

## 2HAPI, A Comprehensive, Internet-Based Microarray Data Analysis System

*J. Lynn Fink, Michael Gribskov*

*University of California, San Diego / San Diego Supercomputer Center*

*jlfink@sdsc.edu, gribskov@sdsc.edu*

### Abstract

Simultaneous measurement of the expression levels of thousands of genes has been made possible by microarray technology. With this technology, we are able to study gene interactions and determine individual gene functions on a genomic scale. This is of particular interest because of the abundance of data gathered from genome sequencing projects. However, the amounts of data generated by microarray use are so large that computational analytical tools are necessary. Several companies and researchers have developed software to fulfill this need, but most of these packages are limited in their analytical capabilities and hinder an integrated approach to analysis. These packages can also be constraining due to hardware and software requirements and, in the case of commercial software, the high cost of licenses. 2HAPI, presented here, is a system for computational microarray data analysis that attempts to create an integrated analytical environment that is highly accessible, fully-featured, and free to academic users. 2HAPI will include a user interface that mediates data upload, exploration, and transformation. Since the field of microarray data analysis is just starting to develop, it is unknown which algorithms are most appropriate for this type of data. The system will therefore include both clustering algorithms that have been previously applied to gene expression data as well as existing algorithms that have not yet been used for gene expression applications. 2HAPI will also include tools that allow the user to create subsets of the data based on criteria of interest and visualization features that describe the expression patterns and any clusters or subsets to which they belong. In order to provide the user with additional information about the genes included in an experiment, links to ancillary data such as GenBank, sequence analysis tools, and other informational databases will be displayed. Underlying these tools and algorithms will be a relational database that contains the data and the various manipulations performed on it by the user. 2HAPI will be internet-based to provide maximum accessibility. 2HAPI is also designed with the notion that the user need not be a computer scientist or statistician.

In order to demonstrate the capabilities of 2HAPI, the system will be tested with real data sets. The publicly available yeast cell cycle oligonucleotide array data has been analyzed by multiple groups. These analyses will provide a useful standard for comparison of 2HAPI's analytical capabilities. A novel data set involving *Mycobacterium tuberculosis*-infected human macrophages will also be used in order to demonstrate that will 2HAPI can provide a modular and extensible way for researchers to analyze microarray data.