Introduction

A human genome typically contains close to 25000 genes. The genome includes both the [genes](http://en.wikipedia.org/wiki/Gene) and the [non-coding sequences](http://en.wikipedia.org/wiki/Non-coding_DNA) of the DNA. Microarray technology evolved from [Southern blotting](http://en.wikipedia.org/wiki/Southern_blotting), where fragmented DNA is attached to a [substrate](http://en.wikipedia.org/wiki/Substrate_%28biochemistry%29) and then probed with a known gene or fragment. The goal of a microarray experiment is to simultaneously measure the expression levels of thousands of genes in a collection of cells.

Microarray is a powerful tool for genome analysis. It gives the global view of the genome analysis in a single experiment. Data analysis in the Microarray is a vital part as this part influences the final result. Each microarray study comprises multiple microarray experiments, each microarray study would give tens of thousands of data points. Since the volume of data growing exponential, the analysis becomes a challenging task. In general the greater the volume of data, the more chances arise for erroneous results. Handling such large volumes of data requires high end computational infrastructures and programs that can handle multiple data formats.

Selection of relevant genes for sample classification is a common task in most gene expression studies, where researchers try to identify the smallest possible set of genes that can still achieve good predictive performance (for instance, for future use with diagnostic purposes in clinical practice). Many gene selection approaches use univariate (gene-by-gene) rankings of gene relevance and arbitrary thresholds to select the number of genes, can only be applied to two-class problems, and use gene selection ranking criteria unrelated to the classification algorithm. In contrast, random forest is a classification algorithm well suited for microarray data: it shows excellent performance even when most predictive variables are noise, can be used when the number of variables is much larger than the number of observations and in problems involving more than two classes, and returns measures of variable importance. Thus, it is important to understand the performance of random forest with microarray data and its possible use for gene selection.

Learning

Model evaluation

Feature Selection and evaluation

a) Filter: t-test, F-test, kruskal- test, etc

Feature Selection and

evaluation

Learning

Model evaluation

b) Wrapper: Binary Particle Swarm Optimization, Genetic Algorithms, etc

c) Embedded: rfeSVM, random forest (trees), elastic net, lasso, etc

Learning

Learning

Subset generation

Model evaluation

Feature Selection and evaluation