Methods for comparing multiple microbial communities.

james robert white, whitej@umd.edu

Advisor: Mihai Pop, mpop@umiacs.umd.edu

October 1st, 2007

Abstract

We propose the development of new software to statistically determine differentially abundant taxa between two populations. Using only randomly selected 16S rDNAs from environmental samples, our goal is to assign each sequence to its appropriate taxon and analyze a taxa abundance matrix to find significantly overrepresented or underrepresented groups between two populations. Our problem is analogous to finding differentially expressed genes, and we aim to modify and implement methods already in practice in the microarray community. Specifically, we shall use the **false discovery rate** (FDR) and its corresponding measurement, the **q-value**, to control the number of false positives that frequently occur when performing many hypothesis tests.

Background

As the field of metagenomics continues to explode, an increasing number of studies focus on species identification and diversity within particular environments. Methods of comparing multiple environments have been largely based on small subunit ribosomal RNA (SSU rRNA), particularly, the 16S rDNA gene. This gene is well conserved among species and found in all known microbes. Due to its high rate of conservation, this gene can be easily read by scientists, and acts as a tag for each organism. By randomly selecting and reading these genes from an environment, one can measure the relative abundances of each species within the environment.

A multitude of metagenomics projects has led to statistical software tools such as DOTUR (Schloss and Handelsman 2005), S-libshuff (Schloss et al. 2004), SONs (Schloss and Handelsman 2006a), UniFrac (Lozupone and Knight 2005), and TreeClimber (Schloss and Handelsman 2006b). Though these packages provide some information about community structure overlap and phylogenetic diversity, they are not designed for comparing hundreds of different communities simultaneously, and often fall short of providing researchers with enough information to make conclusions about environment composition.

We seek to find out not if two populations are different, but exactly how they differ. Our objective is to determine which species in two populations are differentially abundant, that is, make up different proportions of the organisms in the environments. Our problem is analogous to finding differentially expressed genes between two populations, a problem that has been researched for the past 10 years. Consequently, statisticians have developed new methods for performing many hypothesis tests simultaneously.

A result of a typical statistical test is a p-value, a measurement of confidence for rejecting or accepting a null hypothesis. Each test has an individual p-value, and usually a threshold α is imposed to reject a subset of all tests. Thresholding by α means that we reject any test with a p-value less than α, which means that we expect a fraction of α of all significance tests to be false positives (type I error). If we are performing 1000 significance tests, then if $\alpha = 0.05$, we expect 50 false positives. This is far too high for our experimental methodologies to handle.

Recently, statisticians have succeeded in reducing the number of false positives by using the **false discovery rate**, which is defined to be the proportion of rejected null hypotheses that are false positives (Benjamini and Hochberg 1995). For each test, there is an individual measurement of the false discovery rate known as the **q-value** (Storey and Tibshirani 2003). The difference between the q-value and the p-value is subtle but important. Computing these two types of values ranges in difficulty. Some methods of generating p-values involve large-scale permutation algorithms. The algorithm for computing q-values uses p-values. Thresholding by q-values instead of p-values leads to many fewer false positives, thereby improving later analyses.

Strategy

The goal of this project is to apply the false discovery rate to metagenomic analysis in order to determine differentially abundant taxa between two environment populations. We shall develop software that takes a species abundance matrix as input, and outputs a fully automated analysis of this matrix, isolating differentially abundant taxa using the q-value algorithm.

The most intensive computational part of this software will likely be the generation of p-values, which can involve many permutations. Specifically, if we are performing B permutations on M taxa, we will need to create a BxM matrix to store results. However, no operationally expensive procedures are used on this matrix, so our main concern will be space. The q-value algorithm is non-trivial, but does not require large computational resources.

Hierarchical clustering algorithms will be implemented in order to cluster differentially abundant taxa. Additionally, I will cluster species into higher groups such as phyla, orders, families, etc. and perform hypothesis tests on these larger categories of life. Sometimes at species resolution, life abundances are too specific, and only when one views from high categories can the true pattern be detected.

Preliminary coding will be done in Matlab or R as proof of concept. Since most biological researchers do not have Matlab or the expensive additional packages it often requires, it may be better to start in R and transition to C++. R is a free statistical software package with many visualization features. We aim to make our software extremely easy to install and operate on any Unix platform.

Resources

The Center for Bioinformatics and Computational Biology (CBCB) on the University of Maryland campus has more than adequate computational resources. Large-scale computation for validation trials may be performed (if necessary) on one of several dual or quad processors with 8GB or 32GB RAM, respectively. Regular computation will be performed on either a Dell system (Linux) with dual processors and 3.0GHz (2GB RAM), or a MacBook Pro (Mac OS X) with a dual core processor and 2.2GHz (2 GB RAM). Our software should be able to run quickly on these machines, and should be accessible through any Unix platform.

Validation

Validation of statistical calculations will use SAS or R (statistical software packages). To validate the method, we shall apply our software to a classic set of microarray experiments that are commonly used in the analysis of new methods (Hedenfalk et al.

2001). After predicting differentially expressed genes from this dataset, we will compare to other predictions by widely accepted programs, including ones that use methods similar to our own. A large number of ubiquitous differentially expressed genes will validate our methods.

Finally, we shall apply our software to a set of metagenomic samples from obese and lean human subjects (Ley et al. 2006). We shall see if our method reveals the same important conclusion as the original study. Using our software for this type of analysis will save scientists a great deal of time.

Proposed schedule

2007

September

• Create project concept.
• Meet with Dr. A. Zimin and Dr. R. Balan to approve project.

October

• Present project proposal (20 minutes).
• Finish project proposal.
• Acquire validation datasets (Hedenfalk and Ley studies).
• Acquire microarray analysis packages.

• Implement standard multiple hypothesis t-test procedure for input species abundance matrix.
• Implement non-parametric method for non-normal t statistic estimation (p-value estimation).
• Begin implementation of clustering algorithms.
• Complete implementation of q-value algorithm by the end of October.

November

• Complete clustering algorithms for differentially abundant taxa.
• Complete clustering algorithms for higher orders of life.

December

• Finish statistical calculation validation using SAS or R.
• Deliver midpoint report.
• Midpoint presentation (20 minutes).

2008

January

• Complete transition of software to C++ code if not already done so.
• Consider statistical methodology given sampling issues.

February
• Develop documentation for software.
• Begin final validation procedure for microarray data.

March
• Apply software to human gut metagenomic sampling data.
• Begin final report write-up.

April
• Complete final draft of report including edits from advisor.

May
• Deliver final report.
• Final presentation (40 minutes).

References

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met 57(1): 289-300.

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M et al. (2001) Gene-expression profiles in hereditary breast cancer. The New England journal of medicine 344(8): 539-548.

Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444(7122): 1022-1023.

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology 71(12): 8228-8235.

Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and environmental microbiology 71(3): 1501-1506.

Schloss PD, Handelsman J (2006a) Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. Applied and environmental microbiology 72(10): 6773-6779.

Schloss PD, Handelsman J (2006b) Introducing TreeClimber, a test to compare microbial community structures. Applied and environmental microbiology 72(4): 2379-2384.

Schloss PD, Larget BR, Handelsman J (2004) Integration of microbial ecology and statistics: a test to compare gene libraries. Applied and environmental microbiology 70(9): 5485-5492.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100(16): 9440-9445.