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:


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:


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:

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:


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:


Discuss possible explanations for the almost-nonexistent 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 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.