

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

Vetria L. Byrd

Department of Computer and Information Sciences  
University of Alabama at Birmingham  
Birmingham, AL, U.S.A.

Tarynn M. Witten

Center for the Study of Biological Complexity  
Virginia Commonwealth University  
Richmond, Virginia, U.S.A.

*Abstract – The 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 [1] and software environments capable of handling the complexities and sizes of these various “omic” datasets [2]. This is particularly true for microarray data. Microarray technology allows for the simultaneous genomic analysis of entire organismal genomes [3] [4]. The resulting datasets are high-dimensional, complex and frequently difficult to interpret [5]. We decided to examine the need for advanced microarray data analysis software tools. A survey research instrument entitled “Needs Assessment for Scientific Visualization of Microarray Data” was created. The survey research instrument was distributed to a non-random, snowball sample set of researchers and biomedical life scientists currently using microarray methods in their day-to-day research. Results of the survey revealed microarray users are not satisfied with visualization tools that are currently available.*

Keywords: microarray tools, high-dimensional, visualization, microarray data, microarray survey

## 1.0 Introduction

Advances in microarray technology not only initiated a change in the way biological experiments are performed and analyzed, but this new technology has 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 [6] [7]. These analyses are further complicated by the existence of correlative data such as patient medical records that need to be co-analyzed with the microarrays. Microarray technology allows for the simultaneous analysis of several genes or entire genomes at a time [8][9] making the resulting dataset inherently high-dimensional and complex [5]. 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.

The sheer size of microarray data generated by each microarray experiment [6] [7], along with collateral data 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.

In this paper we assess visualization features/tools desired by microarray users that are not available in current software tools, and the need for visualization tools that will not only allow researchers to expand upon current visualization techniques but also integrate microarray data with other biological data. The goal of this research effort was to assess existing visualization tools and to determine what - if any - unmet visualization needs exist. An exhaustive evaluation and/or assessment of existing microarray software - currently available as freeware, shareware, open source and commercial packages are well beyond the scope of this paper. The reader is directed to [10] and [11] for some partial software reviews. Towards achieving our research goal, we developed a microarray tool survey research instrument.

## 2.0 Methods

### 2.1 Survey Design

A survey research instrument entitled “Needs Assessment Survey for Scientific Visualization of Microarray Data” was administered from two sites: Virginia Commonwealth University (Site 1: IRB# 5065) and The University of Alabama at Birmingham (Site 2: IRB protocol X050822007). The survey research instrument was designed and distributed to a combined total of  $n = 972$  potential participants. Surveys were mailed to potential participants chosen from life and biomedical/health scientists, research institutes, biotech companies and other researchers who work with microarray data analysis.

Participation in the survey research study was completely voluntary and there was 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 25 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, 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 data, how long they had been using microarray analysis tools as well as how their current microarray analysis tools are used for image analysis [12][13], data mining [14], annotation [15] and/or for statistical analysis [16].

The Computing Environment Section asked participants to indicate which operating system(s) were currently used in their working environments.

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. Likert scales were used to measure interest and values ranged from 1 (most interested) to 4 (least interested). Participants were asked to indicate which microarray technologies, cDNA and/or oligonucleotide, were 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 tools, database tools, annotation tools, integrated packages and specific packages were used and of these tools and packages which ones were the primary tool(s) packages currently in use. Participants were asked to rank order on a Likert scale of 1 (most used) to 10 (least used), a list of common visualization tools in terms of their frequency of use. We hypothesized that there was an unarticulated or possibly even unrealized need for a visualization tool/tools or visualization function(s) that were not currently available in microarray analysis and visualization software. Participants were asked to indicate if they were 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 then asked to describe the desired tool(s)/function(s) that they felt would allow them to visualize datasets to their satisfaction. Considerable amounts of space were provided in order to allow participants to qualitatively 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.

### 2.2 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 currently represent the bulk of general microarray users and general microarray analyst’s working environment. From our review of the literature, 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” [17]. 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, as well as some computational fields like biostatistics, physics and computer science. A similar method was used to select survey participants from national laboratories. Ideally, we wanted to have an equal number of participants from each organization. However, because the source of participants was from 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.

A C++ computer program was written to read in potential 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 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. 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 the software was written to over-sample university and national laboratory organizations to make up the difference from the research institutes and biotech companies. The percentage of survey participants selected from each organization included research universities (40.4%), research institutions (8.4%), national laboratories (40.4%) and biotech companies (10.8%).

Snowball sampling was used to acquire additional potential participants. Survey recipients were asked to communicate, to the project researchers, names of other individuals who either perform microarray analysis or have individuals in their labs who do microarray analysis. This 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 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. As such, we provide no statistical analyses in this presentation. An alternative way to view this research survey is to recast it as a “paper” version, at a larger scale, of a focus group.

### 3.0 Results

A combined total of 972 surveys were mailed:  $n = 500$  from Site 1 and  $n = 472$  from Site 2. A total of 61 surveys (6%) were returned: 20 surveys were returned (32.8%) from Site 1 and 41 surveys (67.2%) were returned from Site 2. Returned surveys were categorized as completed, blank with explanation and blank (with no explanation). Completed surveys were surveys which respondents provided answers either by check marks or written responses. Some surveys were returned blank with a note attached indicating why the participant did not complete the survey. Comments from respondents regarding why they did not complete the survey will be summarized in the discussion section.

The highest level of education reported was Doctorate Degree (85.2%), Master’s Degree (4.9%) followed by Bachelor’s Degree (6.6%). 1.6% reported MD/PhD degrees. Respondent’s ages at last birthday ranged from 20 to 60 years. The largest group of respondents was in the 40-49 age group (28.1%) followed by an equal number of respondents in the 30-39 (26.3%) and 50-59 (26.3%) age groups. Most respondents were new to microarrays and microarray technology. Most (37.7%) reported 0 – 3 months time spent working with microarrays and 39.2% reported 0 –3 months times spent using microarray analysis tools. The primary job title reported was “Scientist/Investigator” with the largest number of respondents indicating “University” as their place of employment.

Respondents were asked to indicate the primary source of their current microarray software: commercial packages available for a fee, open source software available free or custom software written specifically for and by individual research establishments and/or research groups. 32.4% of the respondents indicated open source mediums as their primary source of microarray software. Commercial software packages were listed as the next likely source for microarray software. Combinations of the software sources are commonly used and represented 20.6% of the responses. Those reporting combinations of primary source of microarray software reported primarily using both open source and custom written software. Some respondents (2.9%) specifically stated “none” as their primary source of microarray software.

Results showed, among respondents, most microarray analysis tools are used for statistical analysis of microarray data (40.98%) followed by database analysis (31.15%), image analysis (24.59%) and annotation (22.95%).

While PC and MAC OS were the most commonly reported computing platform for microarray analysis, UNIX/LINUX environments were indicated as the preferred platform for those researchers who develop custom software.

To assess which microarray analysis level researchers were interested in, respondents were asked to rank on a Likert scale from 1 (most interested) to 4 (least interested) microarray analysis levels (probe level, expression level, cellular level and transcriptomic (mRNA)) according to their level of interest. Respondents were most interested in expression level analysis and least interested in probe level analysis.

Respondents were asked to indicate which of the microarray technologies, spotted cDNA microarrays or oligonucleotide microarrays, are in use in their working environment and which of these technologies is their primary technology in use. Oligonucleotide microarray technology was most prevalently used. Eighteen respondents indicated using spotted technology, 27 respondents indicated using oligonucleotide technology and one respondent indicated using some other type of microarray technology (spotted amplicons not from cDNA). There were 15 missing responses.

### **3.4 Microarray Analysis Tools**

Respondents were given a list of software applications written for microarray analysis from which they were asked to indicate which tool(s) were currently used in their environment for microarray analysis. The software applications were categorized as image analysis tools, database tools, annotation tools, integrated and specific microarray packages. Respondents were asked to check all the applications in use in their process of microarray analysis. The list of tools and packages provided in the survey was not an exhaustive list of current microarray software but a representative of a small sample of what was available when the survey was designed and administered. The lists of applications were obtained from reviewing the literature on microarray analysis and from several internet web sites that resulted from a web search for microarray software via the World Wide Web. For each category of applications respondents were given the option to write in the name of any tools currently in use in their environment but not listed in the survey. The categories and most frequently listed applications are listed below.

Image analysis tools included \*Affymetrix, Array Vision, GenePix Pro, Microarray Suite,\*R, ScanAlyze, Spot, and TIGR Spotfinder. Database tools included Acuity, ArrayDB, BASE, \*Custom Software, Genetrafic, MADAM and SMD. Annotation tools included DAVID, GoMiner, and RESOURCERER. Integrated packages included Bioconductor, BRB Array Tools, Cluster and TreeView, dChip, GeneSpring, \*Genetrafic, J-express Pro, MeV: MultiExperiment, and XpressionNTI. Specific packages included Arraystats, Clusfavor, FreeView, Geneclust, GeneClust2, MaanovaMIDAS, PAM, SAM, and sma. The asterisk (\*) indicates those applications that were not listed in the survey but were written in by survey respondents. From the responses TIGR Spotfinder was most frequently listed as the primary image analysis tool, MADAM was most frequently listed as the primary database tool, and DAVID was most frequently listed as the primary annotation tool. Responses for primary integrated and specific packages varied equally.

### **3.5 Microarray Visualization Tools**

Respondents were asked to rank order a list of graphical methods (visualization tools), commonly used for representing microarray data, in terms of their frequency of use. The Likert scale ranged from 1 (most used) to 10 (least used). Responses to this ranking varied. Some respondents ranked all of the tools while

some ranked a number of them and left others unranked. Table 1 shows the rankings of 15 visualization tools. Those rankings with 30 or more responses are included in Table 1. The most used tool was pathways (9 out of 69 most used responses). Respondents indicated parallel coordinates was the least frequently used tool. No one tool stood out as the primary visualization tool. The responses for primary visualization tool varied equally.

**Table 1.** Visualization tools in terms of frequency of use

	Frequency of Use				Total
	Most Used	Freq. Used	Reg. Used	Least Used	
Parallel Coordinates	1	1	1	6	9
Heat Maps	6	2	2	4	14
Scatter Plots	8	6	3	1	18
Histograms	7	3	4	1	15
Bar Charts	3	3	1	3	10
Line Charts	1	3	3	3	10
Pie Charts	1	2	0	5	8
Block Views	0	0	2	4	6
Array Layouts	3	2	2	5	12
Physical Position on Genomes	4	2	1	3	10
Pathways	9	1	1	0	11
Ontologies	3	0	3	4	10
Spreadsheets	8	1	2	3	14
Gene-to-Gene Comparison	8	3	1	2	14
Cluster Dendogram	7	4	4	2	17
Total	69	33	30	46	178

Freq. = Frequently      Reg. = Regularly ( $\lambda=0.84$ )

When asked if able to visualize multivariate, high-dimensional data sets to the respondent's satisfaction using the software and visualization tools currently available twenty-three out of 30 respondents (76.7%) said they were not satisfied with their current software's ability to visualize such data. Seven out of 30 respondents (23.3%) indicated they are satisfied with their current visualization software (there were 31 missing responses).

Respondents who indicated they were not satisfied with their current visualization tools were asked to describe the desired tool(s)/function(s) that they felt would allow them to visualize multivariate, high-dimensional data to their satisfaction. Respondents described a tool with features that has 3D Clustering and visualization capabilities, a pathway level genome browser which incorporates intermediate data that shows many types of data/information (*i.e.*, co-expression, differential expression). Respondents described a tool that integrates expression profiles with protein-protein interaction networks, shows the intersection/union of different classes of expressions, allows rapid exploration of relational database contents, and relates microarray datasets to other large functional genomic datasets. Respondents indicated that they want a tool that allows users to explore relationships between expressed genes and metabolic pathways, networks of genes and interacting proteins.

Respondents were asked what visualization features are missing from their current microarray software packages that they would prefer and/or would like to have. Respondents indicated they use several tools to link to other biological data and would like tools that easily allows for such linking. High-throughput visualization was listed as a missing feature. Respondents would like to visualize a large number of arrays in one project/dataset and determine which array(s) are outliers or of poor quality as well as have the ability to do cross dataset comparisons (*i.e.*, gene expression and pathway analysis) together with some confidence measure that those pathways are really involved.

## 4.0 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. The survey was distributed to a cross-section of microarray users representing researchers from universities, national laboratories, biotech companies and research institutions. Although survey participants were specifically targeted based on the likelihood that microarray analysis is done in their area of study/research we anticipated a number of surveys would be (1) not returned at all, (2) returned blank or (3) returned blank with a note attached explaining why they chose not to complete the survey. Among those surveys returned blank with a note attached a large number of notes indicated the respondent did not work with microarrays or do microarray analysis at the moment but did anticipate doing so in the future. There were no indications as to when their work with microarrays would begin.

Based upon these results, we conclude that most researchers are currently relying on open source software and are choosing to write their own software to perform analysis and visualizations specifically tailored to the problems they are trying to solve. While there are users running microarray software on LINUX and SOLARIS platforms, most users either purchase and/or use custom written software for PC and MAC OS computing environments. The results highlight an apparent disconnect between software users and software developers. Respondents who indicated using custom written scripts for their visualization and analysis needs also indicated using LINUX/SOLARIS as their computing environments while respondents who indicated using commercial software packages as their source for microarray software indicated PC and MAC OS as their preferred computing environment.

One of the expected results of the survey was to receive specific and detailed descriptions of the visualization features users would like to have that they don't have with currently used software. Several questions in the survey were designed to give the respondents an opportunity to describe these software features. Respondents were asked if they were satisfied with the visualization features of their current software and if there were any visualization features missing from their current software packages. Respondents were asked to write in their replies to these questions with as much detail as they desired. While the responses received to these questions were descriptive, there were some that were general and vague. There were a noticeable number of responses where the respondents indicated they were not satisfied with the current software but provided no description of the desired tool that would allow them to display and represent their data to their satisfaction. A number of respondents indicated the tools and functions they desire are not available to them in their current software but also indicated they were not sure or not aware if this desired feature was available in any other application or software package. There are so many tools available via commercial and open source mediums it is impossible to be aware of and know with any level of detail what they all have to offer.

As with any tool, respondents want a tool that is affordable, easy to use and understand that accurately reflect true biological changes. Respondents indicated a need for improvement in current tools and features like improved handling of cross hybridization, improved ability to data mine probe readings across hundreds of data sets, improved quality control, and improved visualization of pathways. Once the analysis is done and the data biologically verified and visualized, users would like a tool that allows for easy manipulation of generated images and displays and ease of import/export of this data from one application /tool to another. As articulated by one survey respondent: "Every microarray experiment is different; it is difficult to design a program to be flexible enough to cover all possibilities.

It is our hope that the responses reported in this paper will be insightful and provide an indicator of what the microarray user base needs in order to address their data visualization and analysis needs. For those researchers who are new to microarray analysis we hope the survey provided a glimpse of the different and vast number of software tools available for the different levels and interest in microarray analysis

## 5.0 References

[1] Thomas D. Wu. "Analyzing gene expression data from DNA microarrays to identify candidate genes." *J Pathol* 195:53-65, 2001.

[2] H. Ge, A. Walhout and M. Vidal. "Integrating 'omic' information: a bridge between genomics and systems biology." *Trends in Genetics* 19:551-560, 2003.

[3] Patrick O. Brown, and David Botstein. "Exploring the new world of the genome with DNA microarrays." *Nature Genetics* 21 (Suppl.):33-37, Jan 1999.

- [4] David J. Duggan, Michael Brittner, Yidong Chen, Paul Meltzer and Jeffery M. Trent. "Expression profiling using cDNA microarrays." *Nature Genetics* 21:10-14, Jan 1999.
- [5] Izet M. Kapetanovic, Simon Rosenfeld and Grant Izmirlian. "Overview of commonly used bioinformatics methods and their applications." *Ann. N.Y. Acad. Sci.* 1020:10-21, 2004.
- [6] Mike Eisen. "Making biological sense of genome-wide expression data." *Nature Genetics* 23:18-18, 1999.
- [7] Daniel R. Masys. "Linking microarray data to the literature." *Nature Genetics* 28:9-10, May 2001.
- [8] Roger Ekins and Frederick Chu. "Microarrays: their origins and applications." *Trends Biotechnol.* 17:217-218, Jun 1999.
- [9] Danh V. Nguyen, A. Bulak Arpat, Naisyn Wang and Raymond J. Carroll. "DNA microarray experiments: biological and technological aspects." *Biometrics* 58:701-717, Dec 2002.
- [10] D. K. Liu, B. Yao, B. Fayz, D. D. Womble and S. A. Krawetz. "Comparative evaluation of microarray analysis software." *Molecular Biotechnology* 26:225-232, 2004.
- [11] Purvi Saraiya, Chris North and Karen Duca. "An evaluation of microarray visualization tools for biological insight." *IEEE Symposium on Information Visualization*.
- [12] Zhongming Chen and Lin Liu. "RealSpot: Software validating results from DNA microarray data analysis with spot images." *Physiol. Genomics* 21:284-291, Feb 2005.
- [13] Yee Hwa Yang, Michael J. Buckley and Terence P. Speed. "Analysis of cDNA microarray images." *Briefings in Bioinformatics* 2(4):3412-349, Dec 2001.
- [14] M. Gardiner-Garden and T. G. Littlejohn. "A comparison of microarray databases." *Briefings in Bioinformatics* 2(2):143-158, May 2001.
- [15] Purvesh Khatri and Sorin Draghici. "Ontological analysis of gene expression data: current tools, limitations, and open problems." *Bioinformatics* 21(18):3587-3595, Jun 2005.
- [16] Wei Pan. "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments." *Bioinformatics* 18(4):546-554, Dec 2001.
- [17] Ahmed Fadiel and Frederick Naftolin. "Microarray applications and challenges: a vast array of possibilities." *International Archives of Bioscience* 1111-1121, 2003.

## 6.0 ACKNOWLEDGEMENTS

We would like to thank all of the respondents of the survey without whom this research effort would have been for naught. We would like to thank Dr. 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 at the Virginia Commonwealth University, Center for the Study of Biological Complexity. We would also like to thank the postal groups at both VCU and UAB for their assistance. A heartfelt thank you is extended to Dr. Grier Page without whom continuation of the research instrument at Site 2 would not have been possible. We would like to also thank The UAB Department of Computer and Information Sciences for their assistance and resources.