**BBSI Proposal Summer 2007**

Implementing Significant Analysis of Microarrays (SAM) into BioBike language

Andrew Whittam
Washington & Jefferson College
Information Technology Leadership

Introduction

In recent years microarrays have become the standard way to measure gene expression in cells when exposed to different environmental changes. Since a microarray allows thousands of genes to be measured at the same time the dataset produced is quite large. When faced the overwhelming amount of data produced, scientists must filter their results to draw any conclusions. One such desired test would to be able to distinguish between significant variations of a gene expression due to a tested environment with that of random change. For normal experiments with high repetitions a t-test would be used to determine the significant changes. To take a normal t test of the huge amount of data produced by microarry and with so few repetitions will produce skewed results. The Significant Analysis of Microarrays (SAM) described in a paper by Tusher, Tibshirani and Chu works to improve determining significant differences in individual genes. This method returns a set of genes who's expression had changed significantly changed by taking a t-test f each individual gene. To be able to use the results of SAM and run them through other tests or look at the gene sequences for similarities would allow for more conclusions to be drawn. To be able to use many different free standing programs that apply methods and produce sequences would allow scientist to better understand and apply their results. BioBike is a visual programming language that caters to non programmers and allows them the ability to produce

programs.  BioBike uses both original programming and already exiting programs in concert to allow users to tailor their own programs and experiments to meet their needs. To introduce both microarrays and SAM to BioBike would provide its users more abilities and hopefully lead to more breakthroughs.

## Method

To incorporate the SAM program into BioBike would open up the frontier to all type of microarray analyses and methods.  Stanford labs has created a standalone program for SAM written in the computer language R.  The goal of my summers work is to incorporate this previously written code and algorithm into BioBike.  This will be accomplished through the strip down of the original program into basic input / output program that BioBike would call upon using a shell.  The user would be asked for a few variables such as tolerance and output style, this allows tweaking the method to better suit the data.  Also perhaps an option to display the graph showing the tolerance and outlier genes may be an option.   The miccroarray data stored on the BioBike sever would all have the same format which the SAM function would take in and then return the varying genes to the user.

## Results

Though the major focus of my summer work will be in the implementation of SAM into BioBike the hope is to someday have many different options for users to choose between.  SAM is very popular but it is not the only way to generate a list of significantly changed genes. Along with showing difference in gene expression, normalizing the data and reducing background noise are other possible tool additions to

BioBike.  Through the addition of significance analysis of microarrays to the BioBike repertoire we are improve not only the capabilities of the programming language but also its ability to apply algorithms to different types of data, such as microarrays.

References:

http://www-stat.stanford.edu/~tibs/SAM/ - SAM Website

http://ramsites.net/~biobike/ - BioBike Website

Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu

Proceedings of the National Academy of Sciences, U.S.A. 98:5116-5121.