Study Guide for In-class Exam; Bioinformatics II, Spring 2002

1. Give a brief description of how global gene expression data are obtained through SAGE, spotted chips, and affymetrix arrays. Highlight, and comment upon, the differences among these technologies that you think may be relevant for data preprocessing and analysis.

2. Explain the role of normalization in the preprocessing of expression data from microarrays. Based on the discussion in class and the papers you read, list a few crucial aspects to be considered in designing an effective normalization procedure. One possible paper to consider for this question is the paper you selected from the "Normalization" list for homework #1.

3. Explain the role of dimension reduction and clustering in the analysis of expression data from microarrays. What sort of questions do they allow to address? How do they differ from one another and how are they related? You don't need to provide an exhaustive comparison, select one or two aspects you find important and comment on them (e.g. for what purposes and under what assumptions is it/is it not sensible to perform dimension reduction before clustering). Possible papers to consider for this question are:

   - Holter N.S., Mitra M., Maritan A., Cieplak M., Banavar J.R., and Fedoroff N.V. (2000). Fundamental patterns underlying gene expression profiles: Simplicity from complexity. PNAS 97: 8409-8414
   - Alter O., Brown P.O. and Botstein D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. PNAS 97: 10101-10106.
   - Yeung K.Y., Ruzzo W.L. (2001): Principal component analysis for clustering gene expression data. Bioinformatics 17 (9) 762-744.

4. Discuss the advantages that mixture-based clustering may offer over standard partitioning algorithms, such as k-means, for the analysis of expression data from microarrays. The relevant papers for this question are:

   - Yeung K.Y., Fraley C., Murua A., Raftery A.E. and Ruzzo W. L. (2001), Model-based clustering and data transformation for gene expression data. Bioinformatics, 17 (10) 977-987.
   - Bartolucci F. and Chiaromonte F. (2001). Clustering of expression data from microarrays: a mixture-based approach. Preprint.

5. Discuss possible approaches to the selection of key "structural parameters", such as the number of components to retain in a dimension reduction exercise, or the number of clusters to use in a clustering exercise. Possible papers to consider for this question are:

   - Tibshirani R., Walther G. and Hastie T. (2001). Estimating the number of clusters in a dataset via the Gap statistic. Available from the course web-site.
   - Ben-Hur A., Elisseeff A. and Guyon I. (2002). A stability based method for discovering structure in dlustered data. Available from the course web-site.
   - Yeung K.Y., Fraley C., Murua A., Raftery A.E. and Ruzzo W. L. (2001), Model-based clustering and data transformation for gene expression data. Bioinformatics, 17 (10) 977-987.

6. Give a brief description of the rationale for performing motif searches in the upstream regions of genes that present similar expression profiles over a given set of conditions or a given time course. More generally, comment on reasons and possible means to integrate expression data analyses and sequence data analyses. Possible papers to consider for this question are:

   - Fujibuchi W., Anderson J. and Landsman D (2001) PROSPECT improves *cis*-acting regulatory element prediction by integrating expression profile data with consensus pattern search. Nucleic Acids Research 29 (19), 3988-3996.
   - Vilo J, Brazma A, Jonassen I, Robinson A. and Ukkonen E. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, 384-394. (Passed out in class.)
   - Chiang D.Y., Brown P.O. and Eisen M. B. (2001) Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. Bioinformatics 17 (Suppl. 1), S49-S55.

7. Explain the meaning of *position weight matrix*. Describe how a position weight matrix is derived from a collection of experimentally determined binding sites for a given transcription factor. Describe a pattern-discovery method for position weight matrices that can be applied to a set of upstream regions of genes that are thought to be co-regulated.

8. Give an interpretation of Figures 1-4 of the paper:

   - Jansen R., Greenbaum D. and Gerstein M. (2002) Relating whole-genome expression data with protein-protein interactions. Genome Research 12 (1), 37-46.

   Discuss possible explanations for the almost-nonexistant correlation between yeast two-hybrid data and expression data. What are some possible explanations for the high average correlation with some complexes but not others (e.g., see Fig. 3).

9. How would would the theoretical spectrum change if the protein subjected to MS/MS differed from the database sequence by one amino acid (a mutation)? How would a post-translational modification affect the spectrum? Outline a procedure that tries to identify mutations or post-translational modifications as it searches a protein sequence database with an observed spectrum.