

6 CONCLUSIONS

6.1 Conclusions

In this thesis a new methodology for the unsupervised clustering of microarray data was developed, implemented and tested. Central to this methodology is the reinterpretation of clusters as data points themselves. Far from being the end-all of data analysis, the partition created by a given clustering configuration is unreliable and instable. However, the combination of many different partitions, taken *en masse*, is a robust source of information, which can and should be mined for knowledge.

The techniques of exploratory data analysis developed herein were quite effective at discovering putative patterns in leukemia data, and in fact they outperformed existing methods of cluster analysis. However, the author would hope that the techniques not be taken as ‘another form of cluster analysis,’ but rather as an indication that the reinterpretation of clusters as data is a valid and worthwhile endeavor. The methodology presented in this thesis is a first step towards identifying the dominant – not the subjectively ‘relevant’ – clusters across clustering configurations. Different clustering configurations discover different structures; much is possible if these structures can be synthesized.

6.2 Directions for Future Research

Although condensation clustering proved to be successful, computationally it is extremely cumbersome to implement. Moreover, it proved to be susceptible (albeit in a limited way) to many of the same problems that hinder normal clustering, namely the arbitrary specification of parameters. Clustering the clusters has much potential, and was the original motivation for this thesis, but more needs to be known about the shape and characteristics of ‘cluster space’ before such techniques can be perfected. Little theory exists on the metric space of sets, and even less is known about the space of algorithmically-generated clusters. It would be quite instructive to explore other properties of this space, as the author has done to a limited extent with λ and $p_{\mathcal{D}}$.

It would also be relatively simple, and potentially quite useful, to extend the methodology of this thesis to ‘partition space,’ essentially raising all concepts by one more level of complexity. Although this area is even less understood than the cluster space dealt with in this thesis, metrics do exist to measure distance between partitions[43]. It would not be difficult, for instance, to calculate $p_{\mathcal{D}}(P_w, \varepsilon)$ for an arbitrary partition P_w . The discovery of a ‘prevalent partition’ could be extremely useful in subclassification experiments, as well as other forms of analysis that require a proper partition of the underlying data.

Of course, the combination of different algorithms is limited by the algorithms themselves, and if the agglomerate of clustering configurations consistently misses the putative structure in the expression data, surely the synthesis of these configurations will miss the structure as well. Fortunately, cluster analysis is but one technique among many used to analyze expression data. The shotgun-then-synthesize

approach applied here to cluster analysis could potentially be quite useful in other arenas where parameter choices lead to similar variability. This is not to say that brute force analysis should replace a rigorous one, but in data mining it is often the case that little theoretical guidance exists in choosing between different techniques of analysis. In these situations, the synthesis of different approaches provides an attractive alternative to the bickering over which is best.

REFERENCES

1. *Statistics Toolbox User's Guide, Version 4*. 2002, Natick, MA, USA.: The MathWorks.
2. Affymetrix, *GeneChip® Eukaryotic Small Sample Target Labeling Technical Note*. 2001, Affymetrix, Inc.: Santa Clara, CA.
3. Aggarwal, C.C. and P.S. Yu. *Finding generalized projected clusters in high dimensional spaces*. 2001.
4. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*. *Nature*, 2000. **403**: p. 503--511.
5. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. *Proc. Natnl. Acad. Sci. USA.*, 1999. **96**(12): p. 6745--6750.
6. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*. *Proc. Natnl. Acad. Sci. USA.*, 1999. **96**(12): p. 6745-6750.
7. Azuaje, F. and N. Bolshakova, *Clustering Genomic Expression Data: Design and Evaluation Principles*, in *A Practical Approach to Microarray Analysis*, D.P. Berrar, W. Dubitzky, and M. Granzow, Editors. 2003, Kluwer Academic Publishers: Boston, MA. p. 230-244.
8. Ben-Dor, A., et al., *Tissue classification with gene expression profiles*. *Journal of Computational Biology*, 2000. **7**(3-4): p. 559-583.
9. Ben-Dor, A., N. Friedman, and Z. Yakhini. *Class Discovery in Gene Expression Data*. in *Fifth Annual International Conference on Computational Biology*. 2001. Montreal, Quebec, Canada.
10. Ben-Dor, A., R. Shamir, and Z. Yakhini, *Clustering Gene Expression Patterns*. *Journal of Computational Biology*, 1999. **6**(3): p. 281-297.
11. Ben-Hur, A., A. Elisseeff, and I. Guyon. *A Stability Based Method for Discovering Structure in Clustered Data*. in *Pacific Symposium on Biocomputing (PSB2002)*. 2002. Kaua'i, HI.
12. Berkhin, P., *Survey Of Clustering Data Mining Techniques*. 2002, Accrue Software: San Jose, CA.
13. Berry, M.J.A. and G. Linoff, *Data Mining Techniques For Marketing, Sales and Customer Support*. 1996, New York, NY: John Willey & Sons, Inc.
14. Beyer, K., et al. *When Is "Nearest Neighbor" Meaningful?* in *7th International Conference on Database Theory (ICDT99)*. 1999. Jerusalem, Israel.
15. Bittner, M., et al., *Molecular classification of cutaneous malignant melanoma by gene expression profiling*. *Nature*, 2000. **406**: p. 536- 540.
16. Blanco, R., et al., *Selection of Highly Accurate Genes for Cancer Classification by Estimation of Distribution Algorithms*.
17. Bolshakova, N. and F. Azuaje, *Cluster validation techniques for genome expression data*. 2002, Department of Computer Science, Trinity College Dublin: Dublin, Ireland.

18. Bradley, P.S. and U.M. Fayyad. *Refining Initial Points for K-Means Clustering*. in *Proc. 15th International Conf. on Machine Learning*. 1998: Morgan Kaufmann.
19. Brazma, A. and J. Vilo, *Gene expression data analysis*. Febs Letters, 2000. **480**(1): p. 17-24.
20. Constantine, L. and C. Harrington, *Use of GeneChip high-density oligonucleotide arrays for gene expression monitoring*. 2003, Affymetrix Inc.: Santa Clara, CA. p. 1-7.
21. Day, W.H.E., *Complexity theory: An introduction for practitioners of classification*, in *Clustering and Classification*, P. Arabie and L. Hubert, Editors. 1992: River Edge, NJ. p. 138–144.
22. DeRisi, J., V. Iyer, and P. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale*. Science, 1997. **278**: p. 680-686.
23. Deutsch, J.M., *Algorithm for Finding Optimal Gene Sets in Microarray Prediction*. 2001.
24. D'Haeseleer, P., S. Liang, and R. Somogyi, *Genetic network inference: From co-expression clustering to reverse engineering*. 2000.
25. D'Haeseleer, P., et al., *Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data*, in *Information Processing in Cells and Tissues*, R. Paton and M. Holcombe, Editors. 1998, Plenum Publishing. p. 203-221.
26. Dhanasekaran, S., et al., *Delineation of prognostic biomarkers in prostate cancer*. Nature, 2001. **23**(412): p. 822–826.
27. Dubes, R.C., *How many clusters are best? an experiment*. Pattern Recognition, 1987. **6**(20): p. 645-663.
28. Dudoit, S., J. Fridlyand, and T.P. Speed, *Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data*. Journal of the American Statistical Association, 2002. **97**(457): p. 77--??
29. Dudoit, S., et al., *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments*. 2000, Stanford University School of Medicine: Stanford, CA.
30. Dunn, J., *Well separated clusters and optimal fuzzy partitions*. Journal of Cybernetics, 1974. **4**: p. 95-104.
31. Eisen, M., et al., *Cluster analysis and display of genome-wide expression patterns*. Proc. Natl. Acad.Sci.USA, 1998. **95**(25): p. 14863-14868.
32. Everitt, B., *Cluster analysis*. 2d ed. 1980, London: Halsted Press. 136 p.
33. Facility, W.M.K., *Genechip Expression Data Analysis*. 2002, Yale University: New Haven, CT.
34. Fryer, R.M., et al., *Global analysis of gene expression: methods, interpretation, and pitfalls*. Exp Nephrol, 2002. **10**: p. 64-74.
35. Furlanello, C., et al. *Gene Selection and Classification with Support Vector Machines applied to Microarray Data*. in *Primo Workshop Nazionale sulla Bioinformatica*. 2002. Siena, Italy.
36. Garber, M., et al., *Diversity of gene expression in adenocarcinoma of the lung*. PNAS, 2001. **98**(24): p. 13784-13789.

37. Getz, G., E. Levine, and E. Domany, *Coupled two--way clustering analysis of gene microarray data*, in *Natl. Acad. Sci USA*. 2000. p. 12079--12084.
38. Gibbons, F.D. and F.P. Roth, *Judging the quality of gene expression-based clustering methods using gene annotation*. *Genome Research*, 2002. **12**(10): p. 1574 - 1581.
39. Goldstein, D., D. Ghosh, and E. Conlon, *Statistical issues in the clustering of gene expression data*. *Statistica Sinica*, 2002. **12**: p. 219-240.
40. Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*. *Science*, 1999. **286**: p. 531-537.
41. Gordon, A.D., *Classification*. 2 ed. 1999, Boca Raton: Chapman and Hall/CRC. 256.
42. Gordon, G., et al., *Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma*. *Cancer Research*, 2002. **62**(17): p. 4963-4967.
43. Halkidi, M., Y. Batistakis, and M. Vazirgiannis, *On Clustering Validation Techniques*. *Journal of Intelligent Information Systems*, 2001. **17**(2-3): p. 107-145.
44. Halkidi, M. and M. Vazirgiannis. *A Data Set Oriented Approach for Clustering Algorithm Selection*. in *European Conference on Principles of Data Mining and Knowledge Discovery*. 2001. Freiburg, Germany: Springer.
45. Halkidi, M. and M. Vazirgiannis. *Managing Uncertainty and Quality in the Classification Process*. in *Methods and Applications of Artificial Intelligence, Second Hellenic Conference on AI*. 2002. Thessaloniki, Greece: Springer.
46. Hartigan, J.A., *Clustering algorithms*. 1975, New York: Wiley. xiii, 351.
47. Hastie, T., et al., *Gene shaving: a new class of clustering methods for expression arrays*. 2000, Stanford University: Stanford, CA.
48. Hoyle, D.C., et al., *Making sense of microarray data distributions*. *Bioinformatics*, 2002. **18**: p. 576-584.
49. Hsiao, L., et al., *A compendium of gene expression in normal human tissues*. *Physiol Genomics*, 2001. **7**: p. 97-104.
50. Hsiao, L., et al., *Correcting for signal saturation errors in the analysis of microarray data*. *Biotechniques*, 2002. **32**: p. 330-336.
51. Hughes, T., et al., *Functional Discovery via a Compendium of Expression Profiles*. *Cell*, 2000. **102**: p. 109-126.
52. Hunter, L., *Molecular Biology for Computer Scientists*, in *Artificial Intelligence & Molecular Biology*, L. Hunter, Editor. 1993, AAAI/MIT Press.
53. Institute, P., *Hybridisation, washing, staining & scanning of Affymetrix Genechips*. 2002, The Patterson Institute for Cancer Research: Manchester, England.
54. Jaeger, J., R. Sengupta, and W.L. Ruzzo, *Improved Gene Selection For Classification Of Microarrays*, in *Pacific Symposium on Biocomputing, Kauai, Hawaii*. 2003.
55. Jagota, A., *Microarray Data Analysis and Visualization*. 2001, Santa Cruz, CA: Bioinformatics By the Bay Press. 101.

56. Jain, A. and R. Dubes, *Algorithms for clustering data*. 1988, Englewood Cliffs, N.J.: Prentice Hall.
57. Jain, A. and M. Murty, *Data Clustering: A Review*. ACM Computing Surveys, 1999. **31**(3): p. 264-323.
58. Jensen, R., *Personal Communication*. 2003.
59. Jensen, R., et al., *Separating Signal From Noise: Error Models for Oligonucleotide Microarrays*. 2002.
60. Kanungo, T., et al. *The Analysis of a Simple k-Means Clustering Algorithm*. in *Symposium on Computational Geometry*. 2000.
61. Kohonen, T., *Self-organizing maps*. Springer series in information sciences. Vol. 30. 1995, Berlin, New York: Springer. xv, 362.
62. Kothapalli, R., et al., *Microarray results: how accurate are they?* BMC Bioinformatics, 2002. **3**(22): p. 1-10.
63. Kurita, T., *An efficient agglomerative clustering algorithm using a heap*. Pattern Recognition, 1991. **24**(3): p. 205–209.
64. Lander, E.S., *The New Genomics: Global Views of Biology*. Science, 1996. **274**: p. 536-539.
65. Leach, S. and L. Hunter, *Comparative Study of Clustering Techniques for Gene Expression Microarray Data*, in *Currents in Computational Molecular Biology*, S. Miyano, R. Shamir, and T. Takagi, Editors. 2000, Universal Academy Press, Inc.: Tokyo. p. 198-199.
66. Li, W. and Y. Yang, *How Many Genes Are Needed for a Discriminant Microarray Data Analysis?*, in *Methods of Microarray Data Analysis*, S. Lin and K. Johnson, Editors. 2001, Kluwer Academic. p. 137-150.
67. Lipshutz, R., et al., *High density synthetic oligonucleotide arrays*. Nature Genetics, 1999. **21**: p. 20-24.
68. Marcotte, E.M., et al., *A combined algorithm for genome-wide prediction of protein function*. Nature, 1999. **402**: p. 83-86.
69. Meila, M. and D. Heckerman. *An Experimental Comparison of Several Clustering and Initialization Methods*. in *Fourteenth Conference on Uncertainty in Artificial Intelligence*. 1998. San Francisco, CA: Morgan Kaufmann, Inc.
70. Meila, M. and D. Heckerman, *An Experimental Comparison of Model-Based Clustering Methods*. Machine Learning, 2001. **42**(1): p. 9-29.
71. Milligan, G.W. and M.C. Cooper, *An examination of procedures for determining the number of clusters in a data set*. Psychometrika, 1985. **50**: p. 159--179.
72. Monro, G., *The concept of multiset*. Z. Math. Logik Grundlag. Math., 1987. **33**(2): p. 171--178.
73. Morgan, B.J.T. and A.P.G. Ray, *Non-uniqueness and Inversions in Cluster Analysis*. Applied Statistics, 1995. **44**(1): p. 117-134.
74. Murtagh, F., *Multidimensional Clustering Algorithms*, in *Compstat lectures, 4*, J.M. Chambers, Editor. 1985, Physica-Verlag: Vienna. p. 131.
75. Ng, R.T., J. Sander, and M.C. Sleumer. *Hierarchical Cluster Analysis of SAGE Data for Cancer Profiling*. in *BIOKDD01: Workshop on Data Mining in Bioinformatics (with SIGKDD01 Conference)*. 2001.

76. Perou, C.M., et al., *Molecular portraits of human breast tumors*. Nature, 2000. **406**: p. 747-752.
77. Pollard, K.S. and M.J.v.d. Laany, *Resampling-based Methods for Identification of Significant Subsets of Genes in Expression Data*. 2002, U.C. Berkeley Division of Biostatistics: Berkeley, CA. p. 1-34.
78. Quackenbush, J., *Computational Analysis of Microarray Data*. Nature Genetics, 2001. **2**: p. 418-427.
79. Rezaee, R., B.P.F. Lelieveldt, and J.H.C. Reiber, *A new cluster validity index for the fuzzy cmean*. Pattern Recognition Letters, 1998. **19**: p. 237-246.
80. Ross, D.T., et al., *Systematic variation in gene expression patterns in human cancer cell lines*. Nature Genetics, 2000. **24**(3): p. 227-235.
81. Rousseeuw, P., *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*. Journal of Computational Applied Mathematics, 1987. **20**: p. 53-65.
82. Slonim, D., et al. *Class Prediction and Discovery Using Gene Expression Data*. in *Fourth Annual International Conference on Computational Molecular Biology*. 2000.
83. Syropoulos, A., *Mathematics of Multisets*, in *Multiset Processing*, C. Calude, et al., Editors. 2001, Springer-Verlag: Berlin. p. 347-358.
84. Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation*. Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(6): p. 2907-2912.
85. Tang, C., et al. *Interrelated Two-way Clustering: An Unsupervised Approach for Gene Expression Data Analysis*. in *Proc. of 2nd IEEE International Symposium on Bioinformatics and Bioengineering*. 2001. Bethesda, MD.
86. Tavazoie, S., et al., *Systematic determination of genetic network architecture*. Nature Genetics, 1999. **22**(3): p. 281-5.
87. Tibshirani, R., et al., *Clustering methods for the analysis of dna microarray data*. 1999: Department of Health Research and Policy, Stanford University.
88. Traynor, J., et al., *Gene expression patterns vary in clonal cell cultures from Rett syndrome females with eight different MECP2 mutations*. BMC Medical Genetics, 2002. **3**(12).
89. Tusher, V., R. Tibshirani, and G. Chu, *Significance Analysis of Microarrays Applied to the Ionizing Radiation Response*. PNAS, 2001. **98**(9): p. 5116-5121.
90. Vilo, J., et al., *Mining for Putative Regulatory Elements in the Yeast Genome using Gene Expression Data*, in *Proceedings International Conference on Intelligent Systems for Molecular Biology*. 2000, AAAI Press. p. 384--394.
91. Ward, J.H., *Hierarchical grouping to optimize an objective function*. J. American Stat. Assoc, 1963. **58**: p. 236-245.
92. Wen, X., et al., *Large-scale temporal gene expression mapping of central nervous system development*. Proc Natl Acad Sci USA, 1998. **95**: p. 334-339.
93. West, M., et al., *Predicting the clinical status of human breast cancer by using gene expression profiles*. PNAS, 2001. **98**: p. 11462--11467.

94. Wigle, D.A., et al., *Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-free Survival*. *CANCER RESEARCH*, 2002. **62**: p. 3005–3008.
95. Wilks, D.S., *Statistical Methods in the Atmospheric Sciences*. 1995, New York, NY: Academic Press.
96. Wishart, D. *k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values*. in *Proceedings of the German Classification Society*. 2001. publication pending review.
97. Wishart, D., *WHISKY CLASSIFIED: Choosing Single Malts by Flavour*. 2002, London: Pavilion Books. 250.
98. Wu, L.F., et al., *Large-scale Prediction of Saccharomyces Cerevisiae Gene Function Using Overlapping Transcriptional Clusters*. *Nature Genetics*, 2002. **31**: p. 255-265.
99. Yeang, C.H., et al., *Molecular Classification of Multiple Tumor Types*. *Bioinformatics*, 2001. **17**(Suppl. 1): p. S316--S322.
100. Yeung, K.Y., D.R. Haynor, and W.L. Ruzzo, *Validating Clustering for Gene Expression Data*. *Bioinformatics*, 2001. **17**(4): p. 309--318.
101. Zhang, T., R. Ramakrishnan, and M. Livny, *BIRCH: A new data clustering algorithm and its applications*. *Data Mining and Knowledge Discovery*, 1997. **1**(2): p. 141--182.