PH 243A STATISTICAL TECHNIQUES FOR GENE EXPRESSION DATA, Fall 2001

ROUGH SYLLABUS

Lecture 1, Wednesday, September 5: SOME MICROBIOLOGY TERMS, THE MICROARRAY TECHNOLOGY, EXAMPLES OF DATA SETS.

Possible books for the biology and biotechnology:
1) David P. Clark and Lonnie D. Russell, 1997, Molecular Biology made simple and fun, Cache River Press, Vienna, IL 62995, USA, 1-888-862-2243.
2) Edward Alcamo, 1999, DNA Technology (2nd Edition), The Awesome Skill, Academic Press, San Diego.

Lecture 2 and 3, Monday, September 10: GENE EXPRESSION PROFILES ON A SAMPLE OF EXPERIMENTAL UNITS (e.g. subjects).
1) A formal statistical framework: Experimental unit, model, parameter.
2) Testing, selecting statistically significantly differentially expressed genes.
3) Estimation of the true subset of genes. Simultaneous confidence band. Sample size formula. Consistency.
4) Estimation of the true clusters of genes. Consistency. 5) Visualisation of clusters.
6) The bootstrap to establish the variability of the estimated subset and clusters.
7) Visualisation of bootstrap output.
8) Simultaneous clustering of subjects and genes.

Lecture 4 and 5, Monday, September 17: Same.

Lecture 6 and 7, Monday, September 24: OVERVIEW OF CLUSTER ALGORITHMS AND NEW PROPOSALS.
1) kMeans and Partitioning around Medoids (PAM)
2) Principal Components based clustering, e.g. Gene Shaving.
2) Hierarchical clustering.
3) Agglomorative clustering.
4) Hybrid clustering "HOPACH"
5) An improvement on PAM.

Lecture 8 and 9, Monday, October 1: Same.

Lecture 10 and 11. Monday, October 8: FINDING DNA-BINDING SITES of TRANSCRIPTION FACTORS BASED ON A SAMPLE OF GENE EXPRESSION PROFILES IN YEAST
1) Monte-Carlo Cross-validation to select activated binding sites in each experiment.
2) Combining results across experiments.

Lecture 12 and 13, Monday October 15: Same.

Lecture 14 and 15, Monday, October 22: GENE EXPRESSION PROFILES AND A POST-EXPRESSION OUTCOME (e.g. survival, lymph-node involvement) ON A SAMPLE OF SUBJECTS
1) Multivariate regression on all or subsets of genes (CART, linear regression) with Monte-Carlo cross-validation.
2) Marginal regressions on each gene.
3) A new supervised clustering method, and subsetting genes.
4) Bootstrap.

Lecture 16 and 17, Monday, October 29: Same.

Lecture 18 and 19, Monday, November 5: GENE EXPRESSION PROFILES AND PRE-EXPRESSION VARIABLES (e.g. time since surgery, type of cancer, type of treatment) ON A SAMPLE OF SUBJECTS
1) if pre-expression variable is randomized.
2) if pre-expression variable is confounded: Marginal Structural Models, Causal Inference.
3) Supervised clustering.

Lecture 20 and 21, Monday, November 12: LONGITUDINAL STUDIES WITH GENE EXPRESSION DATA: HOW TO DEAL WITH 1) CENSORING 2) TIME-DEPENDENT CONFOUNDING AND 3) CURSE OF DIMENSIONALITY.
Book: van der Laan, M.J., Robins, J.M. (2001), Unified Methods for Censored Longitudinal Data and Causality, to appear, Springer

Lecture 22 and 23, Monday, November 19: LONGITUDINAL (i.e. repeated over time) GENE-EXPRESSION PROFILES

Lecture 24 and 25, Monday, November 26: LONGITUDINAL (i.e. repeated over time) GENE-EXPRESSION PROFILES AND A RIGHT-CENSORED FINAL CLINICAL OUTCOME (e.g. SURVIVAL)

Lecture 26 and 27, Monday, December 3: TRYING TO GET TO THE CAUSAL EFFECTS OF GENES ON THE FINAL CLINICAL OUTCOME

Lecture 18 and 19, Monday, December 10: LONGITUDINAL GENE-EXPRESSION PROFILES WITH BASELINE VARIABLES (e.g. treatment).