

Barcoding Bacteria

Ryan Hayman
Chemistry

Gabriel Wachman
Computer Science

December 18, 2006

Chemistry/Biology Problem

Advances in genotyping assays have resulted in commercially available, cost-effective whole-genome analyses. Once relegated to major genomics centers, genotyping capabilities are now spreading to smaller academic and industrial labs worldwide. With vast numbers of microbial sequences published in on-line databases, there is increasing interest in not only detecting the presence of pathogens, but also determining their relative virulence. Strain differentiation is a prime application for genomic barcoding, whereby microarrays are used to detect a pattern, or signature, based on the presence or absence of multiple genomic sequences. Commercially available software is used to import and align tens of sequences, output as a text file. It is highly time consuming to visually search for informative sections in these alignment files, as they are often thousands to tens of thousands of characters long. Manually developing a phylogenetic tree based on the classification calls of individual SNPs is very inefficient and prone to errors. A computer program designed to choose SNPs from alignment files would greatly simplify the genotyping assay development process.

Computer Science Problems

There are two distinct problems on the computer science side. The first is to determine the best way to identify uniquely a bacterial strain based on a sequence of nucleotides from its genomic DNA. Historically, the chemist has been performing this task in a partially automated fashion that was extremely time consuming. A solution to this problem would free months of the chemist's time in which he could be doing actual experiments. The second problem is to determine if a new, previously unclassified, bacterium is harmful or harmless based on its similarity to known strains.

Computer Science Solution

We have designed a working to solution to the first problem by applying a decision tree algorithm as follows:

1. The DNA sequences are transformed into feature vectors indexed by character positions in the sequence. Each feature is hence a character value for a particular position.
2. Each bacterial strain is given a unique class name.
3. Duplicate feature vectors are removed from the dataset.
4. The decision tree algorithm Id3 is run so that no pruning is done and multi-way splits are used.

Because each *Listeria* strain has a unique class name, and since there are no duplicates, the decision tree puts exactly one strain at each leaf. Note that Id3 uses information gain as a splitting criterion, and while this is not intuitively useful here since we are trying to overfit, it will favor a shorter tree resulting in more efficient use of genomic assays with a fixed number of decisions.

We have successfully implemented and tested this solution on a subset of the data. The remaining data are not in a standard format, and there are sections of DNA that are missing from some strains but present in others. We will have to determine whether we want to deal with such sequences by using a standard technique for missing features, or by simply removing these sequences from all strains.

Since the chemist has requested an *Excel* spreadsheet as the final format of the data, we want to explore implementing the entire process in Visual Basic so that the chemist can simply load a spreadsheet of DNA sequences, run a macro, and make immediate use of the resulting spreadsheet.

Currently, our process does not take into account the case where a chemist tests for the presence of a particular nucleotide in a specific position, and the test itself fails. In this case, having only one decision tree would require performing the chemical test again, which is expensive and time consuming. It may be possible to differentiate the bacteria in several ways, allowing the chemist to incorporate redundancy into the assays. A modification to the standard Id3 algorithm could simply implement every possible decision tree and then merge them.

We have not worked on the second problem yet as there are a few obstacles to overcome. First, there are only 30 strains of bacteria to work with, and as such it may be unrealistic to generate any kind of classifier that we can expect to generalize well to future examples. Second, the data have not yet been tagged as “harmful” or “harmless,” as this will require an extensive literature search. Should we be able to overcome these two obstacles, there are numerous machine learning algorithms that would be able to generate useful classifiers, if one exists.

Implications for Chem / Bio Solution

The program developed will greatly simplify the ability to develop barcodes based on two-color SNP genotyping assays. Months of manual work can be accomplished in *under 1 second*, saving time for the chemist to develop more complicated and informative barcoding schemes or to create multiplexed assays with multiple barcodes. The ability to detect and differentiate tens of pathogens on a single optical fiber-based assay will be critical to accomplish goals set for an important collaborative saliva diagnostics grant (NIDCR/NIH).

Next Steps

Several classification problems will be used to test the program. The genes responsible for the virulence of other bacteria such as *E. coli* and *Salmonella* spp. have been more thoroughly studied and will be ideal for follow-up work. A novel barcode for *H. influenzae*, an important pathogen in chronic inflammatory respiratory diseases such as Asthma, COPD, and emphysema, will be developed based on tens of 16S RNA sequences.