Eric Slud 8/29/01

Reading List for AMSC 699 Genomics Seminar

I. Background Texts.

I.1. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) **Biological Sequence Analysis:** *Probabilistic models of proteins and nucleic acids.* Cambridge: Cambridge Univ. Press.

Contains introductory material on DNA sequence similarity, followed by extensive discussion of dynamic-programming algorithms for optimal matching of pairs of DNA subsequences, with respect to criteria motivated by Markov and hidden-Markov probabilistic models for how sequences are generated.

I.2. Scientific American articles on fundamental molecular biology (some of which are collected in Scientific American paperback book series); *current* elementary (e.g., high-school biology texts, or e.g. the text in BIOL 105,

Campbell, Reece, and Mitchell, **Biology, 5th ed.**

I.3. Textbook material on multivariate classification and clustering or pattern recognition, like:

Duda, R., Hart, P. and Stork, D. (2000) **Pattern Classification**, 2nd ed. New York: Springer-Verlag.

Basic material of this sort, from a statistical point of view, can be found in (application-oriented) books on 'Multivariate Analysis, e.g.,

Johnson, R. and Wichern, D. (1998) **Applied Multivariate Statistical Analysis**, 4th ed. Saddle River: Prentice-Hall.

I.4. Hartigan, J. (1975) **Clustering Algorithms.** New York: Springer-Verlag. *Still a classic text for clustering, updated somewhat in*

Arabie, P., Hubert, L., and DeSoete, G. eds. (1996) **Clustering and Classification**. Singapore: World Scientific.

I.5. **Encyclopedia of Statistical Sciences**, eds. Johnson, Kotz, and Read, articles on: Clustering, Supervised Learning, (Nonparametric) Multivariate Classification, Multiple Comparisons.

II. Background (nonmathematical).

II.1. Hamadeh, H. and Afshari, C. (2000) *Gene chips and functional genomics.* Amer. Scientist **88**, 508–515.

Basic introduction, with very light prerequisites in biology and chemistry.

II.2. The article, *Initial sequencing and analysis of the human genome*, by the International Human Genome Sequencing Consortium (hundreds of authors !), **Nature** v. 409 (Feb. 15, 2001), 860-921, can be down-loaded from **http://www.nature.com** It has a lot of information related to the human genome sequence, which is ultimately highly relevant to the processing of microarry data and to 'functional genomics' generally, but it is difficult and technical. There was also a Supplement to Nature Genetics (vol. 21) in January 1999, devoted to Microarray techonology and analysis. If you know the basics of DNA, RNA's and transcription, and are willing to struggle a little with biochemistry-based nomenclature, then these articles may be the best comprehensive introduction to non-mathematical aspects of Microarrays.

III. Miscellaneous Papers involving Stat Methodology.

III.1. Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci.* **95**, 14863-8.

Very widely cited paper on automatic hierarchical cluster analysis of microarray data. Emphasizes reordering of genes and samples based on clustering information, followed by graphical display of resulting data to aid in uncovering or confirming biological structure.

III.2. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and E. Lander. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531-7.

Another seminal paper showing how automatic clustering is relation to classification of cancers.

III.3. Alizadeh, A, Eisen, M., ..., Botstein, D., Brown, P., and Staudt, L. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-11.

Paper primarily applying hierarchical clustering technique to lymphoma data. Particularly interesting for issues related to simultaneous clustering of genes and classification/clustering of cell lines.

III.4. Kerr, M., and Churchill, G.. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **77**: 123-8.

III.5. Lee, M., Kuo, F., Whitmore, G. and J. Sklar. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Nat. Acad. Sci.* **97** 9834-9.

III.6. Dudoit, S., Fridlyand, J., and Speed, T. (2000) Comparison of discrimination methods for the classification of tumors using gene expression data. Tech. report to appear in JASA.

Survey of statistical techniques for discrimination and classification, along with illustrations on several real datasets, including bootstrap simulation experiments to illustrate correct classification rates on these datasets. Other papers by these authors can be found at the website W.5 below.

III.7. Lazzeroni, L., and Owen, A. (2000) Plaid models for gene expression data. Stanford Univ. preprint.

Interesting attempt to model structure simultaneously with respect to genes and samples, by means of additively superposed layers. Methods have heavy flavor of optimization and algorithmics, not statistical theory. Downloadable from

http://www-stat.stanford.edu/~owen/reports/plaid.pdf

III.8. Ideker, T., Thorsson, V., Siegel, A. and Hood, L. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.* **7**: 805-17.

III.9. Brown, M., Grundy, W., LIN, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M. Jr., and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Nat. Acad. Sci.* **97**: 262-7.

III.10. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q. Kitareewan, S., E. Dmitrovsky, E., Lander, E., and Golub, T. (1999) Interpreting patterns

of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci.* **96**: 2907-12.

III.11. Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chen, W., Botstein, D., and Brown, P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* **1**: 0003.1-0003.21.

IV. More papers from CMU via K. Sellers & E. Russek-Cohen

IV.1. Hastie, T., R. Tibshirani, D. Botstein, and P. Brown. Supervised harvesting of expression trees. Genome Biol. 2: 0003.1-0003.12, 2001.

IV.2. Burke, H. B. Discovering patterns in microarray data. Mol. Diagn. 5: 349-57, 2000.

IV.3. Aach, J., W. Rindone, and Church, G.. Systematic management and analysis of yeast gene expression data. Genome Res. 10: 431-45, 2000.

IV.4. Akutsu, T., S. Miyano, and S. Kuhara. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. J. Comput. Biol. 7: 331-43, 2000.

IV.5. Bard, J. B. A bioinformatics approach to investigating developmental pathways in the kidney and other tissues. Int. J. Dev. Biol. 43: 397-403, 1999.

IV.6. Ermolaeva, O., M. Rastogi, K. D. Pruitt, G. D. Schuler, M. L. Bittner, Y. Chen, R. Simon, P. Meltzer, J. M. Trent, and M. S. Boguski. Data management and analysis for gene expression arrays. Nat. Genet. 20: 19-23, 1998.

IV.7. Friedman, N., M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. J. Comput. Biol. 7: 601-20, 2000.

IV.8. Holter, N. S., M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc. Nat. Acad. Sci. 97: 8409-14., 2000.

IV.9. Holter, N. S., A. Maritan, M. Cieplak, N. V. Fedoroff, and J. R. Banavar. Dynamic modeling of gene expression data. Proc. Nat. Acad. Sci. 98: 1693-8., 2001.

IV.10. Kadota, K., R. Miki, H. Bono, K. Shimizu, Y. Okazaki, and Y. Hayashizaki. Preprocessing implementation for microarray (PRIM): an efficient method for processing cDNA microarray data. Physiol. Genomics 4: 183-8., 2001.

IV.11. Kanehisa, M., and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28: 27-30, 2000.

IV.12. Lemkin, P. F., G. C. Thornwall, K. D. Walton, and L. Hennighausen. The microarray explorer tool for data mining of cDNA microarrays: application for the mammary gland [In Process Citation]. Nucleic Acids Res. 28: 4452-9, 2000.

IV.13. Long, A. D., H. J. Mangalam, B. Y. Chan, L. Tolleri, G. W. Hatfield, and P. Baldi. Gene Expression Profiling in Escherichia coli K12: Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. J. Biol. Chem. 20: 20, 2001.

IV.14. Raychaudhuri, S., J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput.: 455-66, 2000.

IV.15. Raychaudhuri, S., J. M. Stuart, X. Liu, P. M. Small, and R. B. Altman. Pattern recognition of genomic features with microarrays: site typing of Mycobacterium tuberculosis strains. Proc. Int. Conf. Intell. Syst. Mol. Biol. 8: 286-95, 2000.

IV.16. Raychaudhuri, S., P. D. Sutphin, J. T. Chang, and R. B. Altman. Basic microarray analysis: grouping and feature reduction. Trends Biotech. 19: 189-93., 2001.

IV.17. Sharan, R., and R. Shamir. CLICK: a clustering algorithm with applications to gene expression analysis. Proc. Int. Conf. Intell. Syst. Mol. Biol. 8: 307-16, 2000.

IV.18. Sherlock, G. Analysis of large-scale gene expression data. Curr. Opin. Immunol. 12: 201-5, 2000.

IV.19. Sinha, S., and M. Tompa. A statistical method for finding transcription factor binding sites. Proc. Int. Conf. Intell. Syst. Mol. Biol. 8: 344-54, 2000.

IV.20. Toronen, P., M. Kolehmainen, G. Wong, and E. Castren. Analysis of gene expression data using self-organizing maps. FEBS Lett. 451: 142-6, 1999.

V. Miscellaneous Other Papers involving Stat Methodology.

V.1. Aldous, D. (2001) Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* **16**, 23–34.

V.2. Smith, T. and Waterman, M. (1981) Identification of common molecular subsequences.

V.3. Karlin, S. and Altshul, S. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Nat. Acad. Sci.* **87**, 2264–68.

V.4. Zhang, M. (1998) Statistical features of human exons and their flanking sequences. *Hum. Molec. Genet.* **7**, 919–32.

V.5. Baggerly, K., Coombes, K. et al. (2001) Identifying differentially expressed genes in cDNA microarray experiments. Preprint, Department of Biostatistics, M.D. Anderson Cancer Center.

V.6. Wolfinger, R., Gibson, G., Wolfinger, E. et al. (2000) Assessing gene significance from cDNA microarray expression data via mixed models. Preprint, Applications Div., SAS Institute.

W. Web-sites.

W.1. Web-site for microarray datasets to analyze: **http://geaw.nci.nih**

W.2. From I.1: **http://www.americanscientist.org/articles/00articles/hamadeh.html**

W.3. Re III.3: **http://llmpp.nih.gov/lymphoma**

W.4. Re III.4: **http://www.jax.org/research/churchill** , page for Gary Churchill statistical experimental design research in microarrays.

W.5. Re III.6, web-site for T. Speed Microarray working-group (**http://oz.Berkeley.EDU/users/terry/**) includes links to several other useful sites.

W.5. Research course materials on clustering and genomics: **http://www-stat.stanford.edu/~owen/courses/399/**