30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

An interactive visualization tool and data model for experimental design in systems biology

Shray Kapoor, Chang Feng Quo, Alfred H. Merrill, Jr., and May D. Wang

Abstract—**Experimental design is important, but is often under-supported, in systems biology research. To improve experimental design, we extend the visualization of complex sphingolipid pathways to study biosynthetic origin in SphinGOMAP. We use the ganglio-series sphingolipid dataset as a test bed and the Java Universal Network / Graph Framework (JUNG) visualization toolkit. The result is an interactive visualization tool and data model for experimental design in lipid systems biology research. We improve the current SphinGOMAP in terms of interactive visualization by allowing (i) choice of four different network layouts, (ii) dynamic addition / deletion of on-screen molecules and (iii) mouse-over to reveal detailed molecule data. Future work will focus on integrating various lipid-relevant data systematically *i.e.* SphinGOMAP biosynthetic data, Lipid Bank molecular data (Japan) and Lipid MAPS metabolic pathway data (USA). We aim to build a comprehensive and interactive communication platform to improve experimental design for scientists globally in high-throughput lipid systems biology research.**

Keywords—Experimental design, graph layout algorithms, interactive visualization, sphingolipid pathways

I. Introduction

Experimental design is an important, but often under-supported aspect in systems biology research. On one hand, high-throughput technologies such as DNA microarrays and mass spectrometry are continually advanced to improve the collection of high-resolution quantitative and temporal data. On the other hand, computational science is evolving in tandem to produce a diverse spectrum of bioinformatics tools and algorithms to analyze increasing volumes of data post-experiment. In this paradigm, there are limited resources to aid the design of large-scale experiments that involve systematically collecting heterogeneous data from complex biological systems over sustained periods.

Thorough and informed experimental design is critical for successful systems biology research. Because of technological and analytical advancements, there is an increasing momentum shift from studying biological systems in relative isolation to high-throughput systems biology research. Where investigators could adequately articulate and design experiments by hand in the paradigm of single-variable experiments, they now require computational tools to visualize and handle multi-variable experiments. Consequently, a measure of good experimental design may be how well it reduces computational load during data analysis post-experiment. In other words, the computational load for a healthy and complete research process is balanced throughout experimental design, data acquisition and post-experiment analysis.

Such experimental design can be achieved by increasing the quantity and quality of *a priori* knowledge available. With increasing efforts for community annotation and sharing of scientific data, it is certain that researchers have no lack of *a priori* data if they know where to look. On the other hand, the quality of such knowledge is not guaranteed *i.e.* researchers may not know how to look. For the purpose of ensuring both quantity and quality, multiple standards have been proposed for various biological data [1-4].

To improve the quality of *a priori* knowledge, we propose an interactive visualization data model for experimental design in systems biology, using the ganglio-series of sphingolipids dataset as a test bed. We extend and improve the current SphinGOMAP [5] in terms of interactive visualization by implementing (i) dynamic network visualization, (ii) choice of four different network layouts and (iii) mouse-over to reveal detailed molecule data. These features enhance user experience in dealing with high-volume, large-scale systems biology data.

II. Methods

We design and implement our interactive visualization tool based on molecular data of sphingolipids from SphinGOMAP [5] and programmatic visualization library from JUNG [6]. Furthermore, we extend our tool to include

Manuscript received April 16, 2008. This work is supported by grants from The Parker H. Petit Institute for Bioengineering and Bioscience (IBB), Johnson & Johnson, Bio Imaging Mass Spectrometry Initiative at Georgia Tech, National Institutes of Health (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MW), and Microsoft Research.

S. Kapoor is with the College of Computing, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: skapoor6@mail.gatech.edu).

C. F. Quo is with the Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: gtg801f@mail.gatech.edu).

A. H. Merrill, Jr. is with the School of Biology, Georgia Institute of Technology, Atlanta, GA 30332 USA (email: al.merrill@biology.gatech.edu).

M. D. Wang (corresponding author) is with Wallace H. Coulter Department of Biomedical Engineering; the School of Electrical and Computer Engineering; the Parker H. Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA 30332 USA and Hematology and Oncology, Winship Cancer Institute, Emory University, Atlanta GA 30322 USA (phone: (404)385-2954; e-mail: maywang@bme.gatech.edu).

978-1-4244-1815-2/08/$25.00 ©2008 IEEE.

a data model for potential interactions with existing lipid resources such as LipidBank [7] and Lipid MAPS [8].

A. *SphinGOMAP* [5]

SphinGOMAP is a pathway map to organize and visualize sphingolipids. The expressed objective of SphinGOMAP is to "promote dialog about the 'knowns' and 'unknowns' of sphingolipid biosynthesis and lead to experiments to refine this model"[5]. Thus, SphinGOMAP is an active document that evolves with emerging scientific discovery.

The current SphinGOMAP can be improved in terms of interaction and visualization. First, from the SphinGOMAP website, release 2.0 in October 2007 displays ~450 compounds in a static file as a Microsoft PowerPoint slide or JPEG image. In its present form, the static images do not allow users to interact freely with the map. Second, furthermore, the various families (series) of sphingolipids have to be presented separately in different files. This is because the density and scale of sphingolipid networks is too large to allow a panoramic view and yet provide sufficient detail at the same time. Third, the data in SphinGOMAP contains only molecular structure, category, common name and LipidBank ID. Thus, the connectivity of the pathway map is compromised for clarity and some detail.

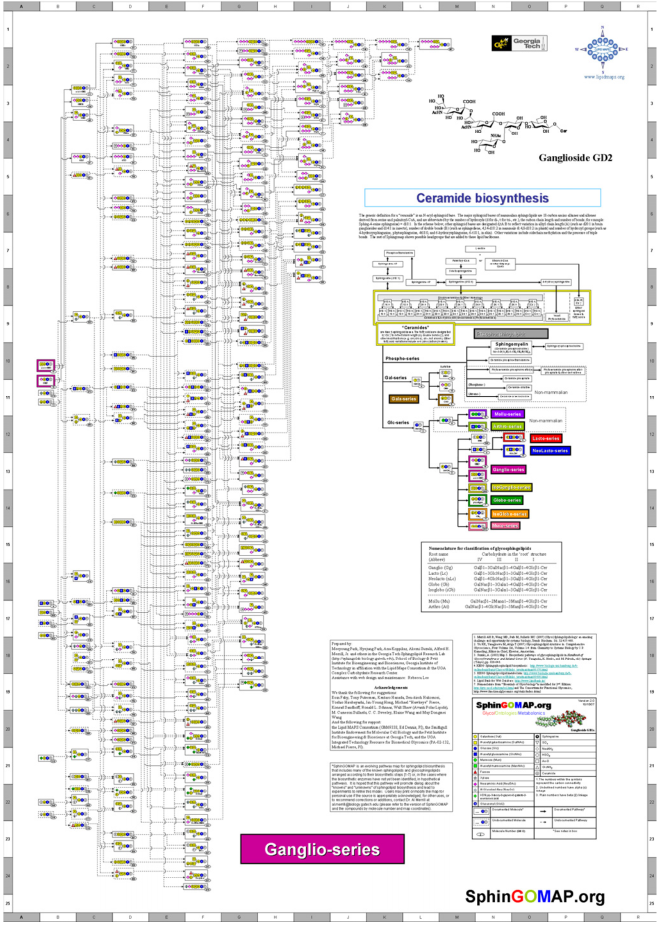


Fig. 1. Current visualization of the ganglio-series of sphingolipids in SphinGOMAP [5]. This is available as a Microsoft PowerPoint slide or static JPEG image file that allows limited interactivity and compromises map connectivity for clarity and some detail.

On the other hand, a simple and intuitive legend to denote biochemical functional groups for sphingolipids and its derivatives is deployed in SphinGOMAP. The legend helps users recognize distinct structural moieties at a glance. This is especially helpful to determine recurring patterns in biosynthetic inheritance.

We address these problems with interaction, data separation, and limited detail using a data model and software tool. We focus on the ganglio-series of sphingolipids from SphinGOMAP as our test data. This software tool is implemented using an open-source visualization toolkit.

B. *JUNG - Java Universal Network/Graph Framework* [6]

The JUNG API for Network/Graph visualization provides common and extensible software library for analysis and visualization of data that can be represented as a graph or network. In this work, we integrate the Kamada-Kawai [9], Fruchterman-Reingold [10], ISOMLayout [11-12] and CircleLayout algorithms provided by the JUNG software library. We use the JAVA graphics API to render the structure of the molecules.

The vertex of molecules is generated based on the internal chain structures. These chains are generated *a priori* during database filtering. In database filtering, we extract the molecular structure as formulas and reorder them based on the cardinality of molecules in a chain. Starting from a shorter chain, the derivatives are generated and mapped to its parent.

An in-memory tree is built in two steps to generate this type of hierarchy. In the first step, we select single-chain molecules to generate the basic hierarchy, while in the second step, all molecules that have branches are broken into different chains. For instance, a molecule “Galb 1-3GalNAcb 1-4(NeuAca2-3)Galb 1-4GlcCer” (b – beta), is broken down into two chains, Galb 1-3GalNAcb 1-4Galb 1-4GlcCer and NeuAca 2-3Galb 1-4GlcCer. We search for these two chains throughout the basic hierarchy built in the first phase to extract the parent chain. Note that every branched molecule will have more than one parent.

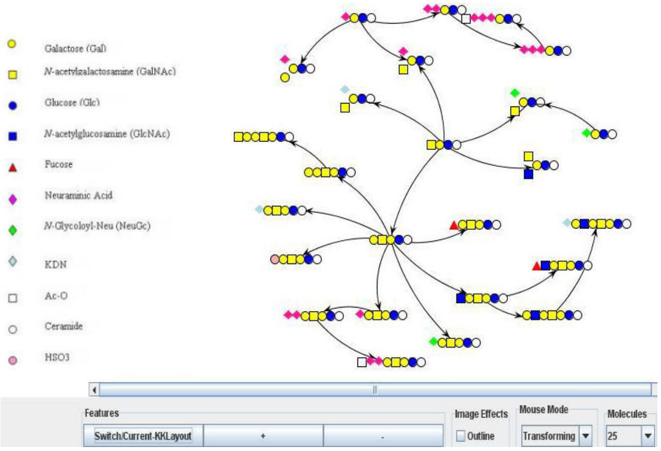
Once this relationship is built, it is sent as an input (Hashtable structure) to the visualization module. The visualization module extracts the molecular structure and renders it as an image on a label, which acts as a vertex for the graph structure.

C. *LipidBank* [7] *and Lipid MAPS* [8]

LipidBank and Lipid MAPS are global leading sources for lipid data that originated from Japan and USA respectively. On one hand, LipidBank contains primary data for an extensive number of lipids (~6000 molecules) in terms of molecular structure, scientific and common names, spectral information and literature. On the other hand, Lipid MAPS is focused on more secondary data in terms of lipid interactions within mammalian cells by "characterizing the global changes in lipid metabolites ('lipidomics')"[7]. Lipid MAPS contains not only structural data and annotations of biologically active lipid molecules, information about lipid metabolic pathways, experimental protocols, standards and time-course results are also available. Lipid MAPS is also linked to public databases for relevant molecules such as lipid-associated proteins.

III. Results

An overview of the visualization is presented in Figure 2. In this section, we report noteworthy features such as (a) dynamic network visualization with increasing complexity, (b) choice of different network layouts and (c) mouse-over to reveal detailed molecule data. The intuitive legend from SphinGOMAP for describing biochemical functional groups on the sphingolipid molecules is preserved in our visualization.



Mouse Mode: Transforming

Molecules: 25

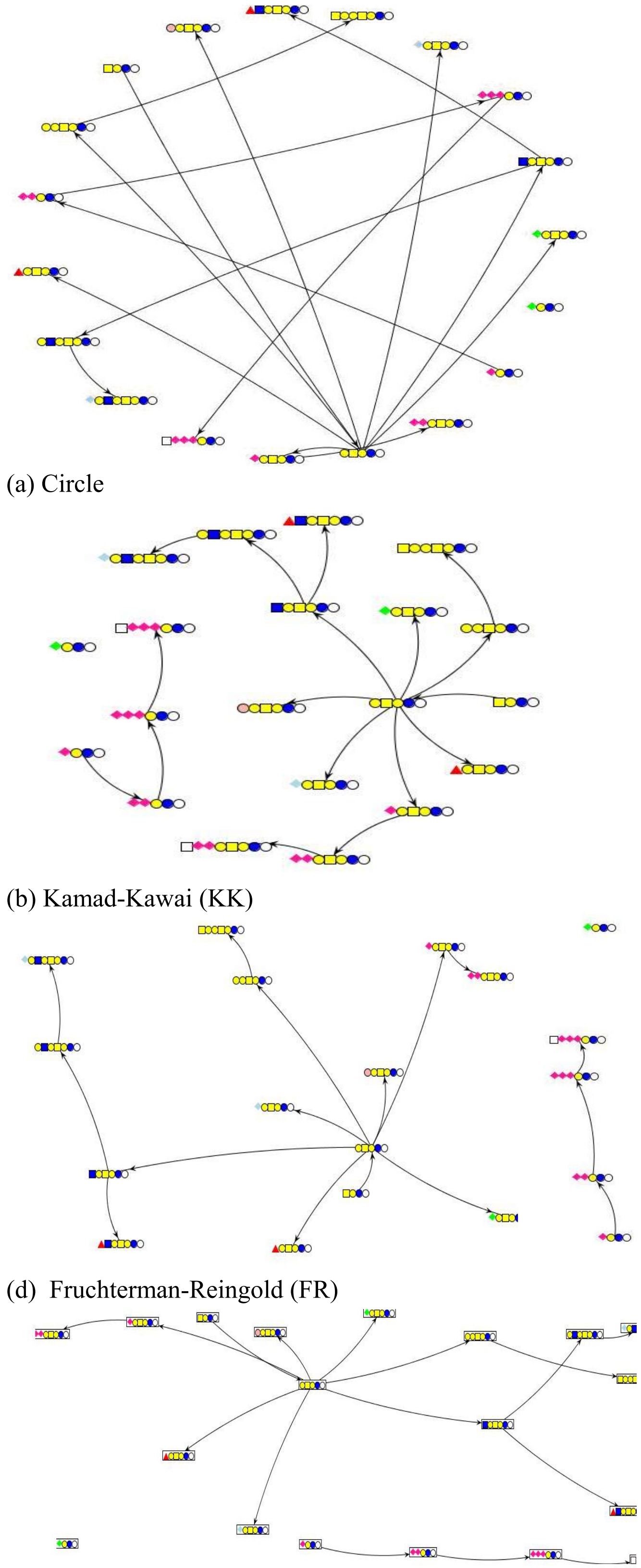
Fig. 2. Screen capture of interactive visualization of sphingolipid ganglio-series (25 molecules selected). Noteworthy features include spontaneous addition / deletion of molecules with increasing network complexity, choice of four different network layouts, and listing of detailed molecule data on mouse-over.

A. Dynamic network visualization with increasing complexity

The complexity of the networks is visualized "on-the-fly" so users may add or delete molecules for visualization spontaneously. To do this, all molecules are filtered from the database *a priori*. We derive the parent-derivative relationship between molecules as a hierarchical structure in memory, starting from the smallest chain to cover all molecules of the given series within the SphinGOMAP database. Thus, molecules may be added or deleted spontaneously in a pre-determined hierarchical order *i.e.* ordered sequence.

B. Choice of network layouts

We implement a choice of four different network layouts as seen in Figure 3. Of the four layout algorithms used, the Kamada-Kawai algorithm was most comprehensible even when the network complexity grew beyond 25 molecules. Circle layout works well with smaller numbers of molecules but becomes more complicated with a much larger number of molecules, for instance, 60 molecules.



(a) Circle

(b) Kamad-Kawai (KK)

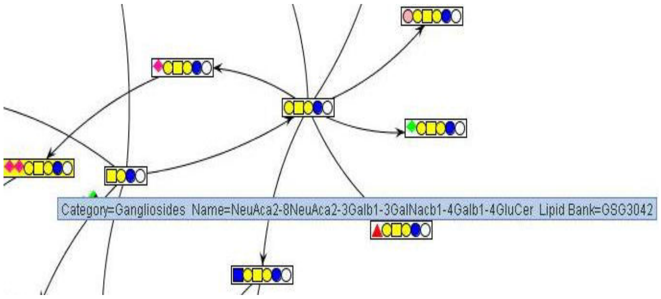
(d) Fruchterman-Reingold (FR)

Fig. 3. Circle (a), Kamad-Kawai (b), ISOM (c), and Fruchterman-Reingold (d) layouts of 25 molecules. Different layouts reveal different network features that may lead users to identify and implement suitable alterations in experimental design.

The self organizing graph layout algorithm (ISOM) renders the graph more widely using as much space as it needs for specifically rendering each relationship. As a result, the network structure becomes clearer, but less manageable. The Fruchterman-Reingold (FR) layout is similar to the ISOM layout and works fine with 25 molecules when the maximum edge length is smaller than 5 cm on a screen. However, it becomes unwieldy, in terms of providing a panoramic view, with a much larger number of molecules. Thus, different layouts may reveal different network features depending on the level of network resolution that users may desire.

C. *Mouse-over*

The mouse-over feature, presented in Figure 4, allows users to view detailed molecule data: common name, chemical category and Lipid Bank ID. Using this feature to incorporate more detailed sphingolipid molecule information from Lipid Bank and Lipid MAPS, we are eager to link biosynthetic network data from SphinGOMAP with primary and secondary data.



Category=Gangliosides Name=NeuAca2-8NeuAca2-3Galb1-3GalNacb1-4Galb1-4GluCer Lipid Bank=GSG3042

Fig. 4. Zoomed image of mouse-over feature. Detailed molecule data is revealed; this feature allows data from Lipid Bank and Lipid MAPS to be incorporated and linked for a more comprehensive interaction with specific sphingolipid molecules and pathways.

IV. Discussion

Collecting large-scale biological network data is not trivial, especially as immeasurable effort and resources are, and will continue to be, invested in performing experiments and gathering observations. Thus, data organization must rightfully receive equal emphasis, if not more, with the increasing focus on systems biology enabled by high-throughput technology.

To facilitate experimental design for systems biology research, we extend and improve the current SphinGOMAP in terms of interactive visualization. We do this by implementing (i) dynamic network visualization, (ii) choice of four different network layouts and (iii) mouse-over to reveal detailed molecule data. These features enhance user experience in dealing with high-volume, large-scale systems biology data. Thus, this work contributes to systems biology research by improving visualization, interactivity and usability of massive, complex biological networks.

Specifically for experimental design, a typical user may use our tool: (a) based on the feature of dynamic network visualization, to track the biosynthetic origins of a specific molecule, or to place it in the wider context of downstream derivatives or parallel molecules, i.e. "cousins", in terms of chemical inheritance; (b) to reveal, or confirm, previously unrecognized biosynthetic relations, using the various different network layouts, and (c) to examine / juxtapose research findings with the knowledge of metabolic data that could account for differences between experimental data and *a priori* network models by linking with Lipid MAPS.

Our implementation represents concrete improvement from previous visualizations in terms of data organization and representation built on software tools. Using this tool, we aim to close the loop for information flow in terms of directing *a priori* knowledge and community feedback from past experiments into better experimental design. Future work will focus on increasing the data collaboration between SphinGOMAP, Lipid Bank and Lipid MAPS. By linking biosynthetic origin data, molecular data and metabolic pathway data, scientists can look forward to a synergistic interaction with more aspects of lipid molecule information.

Acknowledgment

The authors thank Christopher Haynes, Mitchell Parry Jeremy Allegood and Todd Stokes for technical assistance and discussions; Lipid MAPS for the use of sphingolipid biosynthetic data via SphinGOMAP.

References

[1] A. Brazma et al, "Minimum information about a microarray experiment (MIAME) – toward standards for microarray data," *Nature genetics*, vol. 29(4), pp.365-71, Dec 2001.
[2] N. Le Novere et al, "Minimum information requested in the annotation of biochemical models (MIRIAM)," *Nature biotechnology*, vol. 23, pp.1509-1515, Dec 2005.
[3] C. F. Taylor et al, "The minimum information about a proteomics experiment (MIAPE)," *Nature biotechnology*, vol. 25, pp. 887-893, Aug 2007.
[4] C. F. Taylor et al, "Promoting coherent minimum reporting requirements for biological and biomedical investigations: the MIBBI project," *Nature biotechnology* (in press). Available: http://mibbi.sourceforge.net/ (URL).
[5] SphinGOMAP. Available: http://www.sphingomap.org (URL).
[6] Java Universal Network / Graph Framework (JUNG). Available: http://jung.sourceforge.net (URL).
[7] Lipid Bank. Available: http://lipidbank.jp (URL).
[8] Lipid Metabolites and Pathways Strategy (Lipid MAPS). Available: http://www.lipidmaps.org (URL).
[9] T. Kamada, and S. Kawai , "An algorithm for drawing general undirected graphs," *Information Processing Letters,* vol. 31, pp. 7-15, 1988.
[10] T. Fruchterman, and E. Reingold , "Graph drawing by force directed placement," *Software Practise and Experience,* vol. 21(11), pp.1129-64, Nov 1991.
[11] B. Meyer, Self-Organizing Graphs, "A neural network perspective of graph layout," *Proceedings of the 6th International Symposium on Graph Drawing*, pp.246-262, Aug 01, 1998.
[12] B. Meyer, "Competitive learning of network diagram layout," *IEEE Symposium on Visual Languages*, pp. 56-63, 1998.