STAT 5570/6570 Statistical Bioinformatics Spring 2014

Possible Discussion Topics

Topic 1

A study is conducted to evaluate differences in the human gut induced by differences in diet. Twenty subjects are randomly assigned to two diets (ten subjects each), and fecal samples from each subject (before and after a set time period on the assigned diet) are obtained. Using these samples, measures of the abundance of hundreds of bacteria species are obtained. A phylogenetic tree of all these hundreds of bacteria species is available. The objective is to identify regions (branches) of this phylogenetic tree that are significantly affected (up or down) by differences in diet.

Topic 2

In a separate study of many participants, the abundance of hundreds of bacteria species in many participants were obtained. Are there unique combinations of bacteria species? How many?

Topic 3

Given a matrix of pairwise distances between subjects and a single covariate (like Trt/Ctl) on each subject, how to quantify and test the covariate's effect (or dependence) on distance?

Topic 4

Is adjustment for multiple hypothesis testing really necessary?
(Read: Konishi 2011, "Microarray test results should not be compensated for multiplicity of gene contents", http://www.biomedcentral.com/1752-0509/5/S2/S6)

Topic 5

Consider a gene expression experiment involving multiple treatment groups – say T1, T2, T3. We have used contrasts (single and multiple) to define tests of differential expression between groups. What if we allow the claim of differential expression to be multi-dimensional (or multivariate)? For example, it is possible for a gene's expression to be significantly higher in T2 than T1, and significantly higher in T3 than in T2. We can use contrasts to identify individual genes satisfying such specific multivariate differential expression. But how can this be extended to gene set testing? (For example, we want to find GO-BP terms that are significantly more active in T2 than in T1, and significantly more active in T3 than in T2.)

Topic 6

When testing hundreds of GO terms for differential activity, how can a meaningful error rate be controlled while accounting for the known structure (dependency) of the GO graph?

Topic 7

In Notes 3.6 (Case Study part 1), we identified candidate genes that were differentially expressed between depressed and non-depressed HIV patients (recall p-value histogram at right). Can we also characterize these genes?