*, 48(2):235–254, 2005.

[146] Marc J van de Vijver, Yudong D He, Laura J van't Veer, Hongyue Dai, Augustinus A M Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Leonie Delahaye, Tony van der Velde, Harry Bartelink, Sjoerd Rodenhuis, Emiel T Rutgers, Stephen H Friend, and René Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, Dec 2002.

[147] L J van 't Veer, H Dai, MJ van de Vijver, YD He, AAM Hart, M Mao, HL Peterse, K van der Kooy, MJ Marton, AT Witteveen, GJ Schreiber, RM Kerkhoven, C Roberts, PS Linsley, R Bernards, and SH Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, Jan 2002.

[148] V Vapnik. *The nature of statistical learning theory*. Springer, New York, 1995.

[149] V Vapnik. *Statistical learning theory*. Wiley, New York, 1998.

[150] Lodewyk F A Wessels, Marcel J T Reinders, Augustinus A M Hart, Cor J Veenman, Hongyue Dai, Yudong D He, and Laura J Van't Veer. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21(19):3755–62, Oct 2005.

[151] M West, C Blanchette, H Dressman, E Huang, S Ishida, R Spang, H Zuzan, JA Olson, JR Marks, and JR Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA*, 98(20):11462–7, Sep 2001.

[152] Michael L Whitfield, Lacy K George, Gavin D Grant, and Charles M Perou. Common markers of proliferation. *Nat Rev Cancer*, 6(2):99–106, Feb 2006.

[153] H Willenbrock, AS Juncker, K Schmiegelow, S Knudsen, and L P Ryder. Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays. *Leukemia*, 18(7):1270–1277, Jul 2004.

[154] Zhijin Wu and Rafael A. Irizarry. Statistical framework for the analysis of microarray probe-level data. Technical report, Johns Hopkins University, Dept. of Biostatistics Working Papers, March 2005.

[155] Zhijin Wu and Rafael A Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol*, 12(6):882–893, 2005.

[156] Jiong Yang, Haixun Wang, Wei Wang, and P. Yu. Enhanced biclustering on expression data. In *Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on*, 2003.

[157] Eng-Juh Yeoh, Mary E Ross, Sheila A Shurtleff, WK Williams, D Patel, R Mahfouz, FG Behm, SC Raimondi, MV Relling, A Patel, C Cheng, D Campana, D Wilkins, X Zhou, J Li, H Liu, CH Pui, WE Evans, C Naeve, L Wong, and JR Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–43, Mar 2002.

[158] H Zhang, CY Yu, B Singer, and M Xiong. Recursive partitioning for tumor classification with gene expression microarray data. *Proc Natl Acad Sci U S A*, 98(12):6730–5, Jun 2001.

[159] Heping Zhang, Chang-Yung Yu, and Burton Singer. Cell and tumor classification using gene expression data: construction of forests. *Proc Natl Acad Sci U S A*, 100(7):4168–72, Apr 2003.

[160] Qingwei Zhang, Rie Ushijima, Takatoshi Kawai, and Hiroshi Tanaka. Which to use? - microarray data analysis in input and output data processing. *Chem-Bio Informatics Journal*, 4:56–72, 2004.

[161] J Zhu and T Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–43, Jul 2004.