

# Needs Assessment for Scientific Visualization of Multivariate, High-Dimensional Microarray Data

Vetria L. Byrd,<sup>1</sup> and Tarynn M. Witten<sup>2</sup>

<sup>1</sup>Department of Computer Science, University of Alabama at Birmingham, 1300 University Boulevard, 115 Campbell Hall, Birmingham, AL 35294-1170

<sup>2</sup>Center for the Study of Biological Complexity, Virginia Commonwealth University, Suite 111, Trani Life Sciences Center, PO Box 842030, 1000 West Cary Street, Richmond, VA 23284-2030

## ABSTRACT

The availability of genomic data is increasing exponentially. This is magnified by the proliferation of data at all “omic” hierarchy levels. This explosive growth in biological data (currently GenBank contains over 44 billion base pairs and over 40 million sequences) mandates an increasing need for sophisticated mathematical and computational methods and software environments capable of handling the complexities and sizes of these various “omic” datasets. In particular, this is also true for microarray data.

Microarray technology allows for the simultaneous genomic analysis of entire organismal genomes. The resulting datasets are high-dimensional, complex and frequently difficult to interpret. In order to address these microarray dataset and software needs, we have first decided to examine the need for and the subsequent design of advanced microarray data analysis software tools that will allow researchers to use new means and methods of visualizing and analyzing their microarray data. Towards this goal, a survey research instrument entitled “Needs Assessment for Scientific Visualization of Microarray Data” was created and distributed ( $n = 500$ ). The survey was submitted to and approved by Virginia Commonwealth University’s Institutional Review Board (VCU IRB#5065). The survey research instrument was distributed to a non-random sample set of researchers and biomedical life scientists

currently using microarray methods in their day-to-day research. The results of the survey will be statistically analyzed and are anticipated to be instrumental in identifying a set of algorithmic/software needs to be added to the currently available microarray software analysis toolsets.

## INTRODUCTION

Advances in microarray technology not only initiated a change in the way biological experiments are performed and analyzed, but this new technology also created both a need and a demand for visualization tools and techniques that would allow researchers the ability to gain further insight into the enormous amounts of data that is generated from each microarray-based experiment. Microarray technology allows for the simultaneous analysis of several genes or entire genomes at a time, making the resulting dataset inherently high-dimensional and complex. The need to develop and use scientific visualization software packages and methods for microarray data analysis can be seen in the already large number of applications and tools developed in this area. A cursory literature and web search yielded over 100 applications and tools in this area.

As we have just pointed out, there are numerous visualization tools currently available for the analysis and visualization of multivariate microarray analysis. These tools are available in varying software

applications. While not all tools are available in one software package, there are some tools that are common to most. Saraiya *et al.*, (2004), evaluated five microarray visualization tools for their analytical and visual capabilities. Among all of the software packages evaluated, their results showed that clustergrams, parallel coordinates, heat maps, scatterplots and histograms were the most commonly used visualization modes for microarray data.

The sheer size of microarray data generated by each microarray experiment strongly suggests the existence of a need for more advanced, more highly integrated software tools that will allow researchers to ask deeper, more biologically profound questions as well as to allow the investigator a gateway to explore the data in ways not initially suggested by the standard visualization tools. It is also evident that microarray data contains answers to biological questions that haven't been asked or even formulated. Although extremely useful, even the most common visualization tools don't allow for exploration of the entire microarray data set or for questions that do not fit within the domain of the current software analysis capabilities. Parallel coordinate graphs can become cluttered and difficult to interpret when used to visualize large datasets (Bhasi *et al.*, 2004). Heat maps are a form of hierarchical clustering that could "send biologists down blind alleys if they see visualizations that reinforce their own assumptions about relationships between individual genes (Tilstone, 2003).

The goal of this research project is to enable researchers to assess and visualize high dimensional, multivariate microarray data in a way that is equivalent to (or improved upon) current state of the art techniques for visualizing other types of advanced data.

## METHODS

The goal of this research project was to assess existing visualization tools, determine what - if any - unmet visualization needs exist and to subsequently develop a tool/method/software environment to aid in the analysis and visualization of these complex data sets. An exhaustive evaluation and/or assessment of existing microarray software currently available as freeware, shareware, open source and commercial packages is well beyond the scope of this paper. The reader is directed to Liu *et al.*, 2004 and Saraiya *et al.*, 2004 for some software reviews. Towards achieving our research goal, we designed survey research instrument. The survey design is detailed in the next section.

### Survey Design

A survey research instrument entitled Needs Assessment Survey for Scientific Visualization of Microarray Data (VCU IRB# 5065) was designed and distributed to  $n = 500$  potential participants chosen from life and biomedical/health scientists, research institutes, biotech companies and other researchers who work with microarray data analysis. Sampling methodology will be discussed in a section below.

Participation in the survey research study is completely voluntary and there is no way for any of the information provided to be traced back to an individual participant. Qualitative and quantitative survey questions were designed to allow users to freely express their thoughts regarding what additional tools and software might be needed to assist the survey participants in their analysis of microarray data. Participants were free to skip any questions they did not wish to answer. If they chose not to complete the survey they were asked to write down a

couple of reasons why they chose not to complete it and return the blank survey with their comments attached. The survey consisted of 24 questions organized into the following four sections: Demographics, Computing Environment, Microarray Technologies, and Microarray Analysis Tools.

The Demographics Section was designed to gather basic, non-identifiable participant demographic data such as: highest level of education, degree, age, primary job title, *etc.*, To ascertain the degree to which microarray analysis is performed, participants were asked to indicate how long they have been working with microarray analysis, how long they have been using microarray analysis tools as well as how their current microarray analysis tools are used for image analysis (Chen & Liu 2005, Yang et al., 2001), data mining (Gardiner-Garden 2001), annotation (Khatri 2005) and/or for statistical analysis (Pan 2002).

The Computing Environment Section asked participants to indicate which operating system(s) are currently in use in his/her working environment.

In the Microarray Technologies Section, participants were asked to rank microarray analysis levels (probe level, expression level cellular level and transcriptomic (mRNA) level) according to their interest using a Likert scale range from 1 (most interested) to 4 (least interested). Participants were asked to indicate if spotted cDNA and/or oligonucleotide microarray technologies are used and which of the technologies is the primary microarray technology in use in their microarray analysis environment.

The Microarray Analysis Tools Section consisted of questions designed to determine what image analysis, database, annotation

tools, integrated packages and specific packages are used and of these tools and packages which ones were the primary tool(s) packages in use. We also wanted to know how users visualize their microarray data. To determine the types of visualization tools in use participants were asked to rank order on a Likert scale of 1 (most used) to 10 (least used) visualization tools in terms of their frequency of use. Common visualization tools mentioned in the survey include such tools as parallel coordinates, heat maps, scatter plots, histograms, array layouts, pathways, gene-to-gene comparison and cluster dendrogram, for example. While these tools are extremely valuable in their ability to assist researchers in visualizing and interpreting data, we hypothesized that there is an unarticulated or possibly even unrealized need for a visualization tool/tools or visualization function(s) that is/are not currently available in microarray analysis and visualization software. Participants were asked to indicate if they are able to visualize multivariate, high-dimensional data sets to their satisfaction using the software and visualization tools currently available to them. If they indicated they were not able to do so, they were asked to describe the desired tool(s)/function(s) that they felt would allow them to visualize multivariate, high-dimensional data sets to their satisfaction. Considerable amounts of space were provided in order to allow participants to describe what visualization features were missing from their current software packages and, if they could design their own software visualization package what would be the most important features and/or functions they would include. Additional space was provided to allow and to encourage participants to make any additional comments and suggestions they would like to make.

## Sampling Protocol

In order to obtain a cross-section of the microarray user/analyst population, potential survey targets were selected from different categories of users that represent the bulk of general microarray users and general microarray analyst's working environment. From our review of the, we concluded that scientists who are currently using microarray methods in their day-to-day research can be categorized by organization: research universities, research institutes, national laboratories, and biotech companies.

The search for survey participants in each organization began with an Internet web search with the organization as the search criteria.

BioSpace (<http://www.biospace.com>), a web site that highlights clusters of life science industries was an initial source for Biotech and Research Company survey targets. Additional Biotech and Research Companies were selected from The Virginia Bioscience Directory (Virginia Biotechnology Association, 2003-2004). The study and use of microarrays encompasses not only the obvious fields like molecular biology and genetics but, "its ability to profile changes in gene-expression levels under different conditions makes microarrays the method of choice in many fields" (Fadiel and Naftolin 2003). Survey participants chosen from research universities and research institutes were selected from various fields/departments that use microarray technology like molecular biology, genetics, forensic sciences, to name a few. A similar method was used to select survey participants from national laboratories.

Over 600 potential survey participants were selected as described above. Of the 600+ survey targets, 500 were randomly selected

to receive the survey. Ideally, we wanted to have an equal number of participants from each organization. However, because the source of participants were various web sites, not all designed and formatted for easy access and extraction of contact information, the number of selected survey participants were not evenly distributed across all organizations. Table 1 shows the number of survey participants obtained from each organization category.

<b>Organization</b>	<b>No. of Survey Targets</b>
Research Universities	311
Research Institutes	42
National Laboratories	264
Biotech Companies	54
Total	671

The survey research project IRB proposal was approved for a survey size of 500 survey participants. A C++ computer program was written to read in all survey participants, using vectors to separate them by organization. The program counts the number of records read, prompts the user for the target sample size (in our case  $n=500$ ) and based on this information determines how many survey targets should be selected from each organization that would result in an equal distribution across organizations. If the number of targets needed from each organization is greater than the sample size for a specific organization the program assesses the sample size of all organizations and determines which, if any, organizations can be over sampled to meet the target sample size indicated by the user. Table 1 shows research institutions and biotech company organizations were noticeably under sampled compared to research universities and national laboratories. Because this was known a priori, along with the sample target size ( $n = 500$ ) the software was written to over sample university and

national laboratory organizations to make up the difference from the research institutes and biotech companies. Table 2 shows the number of survey participants selected from each organization after over sampling.

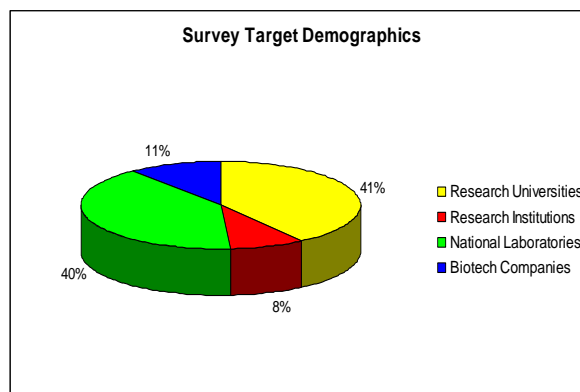
<b>Organization</b>	<b>No. of Survey Targets</b>
Research Universities	202
Research Institutions	42
National Laboratories	202
Biotech Companies	54
Total	500

Snowball sampling was used to acquire additional participant names. Survey recipients were asked to communicate, to the project researchers, the names of other individuals who either perform microarray analysis or have individuals in their labs who do microarray analysis. The snowballing technique comes at the expense of introducing further bias, as the technique itself reduces the likelihood that the sample will represent a good cross-section of the user population.

In addition to the inherent bias of the snowball sampling method, all selected participants have a vested interest in the outcome of the results, thus adding an additional bias to the sample population. In essence, the sample is not representative of the complete microarray user base. Rather, it is representative of only those participants who responded in some way to the survey. The implemented sampling method does not provide a statistically accurate view of the microarray user population within the chosen sampling block. An alternative way to view this research survey is to see it as a “paper” version, at a larger scale, of a focus group.

## RESULTS

The Needs Assessment Survey for Scientific Visualization of Microarray Data was approved by the Virginia Commonwealth University Institutional Review Board (IRB# 05065) and mailed to  $n = 500$  randomly selected survey targets. Results of the survey are pending. Basic survey target organizational affiliation demographics are illustrated in Figure [1] of this paper.



**Figure 1 Survey Participant Demographics**

## DISCUSSION

The goal of this research was to assess the current scientific visualization needs of the microarray user community in order to provide them with a medium through which they could contribute their experiential input to the development of the next generation of scientific visualization tools designed to address their data visualization needs. Surveys were distributed to scientists and to researchers who have a vested interest in the potential results of the survey; useful tools that will help in their daily analysis of complex data sets. Although results of the survey are pending, it is our hope that the responses will be insightful and provide an indicator of what the microarray user base needs in order to address their data visualization and analysis needs. The survey was distributed to a cross-section of

microarray users representing researchers from universities, national laboratories, biotech companies and research institutions.

It is anticipated the duration of the survey response phase will be from two to three months long. Once collected, the responses will be statistically analyzed using SPSS software. The results will be published in a peer reviewed journal.

## ACKNOWLEDGEMENTS

We would like to thank all of the respondents of the survey without who this research effort would have been for naught. We would also like to thank Paul Fawcett for his numerous discussions. We would like to acknowledge support of this project through grant EEC0234104 from the NSF/NIH Bioinformatics and Bioengineering Summer Institute Program. We would also like to thank the VCU postal group for all of their assistance in getting the survey out the door.

## REFERENCES

Bhasi, K, Zhang, L., Zhang, A., Ramanathan, M. (2004). Analysis of Pharmacokinetics, Pharmacodynamics, and Pharmacogenomics Data Sets Using VizStruct, A Novel Multidimensional Visualization Technique. *Pharmaceutical Research* 21(5):777-787.

BioSpace. Web links:  
<http://www.biospace.com>  
[http://www.biospace.com/company\\_index.cfm](http://www.biospace.com/company_index.cfm)

Chen, Z., and Liu L. (2005). RealSpot: software validating results from DNA Microarray Data Analysis with Spot Images. *Physiol. Geonomics* 21:284-291.

Fadiel, A., and Naftolin, F. (2003). Microarray applications and challenges: a vast array of possibilities. *International Archives of Bioscience* 1111-1121.

Fink, A. How to Design Surveys. The Survey Kit. SAGE Publications, Thousand Oaks, CA (1995).

Gardiner-Garden M., and Littlejohn, T. G. (2001). A comparison of microarray databases. *Briefings in Bioinformatics* 2(2):143-158. (Abstract)

GeneTown (2005).  
[http://www.biospace.com/genetown\\_map.cfm](http://www.biospace.com/genetown_map.cfm)

Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* (Advanced access published June 30, 2005).

Liu, D. K., Yao, B., Fayz, B., Womble, D. D., and Krawetz, S. A. (2004). Comparative Evaluation of Microarray Analysis Software. *Molecular Biotechnology* 26(3):225-232.

Pan. W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18(4):546-554.

Saraiya, P., North, C., Duca, K. (2004). An evaluation of microarray visualization tools for biological insight. *IEEE Symposium on Information Visualization*.

Tilstone, C. Vital statistics. (2003). *Nature* 424,610-612.

Virginia Bioscience Directory (2003-2004)  
Virginia Biotechnology Association.

Yang, Y.H. Buckley, M. J., and Speed, T. P.  
(2001). Analysis of cDNA microarray  
images. Briefings in Bioinformatics  
2(4):341-349. (Abstract)