31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

SimpleVisGrid: Grid Services for Visualization of Diverse Biomedical Knowledge and Molecular Systems Data

Todd H. Stokes, *Member, IEEE* and May D. Wang, *Member, IEEE*

Abstract—**Biomedical data visualization is a great challenge due to the scale, complexity, and diversity of systems, system component interactions and experimental data. Standards for interoperable data are a good start to addressing these problems, but standardization of visualization technologies is an emerging topic. SimpleVisGrid builds on Cancer Biomedical Informatics Grid (caBIG) common infrastructure for cancer research, and clearly specifies and extends three standard data formats for inputs and outputs to grid services: comma-separated values (CSV), Portable Network Graphics (PNG), and Scalable Vector Graphics (SVG). Four prototype visualizations are available: 2D array data quality visualization, correlation heatmaps between high-dimensional data and associated meta-data, feature landscapes, and biochemical or semantic network graphs. The services and data model are prepared for submission for caBIG Silver-level compatibility review and for integration into automated research workflows. Making these tools available to caBIG developers and ultimately to biomedical researchers can (1) help with biomedical communication, discovery, and decision-making, (2) encourage more research on standardization of visualization formats, and (3) improve the efficiency of large data transfers across the grid.**

I. Introduction

Visualization of biomedical data encompasses a diverse set of domain specialties, computational platforms, algorithms, and proposed invariant data representations (i.e. symbols or glyphs) [1]. Data visualization enables the discovery of underlying patterns that may indicate key scientific findings or quality problems in the data acquisition step that should be resolved before starting analysis. Additionally, many visualizations enhance clinical decision-making [2], and may be a critical accompaniment to the successful translation of new molecular data acquisition techniques to the clinic and the advent of personalized medicine. In this paper, we present a biomedical visualization classification system that leads us to propose three flexible data format standards to improve data portability and interoperability for bioinformatics and systems biology.

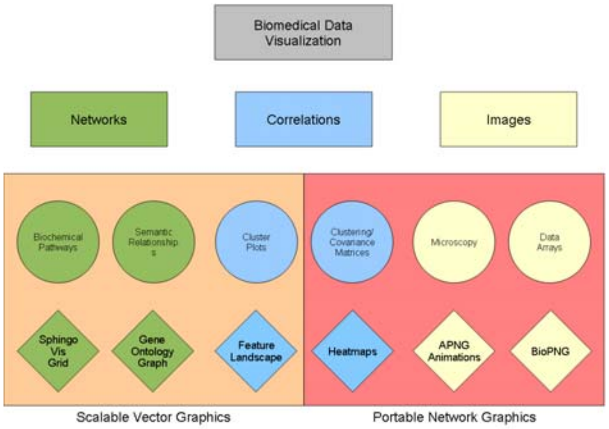


Fig. 1. Organization of biomedical research visualizations into classes. The rectangles indicate the class, circles represent visual representations, and diamonds represent concrete examples of the representations. The choice of whether a visualization is best represented by SVG or PNG is primarily one of data density.

A. Popular Visualization Systems

Visual Statistical Data Analyzer (VISDA) [3] is the first visualization tool to become Cancer Biomedical Informatics Grid (caBIG) certified. Cytoscape [4, 5] is a general network visualization tool that has deservingly received a lot of attention in this field. Cytoscape is an open-source standalone installation and makes some APIs available for other developers, but does not support web-based or grid-based requests. Haploview [6] is useful for comparisons of entire genomes to one another to look for small differences. GeneWindow [7] is an interactive Scalable Vector Graphics (SVG) interface that enables the user to browse gene sequences with a variety of annotation overlays. Matrix2PNG [8] is a simple and useful tool for quickly converting data stored in a matrix into a heatmap. There are some similarities between Matrix2PNG and the BioPNG system presented here. The primary difference is that BioPNG can be used for data transport. It always treats the encoding of data into a PNG as a potential two-way interaction, sacrificing some of the visual appeal of the graphic for the ability to retrieve the original data using only the BioPNG file.

Manuscript received April 7, 2009. This research has been supported by Emory-Georgia Tech NCI Center for Cancer Nanotechnology Excellence (U54CA119338), NCI Bioengineering Research Partnership (R01CA108468), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), National Science Foundation (Graduate Student Research Fellowship to RM), Microsoft Research, and Hewlett-Packard.

T. H. Stokes is with the Electrical and Computer Engineering Department of Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: todd.stokes@bme.gatech.edu).

M. D. Wang is with the Biomedical Engineering Department of Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA. (corresponding author, phone: 404-385-2954; fax: 404-894-4243; e-mail: maywang@bme.gatech.edu).

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

B. *Visualization Classification System*

Attempts have been made to organize visualization efforts into classes of functional problems and structural solutions, including: graphs, tables, maps, diagrams, networks, and icons [9] and matrices, networks and hierarchies [10]. We simplify this view for biomedical research by focusing on networks, correlations, and matrices (including general images and plots) (see Figure 1). This classification system is reduced to visualizations of “non-tangible” data (i.e. data that does not have a clear physical representation), and excludes a whole body of work focused on accurate representation of three-dimensional (3D) objects such as anatomical visualization and molecular conformation visualization. In general, the work of SimpleVisGrid has been to focus on the enormous field of two-dimensional (2D) data representation for interpreting data analysis results, with the minor exceptions of discussing the application of BioPNG to “data cubes” and the use of two-and-a-half-dimensional (2.5D) representation used to view feature landscapes.

C. *Data Scale Impedes Interpretation and Transport*

One reason that so many visualization systems for bioinformatics and systems biology are designed as standalone tools or designed to connect only to local databases is the scale of the data. The idea of plugging visualizations of such large-scale data as 30,000 genes on a microarray, 100,000 proteins in an interaction network, or billions of base pairs on a genome into a grid-based workflow for multiple-sample comparison has been technically infeasible up to this point. One obstacle is that interoperability standards such as MAGE-ML [11] or BioPAX [12], while extremely useful for passing meta-data about experiments around, cannot be extended to raw experimental results because the uncompressed text and required labels are too bulky. A balance must be found between appropriate meta-data to transport in structured formats versus bulk data to transport in standard compressed formats. A recent move in this direction is MAGE-TAB [13], which emphasized the facility of moving data between human-readable spreadsheets and machine-readable data files as a major usability factor for microarray analysis tools.

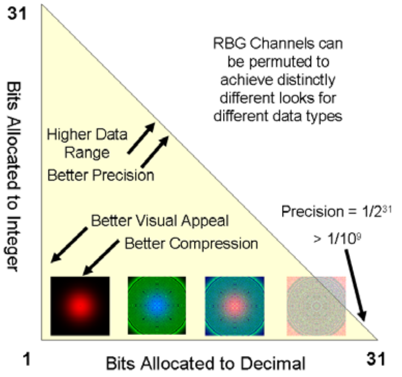

31
RBG Channels can be permuted to achieve distinctly different looks for different data types
Higher Data Range
Better Precision
Bits Allocated to Integer
Better Visual Appeal
Better Compression
Precision = 1/2³¹
> 1/10⁹
1
Bits Allocated to Decimal
31

Fig. 2. Trade-offs between different BioPNG formats. Encoding a value simply requires converting to a binary decimal and splitting the resulting string into bit lengths to fit into the color depth required to store the data. The images along the bottom are all Gaussian surfaces stored at various decimal precision. PHP Source code for BioPNG compression and extraction can be found at the url: http://arraywiki.bme.gatech.edu/index.php/BioPNG_Source_Code

II. Identification of Standard Formats

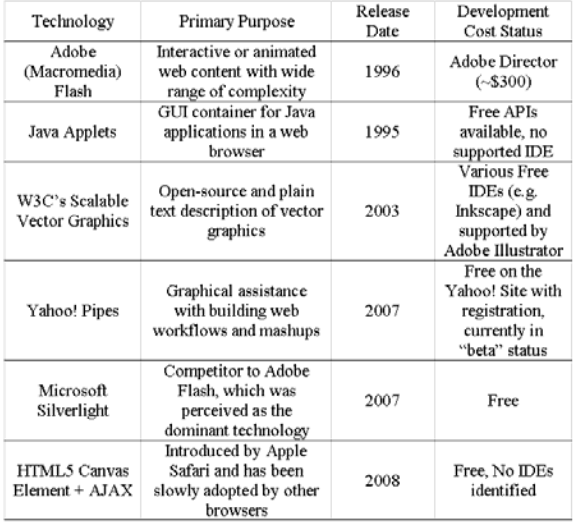
A. *Bio-Portable Network Graphics (BioPNG)*

Portable Network Graphics (PNG) is an open specification for image compression. Since images are simply representations of 2D data matrices, it is somewhat intuitive to think of image formats as candidates for storing data in this form. However, we found no examples of systems making use of image compression to store non-image data. There are two reasons for this: (1) general compression formats such as ZIP have been the obvious choice for compressing general data and (2) image formats tend to have limited bit depth and/or lossy compression algorithms that meant the data you got out might not match the data you put in. In presenting a new community-based microarray repository, ArrayWiki [14], we introduced BioPNG, a method for converting floating point numbers into color channels for storage in a lossless format. This format has many advantages over ZIP: (1) it is natively supported over HTTP, (2) it is presented easily in browsers so the data can be “seen” as it’s received, (3) the compression rate compares favorably to ZIP, often performing 20-30% better, (4) PNG files do not harbor viruses like ZIPs can, (5) meta-data such as scale (x and y dimensions) are easily extracted without parsing, and (6) PNG files can be split, cropped, and joined efficiently, allowing for possibilities of pseudo-random access for large datasets. Figure 2 illustrates the trade-offs of different BioPNG formats.

B. *Scalable Vector Graphics (SVG)*

SVG as a web standard was approved by the World Wide Web Consortium (W3C) in 2003. SVG documents describe 2D graphics in an efficient way and can preserve the underlying scientific data. SVG documents can be annotated with custom tags to store the source data used to generate the visualization. All SVG documents are zoomable (multi-scale) interfaces. This means that the resolution of the display device (or the available screen space in an integrated application) does not affect the readability of the scaled representation of the graphic. Finally, SVG documents can be programmatically manipulated using the Javascript Document Object Model (DOM), providing a means to implement animations and innovative interactive interfaces.

SVG-based bioinformatics tools are becoming more common. Some examples are GeneWindow [7], ArrayXPath [15], the microbial genome viewer [16], caCORRECT [17] and GOMiner [18]. SVG could be the basis for multimodal scientific and educational material in the future [19]. SVG rendering is provided in Microsoft Internet Explorer 7 by means of an unsupported Adobe SVG Plugin but is natively supported in the latest releases of the Mozilla Firefox, Apple Safari, Google Chrome and Opera browsers. SVG is compatible with Asynchronous JavaScript and XML (AJAX) technology. AJAX is a technique for manipulating an HTML-based user interface using background browser processes without causing the noticeable page refresh of classic HTML forms. AJAX has the advantage of being similar in look and feel to locally installed software. However, AJAX without SVG must rely on reloading of server-generated images to improve the interactivity of visualizations.



Technology	Primary Purpose	Release Date	Development Cost Status
Adobe (Macromedia) Flash	Interactive or animated web content with wide range of complexity	1996	Adobe Director (~$300)
Java Applets	GUI container for Java applications in a web browser	1995	Free APIs available, no supported IDE
W3C's Scalable Vector Graphics	Open-source and plain text description of vector graphics	2003	Various Free IDEs (e.g. Inkscape) and supported by Adobe Illustrator
Yahoo! Pipes	Graphical assistance with building web workflows and mashups	2007	Free on the Yahoo! Site with registration, currently in "beta" status
Microsoft Silverlight	Competitor to Adobe Flash, which was perceived as the dominant technology	2007	Free
HTML5 Canvas Element + AJAX	Introduced by Apple Safari and has been slowly adopted by other browsers	2008	Free, No IDEs identified

Fig. 3. Web-based visualization technologies. These have many trade-offs between openness, performance, and capabilities. The choice of SVG as a standard for SimpleVisGrid was made because of unparalleled openness, flexibility for data storage, scripting, and availability on mobile platforms like the iPhone.

III. Implementation of Novel Visualizations

A. *Feature Landscapes*

Feature landscapes are an example of using the SVG format to depict 2.5D visualizations (see Figure 4). Although the peaks and the background appear to have depth, they are positioned on the graphic using pre-built symbols. Each peak symbol contains custom attributes providing an ID for the feature and the exact number of times it appears in the provided data, as this information could never be recovered using only the X,Y location of the peak in the graphics file. We have used feature landscapes in our research to see if tuning certain parameters in feature selection algorithms significantly affects the results. We are developing an algorithm for summarizing the similarity between the feature lists to be published in a separate paper.

B. *Meta-data Correlations*

High-dimensional, high-throughput data acquisition methods have a tendency to suffer from a phenomenon called "batch effect." This is an overall bias applied to the data based on the period of time that the data was acquired. Batch effects may be caused by environmental factors, human error, changes in equipment or changes in lab protocols. It is very important to correct these problems before data analysis models are built, because the bias can become the dominant feature of the data. Our meta-data correlation visualization can be used to detect batch effect, but can also be used to evaluate experimental design problems, such as sudden changes in chip quality, inappropriate randomization procedures, and relationships between clinical factors that might be useful to exploit when analyzing data.

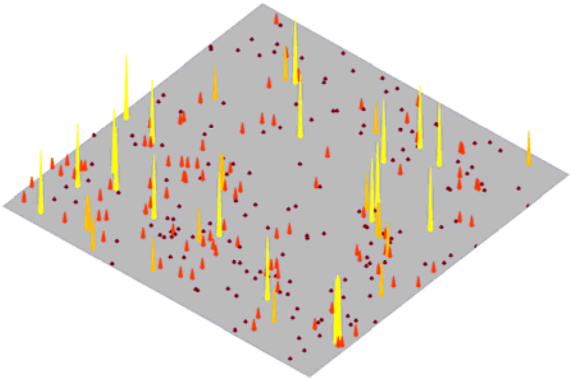


Fig. 4. Example of a feature landscape. This data represents results of using simple feature selection methods on a public microarray dataset. The rendering data format is SVG. Tall, bright peaks were selected often by a variety of methods and short dark peaks were selected by few methods. Grey regions represent features never selected. Feature landscapes may be merged to create other useful comparisons, such as comparing one dataset to another dataset. This landscape contains ~22,000 features total.

C. *Biochemical Pathway Simulation Results*

Network layout problems frequently arise in biomedical visualization. While the problems can be very large, requiring a great deal of computational time to render, many are small enough to be rendered using simple algorithms. A good package for providing flexible and easy-to-learn algorithms is GraphViz [20]. Our network layout visualization service is a wrapper for GraphViz and provides higher-quality SVG documents than those obtained directly from GraphViz, including animated buttons to navigate among time series data and bezel-curved edges. We support many of the standard features of the GraphViz package in addition to automatic generation of a PNG file and embedding the original source data into the SVG document that is returned. In addition to biochemical reactions, the network layout service supports the layout of directed acyclic graphs (DAGs) like the tree structure used to represent relationships between terms in the Gene Ontology.

IV. System Design For caBIG Certification

SimpleVisGrid services are based on a UML Model constructed in Enterprise Architect. Common Data Elements (CDEs) have been identified and semantically annotated using the Semantic Integration Workbench (SIW) provided by caBIG. Where necessary, new concepts have been identified and defined for insertion into the Enterprise Vocabulary Services (EVS) ontology for cancer research. These practices mean that we can build a submission package for Silver-level certification of SimpleVisGrid.

V. Conclusion

We present SimpleVisGrid, a collection of grid services for converting bioinformatics and systems biology data into functional visualizations that reveal important patterns relevant to the quality and comparability of biological data. The data standards used as inputs and outputs to these services improve data portability and interoperability by providing compression and source data in the visual representation.

Acknowledgment

T. H. Stokes thanks Dr. Al Merrill, Dr. Mitch Parry, Richard Moffitt, and Dr. John Phan for valuable discussions concerning the proper use and design of the visualizations presented in this paper. In addition, we thank Baris Suzek, Salvatore Mungal, and Olga Tchuvatkina for providing mentoring on compiling a successful compatibility review package for previous grid services.

References

[1] H. Kitano, A. Funahashi, Y. Matsuoka, and K. Oda, "Using process diagrams for the graphical representation of biological networks," *Nat Biotechnol,* vol. 23, pp. 961-6, Aug 2005.

[2] S. M. Powsner and E. R. Tufte, "Graphical Summary of Patient Status," *Lancet,* vol. 344, pp. 386-389, Aug 6 1994.

[3] J. J. Wang, H. Li, Y. T. Zhu, M. Yousef, M. Nebozhyn, M. Showe, L. Showe, J. H. Xuan, R. Clarke, and Y. Wang, "VISDA: an open-source caBIG (TM) analytical tool for data clustering and beyond," *Bioinformatics,* vol. 23, pp. 2024-2027, Aug 1 2007.

[4] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Research,* vol. 13, pp. 2498-2504, Nov 2003.

[5] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader, "Integration of biological networks and gene expression data using Cytoscape," *Nat Protoc,* vol. 2, pp. 2366-82, 2007.

[6] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics,* vol. 21, pp. 263-265, Jan 15 2005.

[7] B. Staats, L. Qi, M. Beerman, H. Sicotte, L. A. Burdett, B. Packer, S. J. Chanock, and M. Yeager, "Genewindow: an interactive tool for visualization of genomic variation," *Nat Genet,* vol. 37, pp. 109-10, Feb 2005.

[8] P. Pavlidis and W. S. Noble, "Matrix2png: a utility for visualizing matrix data," *Bioinformatics,* vol. 19, pp. 295-296, Jan 22 2003.

[9] G. L. Lohse, K. Biolsi, N. Walker, and H. H. Rueter, "A Classification of Visual Representations," *Communications of the Acm,* vol. 37, pp. 36-49, Dec 1994.

[10] L. R. Novick and S. M. Hurley, "To matrix, network, or hierarchy: That is the question," *Cognitive Psychology,* vol. 42, pp. 158-216, Mar 2001.

[11] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C. J. Stoeckert, Jr., and A. Brazma, "Design and implementation of microarray gene expression markup language (MAGE-ML)," *Genome Biol,* vol. 3, p. RESEARCH0046, Aug 23 2002.

[12] J. S. Luciano, "PAX of mind for pathway researchers," *Drug Discovery Today,* vol. 10, pp. 937-942, Jul 1 2005.

[13] T. F. Rayner, P. Rocca-Serra, P. T. Spellman, H. C. Causton, A. Farne, E. Holloway, R. A. Irizarry, J. M. Liu, D. S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, C. J. Stoeckert, J. White, P. L. Whetzel, F. Wymore, H. Parkinson, U. Sarkans, C. A. Ball, and A. Brazma, "A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB," *BMC Bioinformatics,* vol. 7, pp. -, Nov 6 2006.

[14] T. H. Stokes, J. T. Torrance, H. Li, and M. D. Wang, "ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses," *BMC Bioinformatics,* vol. 9 Suppl 6, p. S18, 2008.

[15] H. J. Chung, M. Kim, C. H. Park, J. Kim, and J. H. Kim, "ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics," *Nucleic Acids Research,* vol. 32, pp. W460-W464, Jul 1 2004.

[16] R. Kerkhoven, F. H. J. van Enckevort, J. Boekhorst, D. Molenaar, and R. J. Siezen, "Visualization for genomics: The microbial genome viewer," *Bioinformatics,* vol. 20, pp. 1812-1814, Jul 22 2004.

[17] T. H. Stokes, R. A. Moffitt, J. H. Phan, and M. D. Wang, "chip artifact CORRECTion (caCORRECT): A Bioinformatics System for Quality Assurance of Genomics and Proteomics Array Data," *Ann Biomed Eng,* vol. 35, pp. 1068-80, Jun 2007.

[18] B. R. Zeeberg, W. M. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein, "GoMiner: a resource for biological interpretation of genomic and proteomic data," *Genome Biology,* vol. 4, p. R28, 2003.

[19] R. H. Landau, D. Vediner, P. Wattanakasiwich, and K. R. Kyle, "Future scientific digital documents with MathML, XML, and SVG," *Computing in Science & Engineering,* vol. 4, pp. 77-85, Mar-Apr 2002.

[20] E. R. Gansner, E. Koutsofios, S. C. North, and K. P. Vo, "A Technique for Drawing Directed-Graphs," *Ieee Transactions on Software Engineering,* vol. 19, pp. 214-230, Mar 1993.