

GCB 535 / CIS 535
Fall 2004

Midterm 2 Review

As before, you should also go over all the homework problems and make sure you understand the concepts behind them. Generally, you should not need to worry about equations, numbers, etc. To encourage you to study your homeworks carefully, we will be including at least one question from the homeworks on the midterm, though possibly in slightly altered form.

A rough guide to what we think you should know is:

1. Biology:

- * Don't forget your biology.

2. Gene finding:

- * ORFs, Ka/Ks, synonymous and non-synonymous mutations, positive selective pressure, stabilizing selective pressure, features useful for gene prediction
- * Using homology / sequence conservation (VISTA plots); pseudogenes; homologs, orthologs, paralogs
- * Linear regression – how to use, how to interpret the model, how to assess statistical significance of the model parameters
- * Artificial Neural Networks; the effects of increasing the number of parameters (overfitting)
- * Hidden Markov Models – components, underlying assumptions,

3. Microarrays:

- * The biology behind gene expression analysis
- * Northern blots
- * What kind of info can we get from Microarray experiments?
- * What's direct and indirect labeling?
- * Spotted arrays (need to use a reference sample) vs Affymetrix arrays (need mismatch probes) – two color vs one color
- * Experimental design – choosing conditions, replicates; dye swap, loop design, common reference; replicates
- * Normalization
- * Gene annotation
- * Clustering – hierarchical (single, average, complete linkage), k-means; PCA; k-nearest neighbors; assessing significance of gene clusters

4. Gene regulatory networks:

- * The biology of regulatory networks
- * Belief nets, Bayes' rule, conditional independence
- * Verification – out of sample tests, cross-validation

5. Proteomics:

- * The proteome
- * Protein separation using gels -- 2-D PAGE gels, limitations of 2-D PAGE
- * liquid chromatography based on chemical characteristics
- * Mass spectrometry to identify proteins (basic procedure, interpreting the results); trypsin digest; m/z; using isotopes to determine charge of the peptide
- * Tandem mass spec (basic procedure, interpreting, b- and y-type ions)
- * Peptide databases; sequence tag, mass fingerprinting
- * Database search - the general approach behind programs like Sequest and Mascot (but none of the details about MoWSe score, cross correlation, fourier transforms)
- * Genomic Peptide Finder – general approach

Fun Questions

1. What's the null hypothesis for the statistical test that examines whether one gene list over-represents a certain type of genes?
2. What is overfitting and why is it a problem?
3. In a tandem mass spectrogram, would you expect to be able to identify equal numbers of b- and y- type ions?
4. Explain why we do dye-swapping in spotted microarray experiments.
5. Explain the difference between a discriminative machine learning model and a generative one. Which term best describes HMMs? Neural Nets? Linear Regression?
6. What are the two dimensions in a 2-D PAGE protein gel, and how does that facilitate protein separation?
7. What is a post-translational modification, and how does it complicate interpreting the peaks on a mass spectrogram?
8. Hierarchical clustering allows the user to end up with a very small number of clusters or a very large number of clusters, depending on where the cutoff line is drawn. Describe the pros and cons of ending up with a large number versus a small number of clusters.
9. What's the advantage of Microarray techniques over using a Northern blot? What biochemical property of DNA molecules serves as the technical foundation for microarrays?
10. What is cross-validation and why is it important?