Bibliography

[1] J. Aach and G. M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001.

[2] S. Aizawa, H. Nakano, T. Ishida, R. Horie, M. Nagai, K. Ito, H. Yagita, K. Okumura, J. Inoue, and T. Watanabe. Tumor necrosis factor receptor-associated factor (TRAF) 5 and TRAF2 are involved in CD30-mediated NFkappaB activation. *Journal of Biological Chemistry*, 272(4):2042–2045, 1997.

[3] K. Akashi, X. He, J. Chen, H. Iwasaki, C. Niu, B. Steenhard, J. Zhang, J. Haug, and L. Li. Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis. *Blood*, 101(2):383–389, 2003.

[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell, Fourth Edition*. Garland, 2002.

[5] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[6] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):10101–10106, 2000.

[7] T. W. Anderson. Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, 1(1):135–141, 1973.

[8] I. P. Androulakis, E. Yang, and R. R. Almon. Analysis of time-series gene expression data: Methods, challenges, and opportunities. *Annual Review Of Biomedical Engineering*, 9:205–228, 2007.

[9] M. Ashburner. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[10] J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

[11] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–503, 2004.

[12] Z. Bar-Joseph, G. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon. A new approach to analyzing gene expression time series data. In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, pages 39–48, New York, NY, USA, 2002. ACM.

[13] Z. Bar-Joseph, G. Gerber, L. Simon, D. K. Gifford, T. S. Jaakkola, and T. S. Jaakkola. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(18):10146–10151, 2003.

[14] Z. Bar-Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon. Continuous representations of time-series gene expression data. *Journal of Computational Biology*, 10(3-4):341–356, 2003.

[15] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21(11):1337–1342, 2003.

[16] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17 Suppl 1:i22–i29, 2001.

[17] D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.

[18] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. DallaFavera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4):382–390, 2005.

[19] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68, New York, NY, USA, 2004. ACM.

[20] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[21] L. H. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique ocurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[22] BDGP. Berkeley Drosophila Genome Project, http://www.fruitfly.org.

[23] N. Beerenwinkel, J. Rahnenfuhrer, M. Daumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. In *RECOMB '04: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pages 36–44, New York, NY, USA, 2004. ACM Press.

[24] T. Beissbarth and T. P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[25] A. Bhandoola and A. Sambandam. From stem cell to t cell: one route or many? *Nature Reviews Immunology*, 6:117–126, 2006.

[26] C. Biernacki and G. Govaert. Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 29(2):451–457, 1997.

[27] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, International Computer Science Institute, 1997.

[28] N. Bolshakova and F. Azuaje. Cluster validation techniques for genome expression data. *Signal Processing*, 83(4):825–833, 2003.

[29] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[30] L. Brehelin. Clustering gene expression series with prior knowledge. In *Algorithms in Bioinformatics, Proceding of the Workshop in Bioinformatics*, number 3691 in LNBI, pages 27–38. Springer Verlag, 2005.

[31] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12):4164–4169, 2004.

[32] L. Busino, M. Donzelli, M. Chiesa, D. Guardavaccaro, D. Ganoth, N. V. Dorrello, A. Hershko, M. Pagano, and G. F. Draetta. Degradation of Cdc25A by beta-TrCP during S phase and in response to DNA damage. *Nature*, 426(6962):87–91, 2003.

[33] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *PSB 2005: Proceedings of the Pacific Symposium on Biocomputing*, pages 418–429, 2000.

[34] L. Bystrykh, E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, T. Wiltshire, A. I. Su, E. Vellenga, J. Wang, K. F. Manly, L. Lu, E. J. Chesler, R. Alberts, R. C. Jansen, R. W. Williams, M. P. Cooke, and G. de Haan. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genetics*, 37(3):225–232, 2005.

[35] G. A. Calin, C. Liu, C. Sevignani, M. Ferracin, N. Felli, C. D. Dumitru, M. Shimizu, A. Cimmino, S. Zupo, M. Dono, M. L. Dell'Aquila, H. Alder, L. Rassenti, T. J. Kipps, F. Bullrich, M. Negrini, and C. M. Croce. MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32):11755–11760, 2004.

[36] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC, 2000.

[37] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793, 1995.

[38] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

[39] O. Chapelle, B. Schoelkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

[40] S. Chaudhuri, M. Drton, and T. S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216, 2007.

[41] C. Z. Chen, L. Li, H. F. Lodish, and D. P. Bartel. MicroRNAs modulate hematopoietic lineage differentiation. *Science*, 303(5654):83–86, 2004.

[42] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65–73, 1998.

[43] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[44] A. Cimmino, G. A. Calin, M. Fabbri, M. V. Iorio, M. Ferracin, M. Shimizu, S. E. Wojcik, R. I. Aqeilan, S. Zupo, M. Dono, L. Rassenti, H. Alder, S. Volinia, C. Liu, T. J. Kipps, M. Negrini, and C. M. Croce. miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13944–13949, 2005.

[45] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553–1567, 2004.

[46] T. H. Cormen, C. E. Leiserson, and C. Rivest, Ronald L. amd Stein. *Introduction to Algorithms*. MIT Press and McGraw-Hill, second edition, 2001.

[47] I. G. Costa, M. C. P. de Souto, and A. Schliep. Validating gene clusterings by selecting informative gene ontology terms with mutual information. In *Advances in Bioinformatics and Computational Biology, Proceedings of the Brazilian Symposium on Bioinformatics*, LNBI, pages 81–92. Springer Verlag, 2007.

[48] I. G. Costa, R. Krause, L. Optiz, and A. Schliep. Semi-supervised learning for the identification of syn-expressed genes from fused microarray and in situ image data. *BMC Bioinformatics*, 8(Suppl 10):S3, 2007.

[49] I. G. Costa, S. Roepcke, C. Hafemeister, and A. Schliep. Inferring differentiation pathways from gene expression. *Bioinformatics*, 2008. Accepted.

[50] I. G. Costa, S. Roepcke, and A. Schliep. Gene expression trees in lymphoid development. *BMC Immunology*, 8(1):25, 2007.

[51] I. G. Costa and A. Schliep. On external indices for mixtures: validating mixtures of genes. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nurnberger, and W. Gaul, editors, *From Data and Information Analysis to Knowledge Engineering, Proceedings of the 29th Annual Conference of the Gesellschaft fur Klassifikation*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 662–669. Springer, 2005.

[52] I. G. Costa and A. Schliep. On the feasibility of heterogeneous analysis of large scale biological data. In *Proceedings of ECML/PKDD 2006 Workshop on Data and Text Mining for Integrative Biology*, pages 55–60, 2006.

[53] I. G. Costa, A. Schönhuth, and A. Schliep. The Graphical Query Language: a tool for analysis of gene expression time-courses. *Bioinformatics*, 21(10):2544–2545, 2005.

[54] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., New York, N.Y., 1991.

[55] C. M. Croce and G. A. Calin. miRNAs, cancer, and stem cell division. *Cell*, 122(1):6–7, 2005.

[56] S. Datta and S. Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19:459–466, 2003.

[57] I. Davidson and S. S. Ravi. Intractability and clustering with constraints. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 201–208, New York, NY, USA, 2007. ACM.

[58] U. de Lichtenberg, L. J. Jensen, A. Fausbull, T. S. Jensen, P. Bork, and S. Brunak. Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, 21(7):1164–1171, 2005.

[59] M. C. P. de Souto, D. A. S. Araujo, I. G. Costa, R. G. F. Soares, T. B. Ludermir, and A. Schliep. Comparative study on normalization procedures for cluster analysis of gene expression datasets. In *Proceedings of the International Joint Conference on Neural Networks*. IEEE Computer Society, 2008. Accepted.

[60] M. C. P. de Souto, R. B. C. Prudencio, R. G. F. Soares, D. A. S. Araujo, I. G. Costa, T. B. Ludermir, and A. Schliep. Ranking and selecting clustering algorithms using a meta-learning approach. In *Proceedings of the International Joint Conference on Neural Networks*. IEEE Computer Society, 2008. Accepted.

[61] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[62] R. Desper, F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schaffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology*, 6(1):37–51, 1999.

[63] R. Desper, F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schaffer. Distance-based reconstruction of tree models for oncogenesis. *Journal of Computational Biology*, 7(6):789–803, 2000.

[64] P. D'haeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23(12):1499–1501, 2005.

[65] J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375, 1994.

[66] S. Dudoit and J. Fridlyand. A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biology*, 3(7):R36, 2002.

[67] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids*. Cambridge Press, 1998.

[68] B. Edgar. Diversification of cell cycle controls in developing embryos. *Current Opinion in Cell Biology*, 7(6):815–824, 1995.

[69] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[70] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall, 1997.

[71] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95:14863–8, 1998.

[72] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. MicroRNA targets in Drosophila. *Genome biology*, 5(1):R1, 2003.

[73] J. Ernst, G. J. Nau, and Z. Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21 Suppl 1:i159–i168, 2005.

[74] J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph. Reconstructing dynamic regulatory maps. *Molecular systems biology*, 3:74, 2007.

[75] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[76] N. Felli, L. Fontana, E. Pelosi, R. Botta, D. Bonci, F. Facchiano, F. Liuzzi, V. Lulli, O. Morsilli, S. Santoro, M. Valtieri, G. A. Calin, C. G. Liu, A. Sorrentino, C. M. Croce, and C. Peschle. MicroRNAs 221 and 222 inhibit normal erythropoiesis and erythroleukemic cell growth via kit receptor down-modulation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(50):18081–18086, 2005.

[77] M. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[78] V. Filkov, S. Skiena, and J. Zhi. Analysis techniques for microarray time-series data. *Journal of Computational Biology*, 9(2):317–30, 2002.

[79] C. Fraley and A. E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[80] C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, 2007.

[81] M. Futschik and H. Herzel. Are we overestimating the number of cell-cycling genes? the impact of background models. In *GCB 2007: Proceedings of the German Conference on Bioinformatics*, volume 115 of *Lecture Notes in Informatics*, pages 2–14, 2007.

[82] A. Gasch, P. Spellman, C. Kao, O. Carmel-Harel, M. Eisen, G. Storz, D. Botstein, and P. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–57, 2000.

[83] A. C. Gavin, M. Busche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. Michon, C. M. Cruciat, M. Remor, C. Hufert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. S. Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[84] A. Gelman and D. Rubin. Markov chain monte carlo methods in biostatistics. *Statistical Methods in Medical Research*, 5:339–355, 1996.

[85] J. Gilthorpe, M. Vandromme, T. Brend, A. Gutman, D. Summerbell, N. Totty, and P. W. J. Rigby. Spatially specific expression of Hoxb4 is dependent on the ubiquitous transcription factor NFY. *Development*, 129(16):3887–3899, 2002.

[86] R. Glynne, G. Ghandour, J. Rayner, D. H. Mack, and C. C. Goodnow. B-lymphocyte quiescence, tolerance and activation as viewed by global gene expression profiling on microarrays. *Immunological reviews*, 176:216–246, 2000.

[87] R. Gonzalez and P. Wintz. *Digital image processing*. Addison-Wesley, 1991.

[88] A. D. Gordon. *Classification: methods for the exploratory analysis of multivariate data*. Chapman & Hall, 1981.

[89] A. D. Gordon. *Classification*. Chapman & Hall, New York, 1999.

[90] GQL. Graphical query language, http://www.ghmm.org/gql.

[91] S. GriffithsJones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database issue):D140–D144, 2006.

[92] J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. Unpublished, 1971.

[93] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[94] S. D. Hatfield, H. R. Shcherbata, K. A. Fischer, K. Nakahara, R. W. Carthew, and H. RuoholaBaker. Stem cell division is regulated by the microRNA pathway. *Nature*, 435(7044):974–978, 2005.

[95] R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800, 1985.

[96] C. L. L. Hendriks, S. V. E. Keränen, C. C. Fowlkes, L. Simirenko, G. H. Weber, A. H. DePace, C. Henriquez, D. W. Kaszuba, B. Hamann, M. B. Eisen, J. Malik, D. Sudar, M. D. Biggin, and D. W. Knowles. Three-dimensional morphology and gene expression in the drosophila blastoderm at cellular resolution i: data acquisition pipeline. *Genome biology*, 7(12):R123, 2006.

[97] A. Hoffmann, A. Levchenko, M. L. Scott, and D. Baltimore. The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science*, 298(5596):1241–1245, 2002.

[98] R. Hoffmann, L. Bruno, T. Seidl, A. Rolink, and F. Melchers. Rules for gene usage inferred from a comparison of large-scale gene expression profiles of T and B lymphocyte development. *Journal of Immunology*, 170(3):1339–1353, 2003.

[99] R. Hoffmann and F. Melchers. A genomic view of lymphocyte development. *Current Opinion in Immunology*, 15(3):239–245, 2003.

[100] R. Hoffmann, T. Seidl, M. Neeb, A. Rolink, and F. Melchers. Changes in gene expression profiles in developing B cells of murine bone marrow. *Genome Research*, 12(1):98–111, 2002.

[101] I. Holmes and W. J. Bruno. Finding regulatory elements using joint likelihoods for sequence and expression profile data. In *ISMB 2000: Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, volume 8, pages 202–10, 2000.

[102] D. Huang and W. Pan. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, 22(10):1259–1268, 2006.

[103] L. J. Hubbert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:63–76, 1985.

[104] W. Huber, A. V. Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:i96–i104, 2002.

[105] G. Hyatt, R. Melamed, R. Park, R. Seguritan, C. Laplace, L. Poirot, S. Zucchelli, R. Obst, M. Matos, E. Venanzi, A. Goldrath, L. Nguyen, J. Luckey, T. Yamagata, A. Herman, J. Jacobs, D. Mathis, and C. Benoist. Gene expression microarrays: glimpses of the immunological genome. *Nature Immunology*, 7(7):686–691, 2006.

[106] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[107] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush,

A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu. Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2(5):345–350, 2005.

[108] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–4574, 2001.

[109] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988.

[110] A. K. Jain and J. Moreau. Bootstrap techniques in cluster analysis. *Pattern Recognition*, 20:547–568, 1987.

[111] K. K. Jain. Biochips for gene spotting. *Science*, 294(5542):621–623, 2001.

[112] T. Joachims. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer/Springer, 2002.

[113] N. Kaminski and Z. Bar-Joseph. A patient-gene model for temporal expression profiles in clinical studies. *Journal of Computational Biology*, 14(3):324–338, 2007.

[114] M. Kanehisa, S. Goto, M. Hattori, K. F. AokiKinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue):D354–D357, 2006.

[115] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

[116] A. Kazanjian, E. A. Gross, and H. L. Grimes. The growth factor independence-1 transcription factor: New functions and new insights. *Critical reviews in oncology/hematology*, 2006.

[117] S. V. E. Keränen, C. C. Fowlkes, C. L. L. Hendriks, D. Sudar, D. W. Knowles, J. Malik, and M. D. Biggin. Three-dimensional morphology and gene expression in the drosophila blastoderm at cellular resolution ii: dynamics. *Genome biology*, 7(12):R124, 2006.

[118] D. Klein, S. Kamvar, and C. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *ICML '02: Proceedings of the 19th international conference on Machine learning*, 2002.

[119] I. S. Kohane, A. J. Butte, and A. Kho. *Microarrays for an Integrative Genomics*. MIT Press, Cambridge, MA, USA, 2002.

[120] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[121] T. Koski. *Hidden Markov Models for Bioinformatics*. Kluwer Academic Publishers, 2001.

[122] R. Krause, C. von Mering, P. Bork, and T. Dandekar. Shared components of protein complexes–versatile building blocks or biochemical artefacts? *BioEssays : news and reviews in molecular, cellular and developmental biology*, 26(12):1333–1343, 2004.

[123] T. Lange, M. H. Law, A. K. Jain, and J. M. Buhmann. Learning with constrained and unlabelled data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 731–738, 2005.

[124] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural Computing*, 16(6):1299–1323, 2004.

[125] S. L. Lauritzen. *Graphical Models*. Oxford University Press, USA, 1996.

[126] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B*, 50:157–224, 1988.

[127] E. Lecuyer, H. Yoshida, N. Parthasarathy, C. Alm, T. Babak, T. Cerovina, T. R. Hughes, P. Tomancak, and H. M. Krause. Global analysis of mrna localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 131(1):174–187, 2007.

[128] T. Lee. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[129] M. Legendre, W. Ritchie, F. Lopez, and D. Gautheret. Differential Repression of Alternative Transcripts: A Screen for miRNA Targets. *PLoS Computational Biology*, 2(5):e43, 2006.

[130] M. Leptin. Gastrulation in drosophila: the logic and the cellular mechanisms. *The EMBO Journal*, 18:3187–3192, 1999.

[131] S. Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[132] X. Li, A. Schliep, and A. Schoenhuth. The Viterbi decomposition. Technical report, Center for Applied Computer Science, University of Cologne, 2004.

[133] L. P. Lim, N. C. Lau, P. GarrettEngele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.

[134] R. J. Lipshutz, D. Morris, M. Chee, E. Hubbell, M. J. Kozal, N. Shah, N. Shen, R. Yang, and S. P. Fodor. Using oligonucleotide probe arrays to access genetic diversity. *BioTechniques*, 19(3):442–447, 1995.

[135] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405(6788):827–836, 2000.

[136] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. E. Darnell. *Molecular Cell Biology*. W. H. Freeman & Co., 4th edition edition, 2000.

[137] Z. Lu and T. Leen. Semi-supervised learning with penalized probabilistic clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 849–856. MIT Press, 2005.

[138] Y. Luan and H. Li. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474–482, 2003.

[139] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[140] P. Matthias and A. G. Rolink. Transcriptional networks in developing and mature B cells. *Nature Reviews Immunology*, 5(6):497–508, 2005.

[141] R. M. McIntyre and R. K. Blashfield. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavior Research*, 15:225–238, 1980.

[142] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York, Basel, 1988.

[143] G. J. McLachlan, R. W. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, 2002.

[144] G. J. McLachlan and D. Peel. Robust cluster analysis via mixtures of multivariate t-distributions. In *SSPR '98/SPR '98: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition*, pages 658–666, London, UK, 1998. Springer-Verlag.

[145] G. J. McLachlan and D. Peel. *Finite Mixture Models.* Wiley Series in Probability and Statistics. Wiley, New York, 2000.

[146] J. McQueen. Some methods of classification and analysis of multivariate observations. In *5th Berkeley Symposium in Mathematics, Statistics and Probability*, pages 281–297, 1967.

[147] M. Medvedovic, K. Yeung, and R. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

[148] M. Meila and M. I. Jordan. Learning with mixtures of trees. *Journal of Machine Learning Research*, 1:1–48, 2001.

[149] G. W. Milligan and M. C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavior Research*, 21:441–458, 1986.

[150] C. Moller-Levet, F. Klawonn, K. Cho, and O. Wolkenhauer. Fuzzy clustering of short time-series and unevenly distributed sampling points. In *Advances in Intelligent Data Analysis V*, volume 2810 of *LNCS*, pages 330–340. Springer Verlag, 2003.

[151] S. Monticelli, K. M. Ansel, C. Xiao, N. D. Socci, A. M. Krichevsky, T. Thai, N. Rajewsky, D. S. Marks, C. Sander, K. Rajewsky, A. Rao, and K. S. Kosik. MicroRNA profiling of the murine hematopoietic system. *Genome biology*, 6(8):R71, 2005.

[152] H. Nakano, S. Sakon, H. Koseki, T. Takemori, K. Tada, M. Matsumoto, E. Munechika, T. Sakai, T. Shirasawa, H. Akiba, T. Kobata, S. M. Santee, C. F. Ware, P. D. Rennert, M. Taniguchi, H. Yagita, and K. Okumura. Targeted disruption of Traf5 gene causes defects in CD40- and CD27-mediated lymphocyte activation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(17):9803–9808, 1999.

[153] B. Nelson and I. Cohen. Revisiting probabilistic models for clustering with pair-wise constraints. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 673–680, New York, NY, USA, 2007. ACM.

[154] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 13*, pages 849–856. MIT Press, 2001.

[155] S. K. Ng, G. J. McLachlan, K. Wang, L. BenTovim Jones, and S. Ng. A Mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, 22(14):1745–1752, 2006.

[156] N. Niederberger, L. K. Buehler, J. Ampudia, and N. R. J. Gascoigne. Thymocyte stimulation by anti-TCR-beta, but not by anti-TCR-alpha, leads to induction of developmental transcription program. *Journal of leukocyte biology*, 77(5):830–841, 2005.

[157] C. Niehrs and N. Pollet. Synexpression groups in eukaryotes. *Nature*, 402(6761):483–487, 1999.

[158] OMIM. Online mendelian inheritance in man, http://www.ncbi.nlm.nih.gov/omim/.

[159] L. Opitz, A. Schliep, and S. Posch. Analysis of fused in-situ hybridization and gene expression data. In R. Decker and H. J. Lenz, editors, *Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft fur Klassifikation*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 157–166, Heidelberg, Germany, 2006. Springer.

[160] J. Y. Pan, A. Guilherme, R. Balan, E. P. Xing, A. J. M. Traina, and C. Faloutsos. Automatic mining of fruit fly embryo images. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 693–698, New York, NY, USA, 2006. ACM Press.

[161] W. Pan. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, 22(7):795–801, 2006.

[162] H. Peng, F. Long, M. B. Eisen, and E. W. Myers. Clustering gene expression patterns of fly embryos. In *Proceeding of the 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro*, pages 1144–1147. IEEE, 2006.

[163] H. Peng and E. W. Myers. Comparing in situ mrna expression patterns of drosophila embryos. In *RECOMB '04: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pages 157–166, New York, NY, USA, 2004. ACM.

[164] O. D. Perez, S. Kinoshita, Y. Hitoshi, D. G. Payan, T. Kitamura, G. P. Nolan, and J. B. Lorens. Activation of the PKB/AKT pathway by ICAM-2. *Immunity*, 16(1):51–65, 2002.

[165] L. Poirot, C. Benoist, and D. Mathis. Natural killer cells distinguish innocuous and destructive forms of pancreatic islet autoimmunity. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21):8102–8107, 2004.

[166] S. H. Powis, I. Mockridge, A. Kelly, L. A. Kerr, R. Glynne, U. Gileadi, S. Beck, and J. Trowsdale. Polymorphism in a second ABC transporter gene located within the class II region of the human major histocompatibility complex. *Proceedings of the National Academy of Sciences of the United States of America*, 89(4):1463–1467, 1992.

[167] PubMed. http://www.ncbi.nlm.nih.gov/sites/entrez/.

[168] W. Qiu and H. Joe. Generation of random clusters with specified degree of separation. *Journal of Classification*, 23(2):315–334, 2006.

[169] L. R. Rabiner. A tutorial on Hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.

[170] H. S. Radomska, A. B. Satterthwaite, N. Taranenko, S. Narravula, D. S. Krause, and D. G. Tenen. A nuclear factor Y (NFY) site positively regulates the human CD34 stem cell gene. *Blood*, 94(11):3772–3780, 1999.

[171] S. H. Ramkissoon, L. A. Mainwaring, Y. Ogasawara, K. Keyvanfar, J. P. McCoy, E. M. Sloand, S. Kajigaya, and N. S. Young. Hematopoietic-specific microRNA expression in human cells. *Leukemia research*, 2005.

[172] M. F. Ramoni, P. Sebastiani, and I. S. Kohane. Cluster analysis of gene expression dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14):9121–9126, 2002.

[173] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. L. Wild, and F. Falciani. Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372, 2004.

[174] H. Redestig, D. Weicht, J. Selbig, and M. Hannah. Transcription factor target prediction using multiple short expression time series from arabidopsis thaliana. *BMC bioinformatics*, 8(1):454, 2007.

[175] A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

[176] K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.

[177] E. V. Rothenberg and T. Taghon. Molecular genetics of T cell development. *Annual Review of Immunology*, 23:601–649, 2005.

[178] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[179] S. Roweis and Z. Ghahramani. A Unifying Review of Linear Gaussian Models. *Neural Comp.*, 11(2):305–345, 1999.

[180] S. Sahu and G. Roberts. On convergence of the em algorithm and the gibbs sampler. *Statistics in Computing*, 9:55–64, 1998.

[181] A. A. Samsonova, M. Niranjan, S. Russell, and A. Brazma. Prediction of gene expression in embryonic structures of drosophila melanogaster. *PLoS Computational Biology*, 3(7):e144, 2007.

[182] J. Schafer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4:Article32, 2005.

[183] M. Schena. Genome analysis with gene expression microarrays. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 18(5):427–431, 1996.

[184] A. Schliep. *Learning Hidden Markov Model topology*. PhD thesis, Center for Applied Computer Science, University of Cologne, 2001.

[185] A. Schliep, I. G. Costa, C. Steinhoff, and A. Schönhuth. Analyzing gene expression time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3):179–193, 2005.

[186] A. Schliep, A. Schönhuth, and C. Steinhoff. Using Hidden Markov models to analyze gene expression time course data. *Bioinformatics*, 19 Suppl 1:i255–i263, 2003.

[187] A. Schliep, C. Steinhoff, and A. Schönhuth. Robust inference of groups in gene expression time-courses using mixtures of HMMs. *Bioinformatics*, 20 Suppl 1:i283–i289, 2004.

[188] T. Schmidt, H. Karsunky, B. Rodel, B. Zevnik, H. P. Elsasser, and T. Moroy. Evidence implicating Gfi-1 and Pim-1 in pre-T-cell differentiation steps associated with beta-selection. *EMBO Journal*, 17(18):5349–5359, 1998.

[189] A. Schönhuth, I. G. Costa, and A. Schliep. Semi-supervised clustering of yeast gene expression. In *Japanese-German Workshop on Data Analysis and Classification*. Springer, 2006.

[190] A. Schulze and J. Downward. Navigating gene expression using microarrays–a technology review. *Nature Cell Biology*, 3(8):E190–E195, 2001.

[191] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[192] E. Segal, M. Shapira, A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–176, 2003.

[193] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19 Suppl 1:i264–i271, 2003.

[194] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19 Suppl 1:i273–i282, 2003.

[195] SGD. Saccharomyces genome database, http://www.yeastgenome.org/.

[196] N. Shental, A. BarHillel, T. Hertz, and D. Weinshall. Computing gaussian mixture models with em using equivalence constraints. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[197] Y. Shi, T. Mitchell, and Z. Bar-Joseph. Inferring pairwise regulatory relationships from multiple time series datasets. *Bioinformatics*, 23(6):755–763, 2007.

[198] M. Shiga, I. Takigawa, and H. Mamitsuka. Annotating gene function by combining expression data with a modular gene network. *Bioinformatics*, 23(13):468–478, 2007.

[199] R. Sokal and F. Rohlf. *Biometry*. W. H. Freeman and Company, New York, 1995.

[200] P. Sood, A. Krek, M. Zavolan, G. Macino, and N. Rajewsky. Cell-type-specific signatures of microRNAs on target mRNA expression. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8):2746–2751, 2006.

[201] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[202] M. A. Steel and D. Penny. Distributions of tree comparison metrics-some new results. *Systematic Biology*, 42(2):126–141, 1993.

[203] C. Steinhoff and M. Vingron. Normalization and quantification of differential expression in gene expression microarrays. *Briefings in bioinformatics*, 7(2):166–177, 2006.

[204] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18 Suppl 2:i231–i240, 2002.

[205] A. Stolcke and S. Omohundro. Hidden Markov model induction by Bayesian model merging. In *Advances in Neural Information Processing Systems 5*. 1992.

[206] J. D. Storey, W. Z. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, 2005.

[207] M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[208] Y. C. Tai and T. P. Speed. A multivariate empirical bayes statistic for replicated microarray time course data. *Annals Of Statistics*, 34:2387, 2006.

[209] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2981–2986, 2004.

[210] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18 Suppl 1:i136–i144, 2002.

[211] D. Tautz and C. Pfeifle. A non-radioactive in situ hybridization method for the localization of specific rnas in drosophila embryos reveals translational control of the segmentation gene hunchback. *Chromosoma*, 98(2):81–85, 1989.

[212] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–5, 1999.

[213] B. Thiesson, C. Meek, D. Chickering, and D. Heckerman. Learning mixtures of dag models. In *UIA 98: Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, pages 504–51, San Francisco, CA, 1998. Morgan Kaufmann.

[214] P. Tomancak, A. Beaton, R.Weiszmann, E. Kwan, S. Shu, E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. Celniker, and G. Rubin. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, 3(12), 2002.

[215] P. Tomancak, B. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. Celniker, and G. Rubin. Global analysis of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 8(7):R145, 2007.

[216] S. Troncale, F. Tahi, D. Campard, J. Vannier, and J. Guespin. Modeling and simulation with hybrid functional petri nets of the role of interleukin-6 in human early haematopoiesis. In *PSB 2006: Proceeding of the Pacific Symposium on Biocomputing*, volume 11, pages 427–438, 2006.

[217] O. G. Troyanskaya. Putting microarrays in a context: integrated analysis of diverse biological data. *Briefings in bioinformatics*, 6(1):34–43, 2005.

[218] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein. A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8348–8353, 2003.

[219] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, 2001.

[220] L. E. Tze, B. R. Schram, K. Lam, K. A. Hogquist, K. L. Hippen, J. Liu, S. A. Shinton, K. L. Otipoby, P. R. Rodine, A. L. Vegoe, M. Kraus, R. R. Hardy, M. S. Schlissel, K. Rajewsky, and T. W. Behrens. Basal immunoglobulin signaling actively maintains developmental stage in immature B cells. *PLoS Biology*, 3(3):e82, 2005.

[221] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. Som toolbox for matlab. Technical report, Department of Computer Science and Engineering at the Helsinki University of Technology., 2000.

[222] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

[223] A. von Heydebreck, B. Gunawan, and L. Fuzesi. Maximum likelihood estimation of oncogenetic tree models. *Biostatistics*, 5(4):545–556, 2004.

[224] L. A. Warren and E. V. Rothenberg. Regulatory coding of lymphoid lineage choice by hematopoietic transcription factors. *Current Opinion in Immunology*, 15(2):166–175, 2003.

[225] B. Weinstock-Guttman, D. Badgett, K. Patrick, L. Hartrich, R. Santos, D. Hall, M. Baier, J. Feichter, and M. Ramanathan. Genomic effects of IFN-beta in multiple sclerosis patients. *Journal of Immunology*, 171(5):2694–702, 2003.

[226] M. L. Whitfield, G. Sherlock, A. J. Saldanha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, and D. Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular Biology of the Cell*, 13(6):1977–2000, 2002.

[227] M. P. Windham and A. Cutler. Information ratios for validating mixture analyses. *Journal of the American Statistical Association*, 87(420):1188–1192, 1992.

[228] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In S. T. S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2003.

[229] T. Yamagata, C. Benoist, and D. Mathis. A shared gene-expression signature in innate-like lymphocytes. *Immunological reviews*, 210:52–66, 2006.

[230] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.

[231] C.-H. Yeang and T. Jaakkola. Time series analysis of gene expression and location data. *International Journal on Artificial Intelligence Tools*, 14(5):755–770, 2005.

[232] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

[233] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

[234] K. Y. Yeung, M. Medvedovic, and R. E. Bumgarner. Clustering gene-expression data with repeated measurements. *Genome biology*, 4(5):R34, 2003.

[235] M. Yuan and C. Kendziorski. Hidden markov models for microarray time course data in multiple biological conditions. *Journal Of The American Statistical Association*, 101(476):1323–1332, 2006.

[236] G. Zhu, P. Spellman, T. Volpe, P. Brown, D. Botstein, T. Davis, and B. Futcher. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, 406(6791):90–4, 2000.

[237] X. S. Zhu, M. W. Linhoff, G. Li, K. C. Chin, S. N. Maity, and J. P. Ting. Transcriptional scaffold: CIITA interacts with NF-Y, RFX, and CREB to cause stereospecific regulation of the class II major histocompatibility complex promoter. *Molecular and Cellular Biology*, 20(16):6051–6061, 2000.

Appendix A

Gene Ontology enrichment

In order to find GO terms with annotations related to a given group (or cluster) of genes, one should look for annotation terms that are over-represented in this group. The probability that this over-representation is not found by chance can be measured with the use of a hyper-geometric Fisher exact test [199]. This test returns for each cluster and gene ontology term a p-value describing how statistically significant a GO term is for describing genes in a particular cluster.

Let n be the total number of annotated genes in GO (reference group), and m be the number of genes annotated with a specific GO term. This will give us m positive genes and $n - m$ negative genes. If we draw k genes from the reference group (or analogously obtain a cluster with k genes), we obtain q positive genes and $k - q$ negative genes, see Table A.1 for a 2X2 contingency table representation of these terms. We are interested in observing how unusually large this value q is, given n, m and k. This can be achieved by calculating a p-value defined by $p(X \geq q)$, where X is defined by $\{\mathbf{P}(x = i)\}_{1 \leq i \leq k}$, and $\mathbf{P}(x = i)$ is defined as below:

$$\mathbf{P}(x = i) = \frac{\binom{m}{i}\binom{n-m}{k-i}}{\binom{n}{k}}$$

In the thesis, when a particular GO term is over-represented for a given cluster, we state GO Term X is enriched in cluster Y, or we found enrichment for GO Term X in cluster Y.

A later correction of the p-values is necessary, because of the effects of multiple testing. For example, if we have 1000 GO terms, and a p-value of 0.1 is used, at least 100 false

Table A.1: *2x2 Contingency Table for genes annotated or not annotated by a given GO term*

	Annotated Genes	Non-annotated Genes	Total
in cluster	q	$k - q$	k
not in cluster	$m - q$	$(n - k) - (m - q)$	$n - k$
Total	m	$n - m$	n

positives are expected. To correct this, we apply a false positive discovery ratio proposed in [175].

Appendix B

Analysis of Gene Expression of Lymphoid Development

Table B.1: *Contingency Table comparing results from* `MixDTrees-Dev` *(columns) versus SOM (lines) for* `TCell`

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	41	24	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4
3	6	38	14	1	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
6	2	1	1	14	2	11	2	2	0	6	0	0	0	0	0	0	0	0	0	1
2	4	12	31	32	25	13	0	0	0	0	0	0	0	0	0	0	0	0	0	1
8	0	1	10	0	13	1	0	0	1	0	0	0	1	0	0	0	0	3	0	2
5	0	0	0	35	8	88	3	34	0	4	0	0	0	0	0	0	0	0	0	0
10	0	0	1	0	0	1	15	6	9	1	0	0	0	1	0	0	1	3	0	0
14	0	0	0	0	0	0	10	7	2	23	9	0	0	0	0	0	3	0	0	0
15	0	0	0	0	0	0	0	0	19	0	0	16	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	3	35	0	49	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	17	0	12	21	1	2	0	0	1	0	0	0
18	0	0	0	0	0	0	0	0	4	0	0	47	18	2	0	0	1	0	0	0
12	2	0	0	0	0	0	0	0	1	0	0	4	11	5	7	4	8	5	2	1
17	0	0	0	0	0	0	0	0	5	0	4	0	7	4	7	1	8	0	2	0
19	0	0	0	0	0	0	0	0	0	0	0	6	15	35	40	4	27	0	0	0
13	2	0	0	0	0	0	0	0	0	0	0	0	0	4	7	34	21	0	6	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	24	23	0	0	0
4	4	0	5	2	0	0	5	0	3	0	0	1	0	0	0	0	0	11	0	0
11	1	0	2	0	3	0	0	1	0	0	1	0	0	0	0	6	2	0	3	10
7	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	12	1	0	11	13

Table B.2: *Contingency Table comparing results from* `MixDTrees-Dev` *(columns) versus SOM (lines) for* `BCell`

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	52	0	0	0	0	4	5	4	0	0	0	3	0	0	0	0	0	0	0	0
4	0	20	12	10	0	4	0	0	1	14	0	0	0	0	0	0	0	0	0	0
7	0	6	64	5	25	2	0	0	22	0	0	0	0	0	0	0	0	0	0	0
2	14	3	8	4	2	40	0	0	0	0	0	1	0	0	0	0	0	0	0	0
15	0	0	1	0	0	1	1	0	0	2	0	0	0	0	0	0	0	0	0	0
6	0	7	43	5	10	42	2	0	3	0	0	0	0	0	0	0	0	0	0	0
5	4	1	0	0	0	1	4	1	0	2	1	0	0	0	0	0	0	0	0	0
3	5	3	0	1	0	5	0	7	0	5	0	0	0	0	0	0	0	0	0	0
8	0	7	3	0	0	0	0	0	9	13	0	0	1	0	0	0	0	0	0	0
9	0	1	1	10	0	0	0	7	1	17	0	0	3	0	0	0	0	0	0	0
16	0	0	0	0	0	0	4	1	0	0	4	2	0	2	1	1	18	0	0	0
20	1	0	0	0	0	1	1	3	0	0	7	14	0	0	0	0	0	0	0	0
10	0	0	0	3	0	0	0	0	8	6	0	0	18	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	7	0	0	0	0	2	6	18	2	6	1	0	1
13	0	0	0	0	0	0	0	1	0	0	1	0	0	0	20	13	1	0	0	8
14	0	0	0	0	0	0	0	2	0	0	2	0	0	0	3	28	3	8	4	6
17	0	0	0	0	0	0	0	0	0	0	1	0	0	1	3	25	24	0	0	3
19	0	0	0	0	0	0	2	0	0	1	5	12	0	0	0	4	0	35	18	0
18	0	0	0	0	0	0	0	0	0	0	14	9	0	0	0	24	8	3	18	0
11	0	0	0	0	0	0	0	3	0	1	0	0	14	1	19	0	1	0	0	8

Table B.3: *MicroRNA enrichment per cluster for* `TCell` *for* `MixDTrees-Dev`

Cluster ID	MicroRNA	p-value
3	miR-222	0.0006906
5	miR-15a	0.0019456
	miR-26a	0.0369906
	miR-24	0.0369906
	miR-221	0.0051746
	miR-181a	0.0244306
7	miR-342	0.0200686
8	miR-26a	0.0013526
10	miR-150	0.0012176
	miR-142-3p	0.0000056
11	miR-16	0.0049776
	miR-146	0.0011936
	miR-181b	0.0049776

Table B.4: *MicroRNA enrichment per cluster for* `BCell` *for* `MixDTrees-Dev`

Cluster ID	MicroRNA	p-value
3	miR-26a	0.0358116
	miR-181c	0.0025866
	miR-181b	0.0358116
5	miR-15b	0.0029956
	miR-15a	0.0029956
	miR-223	0.0029956
	miR-221	0.0323296
6	miR-191	0.0486736
	miR-155	0.0271276
19	miR-342	0.0402686
	miR-142-3p	0.0088346

Appendix C

Notation

All chapters

$\mathbf{1}(e)$	indicator function, which takes value 1 iff e is true
α_k	mixture coefficient of the kth mixture component
$E[X]$	expectation of a random variable X
$\mathcal{L}$	likelihood function
K	number of clusters or components in a mixture model
μ_x	mean value of random variable X
$p(x \mid \theta)$	a probability density function over variable X and parameterized by θ
r_{ik}	posterior probability that observation x_i is assigned to the kth mixture component, i.e., $p(y_i = k \mid x_i, \Theta)$
Σ_x	covariance matrix of random variable X
Θ	set of parameters of a mixture model
θ_k	set of parameters of the kth mixture component
X	an L dimensional continuous random variable
x	an observation vector $(x_1, ..., x_L)$ from X
$\mathbf{X}$	a data set represented by a $N \times L$ matrix, where entry x_{ij} denotes the values of the jth variable from the ith observation
Y	an one dimensional discrete random variable
y	an observation of Y, where $y \in \{1, ..., K\}$ indicates the mixture component (or cluster) the observation belongs
$\mathbf{Y}$	a set of N observations from Y, where $y_i = k$ denotes that the ith observation belongs to the kth mixture component (or mixture)
$\mathcal{Y}$	space of all possible values of $\mathbf{Y}$

Chapter 4

A	transition matrix of a HMM, where a_{uv} represents the probability of going from state u to state v
d_u	duration parameter representing the expected number of visits to state u

M number of states of the HMM
μ_u mean parameter of the emission function of the uth state
π_u probability of visiting state u at time $t = 1$
Q an L-dimensional discrete variable representing the sequence of visited states
q observation from Q, where $q = (q_1, ..., q_t, ..., q_L)$ and $q_t \in \{1, ..., M\}$ represents the HMM state visited at time t.
σ_u^2 standard error parameter of the emission function from the uth state
θ_L parameters of a linear HMM

Chapter 5

$\mathrm{D}(p||p^*)$ relative entropy between the pdfs p and p^*
$\mathrm{H}(X)$ entropy of variable X
$\mathrm{I}(X_u, X_v)$ mutual information of variables X_u and X_v
$p^T(x|\Theta)$ dependence tree pdf
$p(x_u|x_v, \tau_u)$ conditional Gaussian pdf
pa parent map defining the dependence tree structure
$\sigma_{u|v}^2$ standard error of the conditional Gaussian pdf
τ_u parameters of a conditional Gaussian pdf
$w_{u|v}$ regression parameter of the conditional Gaussian pdf

Chapter 6

λ^+ parameter defining the penalty weights of positive constraint violations
λ^- parameter defining the penalty weights of negative constraint violations
W pair (W^+, W^-) representing the positive and negative constraint matrices
W^+ positive constraints matrix, where w_{ij}^+ is the positive constraint value for observations i and j
W^- negative constraints matrix, where w_{ij}^- is the negative constraint value for observations i and j
Z an L-dimensional continuous random variable
z an observation $(z_{i1}, ...z_{il}, ..., z_{iL})$ of Z representing the pixel intensities of an image

Appendix D

Abbreviations

BCell	B cell development data
Bimm	immature B cells
BMC	Bayesian model collection
Bpre	pre B cells
Bpro	pro B cells
BIC	Bayesian information criteria
CL	co-location index
CLP	common lymphoid progenitor
CMP	common myeloid progenitor
CR	corrected Rand index
DAG	directed acyclic graph
DN	CD4-/CD8- double negative cells
DPL	CD4+/CD8+ double positive large cells
DPS	CD4+/CD8+ double positive small cells
DTree	dependence tree
ECR	extended corrected Rand index
ED	equal density
EM	expectation-maximization algorithm
E-Step	expectation step
FACS	fluorescence activated cell sorting
GQL	Graphical Query Language
GO	Gene Ontology
ImaGO	Image Gene Ontology
HemoMIR	hematopoiesis related microRNAs data
HMM	hidden Markov model
HMRF	hidden Markov random fields
KEGG	Kyoto encyclopedia of genes and genomes
KMC	k-means model collection
lHMM	linear hidden Markov model
MAP	maximum-a-posteriori

MCMC	Monte Carlo Markov Chain
mir	microRNA
MixDTrees	mixture of dependence trees
MixDTrees-Dev	MixDTrees with the developmental tree as structure
MixDTrees-Str	MixDTrees with estimated structure
MLE	maximum likelihood estimation
MM	probe mismatch
MoG	mixture of multivariate Gaussians
MoG Full	MoG with full covariance matrix
MoG Diag	MoG with diagonal covariance matrix
M-Step	maximization step
NK	natural killer cells
NMF	non-negative matrix factorization
PC	Pearson correlation
pdf	probability density function
pHSC	pluri-potent, self-renewing hematopoietic stem cells
PM	probe match
PPP	pluripotent progenitor
RMC	random model collection
SCC	strongly connected components
Sens	sensitivity index
SIM	simulated data
SOM	self-organizing maps
SSL	semi-supervised learning
Spec	specificity index
SP4	single positive CD4
SP8	single positive CD8
TCell	T cell development data
TCD4	cd4 T cells
TCD8	cd8 T cells
TDN	double negative T cells
TF	transcription factor
TFBS	transcription factor binding site
TNK	natural killer T cells
TR	transcription regulation data
YCC	yeast cell cycle
VD	Viterbi decomposition