Systems biology visualization tools for drug target discovery

Tianxiao Huan, Xiaogang Wu & Jake Y Chen†
†Indiana University-Purdue University, School of Informatics, Indianapolis, IN, USA

Importance of the field: Post-genome drug development has been driven by the need to study biological perturbations at the molecular system level. Systems biology visualization tools can help researchers extract hidden patterns from complex and large Omics data sets, model disease molecular mechanisms, and identify drug targets and drugs with good pharmacological and toxicological profiles.

Areas covered in this review: This review covers basic concepts in developing and applying information visualization tools to systems biology. We describe a framework and basic data representation schemes for visual data analysis in systems biology. We review major application areas of these visualization tools within drug discovery by focusing on early-stage drug discovery tasks such as disease biology modeling, target identifications and lead identification. We also show case studies and summarize our experience using visualization tools as lessons to our readers.

What the reader will gain: The reader will understand what visualization tools are available for diverse types of systems biology studies in drug discovery and understand how these tools can help advance drug development.

Take home message: In spite of the complexity inherent in systems biology, proper use of information visualization tools may reveal emerging properties hidden in the data and enhance chances of success for drug discovery.

Keywords: bioinformatics, drug discovery, systems biology, visualization tools


1. Introduction

Systems biology has become an important area of interdisciplinary study that can help researchers understand emerging properties of complex biological systems. In systems biology, computational models are developed to interpret biological data and predict system behaviors under perturbation conditions. This has set it apart from traditional molecular biology, which relies primarily on the study of experimental models to generate biological insights [1,2].

Adopting a systems biology approach has become a current trend in drug discovery [3]. Systems biology provides a new paradigm for drug developers to understand the effects of drugs on targets, target-interacting proteins, regulatory control genes, in downstream metabolic pathways, and in affected ‘biomolecular signaling circuits’ of cells, tissues and the entire organism [1]. Setting up molecular disease biology models has also become a prerequisite for finding effective cures, especially when the toxicology and pharmacology of drugs on biological targets can be difficult to understand in conventional cell-based models. In hepatocellular carcinoma (HCC), for example, aberrant activation of at least six signaling cascades, including the Wnt-catenin pathway, the EGFR-RAS-MAPKK pathway, the c-MET pathway, IGF signaling, Akt/mTOR signaling, and the VEGF and PDGF signaling cascades, are involved. More than 40 drugs and mAbs targeting 21 essential proteins related to these pathways are currently being investigated for HCC treatment [4]. To make significant impacts on drug discovery for complex diseases such as cancers,
Systems biology visualization tools for drug target discovery

**Article highlights.**

- Systems biology visualization tools provide drug discovery domain experts with fast-track access to rich biological context of drug targets, with decreased reliance on statistical and informatics experts to interpret the data first.
- High-throughput experiments, such as automated parallel chemical synthesis, high-throughput screening, genome analysis of protein, gene expression microarrays, yeast two-hybrid screens and affinity produced a large amount of primary data. Public databases, such as Protein Data Bank, Drugbank, Human Protein Reference Database and Human Annotated and Predicted Protein Interaction are available to conduct research in drug discovery.
- Visual representations include matrix, ball-and-stick graph and 3D representations. The most popular layout algorithms are radial, organic and hierarchical layouts.
- There are >50 publicly available visualization tools suitable for analyzing data in systems biology, from microarray profiles, molecule interaction network and signal pathway to integrative data analysis.
- Using visualization tools, we could overview the comprehensive drug-protein relationships in breast cancer by building up drug-target and breast cancer-related disease gene-protein interaction networks.

Biomedical researchers increasingly rely on the development of global perspectives. The quality of these perspectives further depends on the quality of analysis of complicating biological pathways, related drug targets, companion biomarkers and optimized chemical compounds with common structure/activity properties of initial hits.

Visualization plays an important role in helping analysts to extract useful information from systems biology studies. Good visualization tools take advantage of human visual perceptive strength in detecting patterns embedded in large, complex and noisy data sets. Specific to systems biology are >40 biological molecular network visualization tools, for example, Cytoscape (5), Osprey [6], Pathway studio [7] and ProteoLens [8], which help lay out molecular interaction networks as 2D graphs. These tools differ in their ability to encode similar molecular association relationships into topological proximities and incorporate additional information on molecular measurements as visual properties of the graph. With proper visual representation of attributes of the data, either raw or derived, visualization tools help domain experts accomplish similar tasks better by separating noise from real patterns.

To drug discovery domain experts, systems biology visualization tools provide fast-track access to rich biological context of drug targets, with decreased reliance on statistical and informatics experts to interpret the data first.

In this article, we present an overview of visualization tools currently used for drug target discovery, the earliest phase of the drug development process and where systems biology studies are making the most significant impact. We summarize major characteristics of these tools and hope to guide readers in choosing the best ones in specific application contexts. We also show several case studies which demonstrate how visualization tools are used by their developers to foster the drug target discovery today.

**2. Information visualization: an overview**

**2.1 Visualization framework**

Visualization is much more than displaying a pretty picture! It is an iterative process that may be built on an integrated data analysis framework (Figure 1). In a visualization-centric data analysis framework, a variety of experimental data are integrated with visualization; in silico models are built to identify good drug targets with visualization and biological hypothesis are validated with visualization. Exploratory data analysis by visualization, or ‘visual analytics’, can complement conventional automated data mining efforts in model fitting, attribute selection, clustering and classification. According to this paradigm, both primary data collected from original biological experiments and derived data processed after statistical filtering can benefit from visualization. Due to the diverse types of systems biology data used for analysis and the wide range of questions to address, single generic systems biology visualization tools cannot satisfy all the data analysis demands of drug discovery. Instead, specialized systems biology visualization tools have been developed, each addressing different sets of visual analytic needs.

**2.2 Systems biology data**

Large amounts of data from a wide range of high-throughput devices have been made publicly available to today’s systems biologists. Combinatorial chemistry, automated parallel chemical synthesis, high-throughput screening and high throughput -X-ray crystallography have contributed large chemical and protein data banks potentially useful for drug discovery [9]. Genome analysis of proteins, gene expression microarrays, yeast two-hybrid screens and affinity purification–mass spectrometry have also been developed for unraveling complex functions and relationships among molecules in biological models [10]. Usually, this information is stored in databases and can be queried during drug development. For example, as of early 2010, Protein Data Bank has collected >62,430 protein structures and its rate of structure collection has increased steadily [11]. DrugBank includes ~4800 drug entries and 2500 non-redundant protein drug targets [12]. Human Protein Reference Database has collected 38,806 literature-curated human protein interactions [13]. Human Annotated and Predicted Protein Interaction (HAPPI) database includes ~140,000 non-redundant, medium-to-high, quality computationally-curated human protein interactions [14]. These databases provide opportunities to integrate and mine molecular measurements and their relationships, with which detailed molecular network biology contexts for disease biology models may be built for drug discovery.
2.3 Visual representations

Different visual data representational schemes have been used in systems biology. The first scheme, matrix representation, is commonly used to represent association patterns of Omics data. In a 2D matrix, each column and each row can be used for a unique attribute, so that each intersecting cell represents a unique value given two attribute conditions, for example, columns for sample conditions, rows for genes and intersecting cells for gene expression values given the sample condition. To visualize a matrix, heatmaps are commonly used (using color on intensity to represent numerical ranges). Because the proximity of adjacent cells can be defined as those having similar column/row attribute values ordered by clustering results, a heatmap visualization of microarray results represented this way can, therefore, reveal sets of correlated gene expression patterns in grouped sample conditions [15]. Matrix representation can also be used to represent the pairwise association relationship between two biological entities as an adjacency matrix for the entity. Cluster analysis of adjacency matrix can directly show characteristic patterns hidden among all the pairwise association relationships in the system of entities. For example, Wu et al. ranked proteins in a yeast protein–protein interaction (PPI) network based on network topology, and gene lethality information and adjacency visualization show intriguing fractal-like patterns after clustering (Figure 2) [16].

A second scheme, ball-and-stick graph representation, is commonly used to represent biological molecular association networks. There is no optimal graph layout for any type of network; however, three types are commonly used in biological network visualization tools: radial, organic and hierarchical layouts [17]. In a radial layout, all the molecules are drawn as nodes at equal distance along circles (2D) or on spherical surfaces (3D), allowing users to identify network 'hubs' (highly connected nodes). In an organic layout, a spring-force physics-based model is established to calculate how close two linked nodes can be pulled together and how far apart two un-linked nodes can be pushed away until all the forces reach mechanical equilibrium. In a hierarchy layout, network nodes are aligned in hierarchical orders on horizontal lines so that edges are directed from nodes on lower horizontal lines to nodes on higher horizontal lines. After the network
is properly laid out, changing the visual attributes of nodes/ edges, including label, size, color and shape, can bring forth the full, rich set of visualization features. Specialized data representation formats such as Graph Modeling Language and PSI Proteomics Standards Initiative—Molecular Interactions enable interoperability of the laid-out network with other annotation features.

A third scheme, 3D representation, brings visualization to its full capacity, and is most applicable for molecular structure modeling [18] and 3D view of multi-scale biological networks [19]. A small molecules—protein docking program typically consists of a search algorithm for sampling the configuration and conformation space of a ligand/receptor pair, and an energy-based scoring function for ranking the potential solutions [20]. Compared with the 2D graph representation tools, these 3D visual modeling tools use more physics-based constraints, and spatial orientation of molecular or atomic structures are more precisely conveyed in the final visual models. In this review, we omit the discussion of tools in this scheme, because they have been extensively covered in many publications and special topic books [21-25].

3. Systems biology visualization tools

To identify good drug targets using systems biology approaches, researchers need to understand disease biology models by investigating the genomics and functional genomics changes between diseased and normal cells, and studying the global effects of drug compounds on the intended drug target. In Table 1, we summarize a variety of visualization tasks, visualization tools that support them and their applications in drug discovery. Next, we describe these visualization tools in detail, by relating them to systems biology studies that can lead to identification of new drug targets or drug compounds.

3.1 Functional genomics of diseases and drug response

Functional genomics studies of disease biology and drug response, particularly those using gene expression profile analysis, allow genome-wide analysis of biological samples at the expression activity level [26]. Gene expression profiling has been used to identify pharmacology and toxicology effects of drugs by revealing the different expression patterns of clinical samples of disease models before and after drug treatment. For example, Stegmaier et al. used gene expression signature to screen AML (acute myelogenous leukemia) differentiation-promoting compounds. They discovered that gefitinib, an EGFR inhibitor, can promote the differentiation of AML [27].

Visualization software plays a critical role in revealing clustering patterns of tens or hundreds of microarray experiments involving thousands of genes. Many visualization tools used in gene expression analysis have been reviewed recently, showing a trend of increased community interest in their development and applications [15,28,29]. Among them are widely used tools, including; GeneSpring (www.chem.agilent.com) and Spotfire.

Figure 2. A heatmap visualization of a breast cancer protein adjacency matrix, after ACO ranking/re-ordering of proteins in the protein interaction networks. A. Result for the protein interaction network constructed by including only protein interactions with a confidence score > 0.50 from the HAPPI database. B. Result for the protein interaction network constructed by including only protein interactions with a confidence score > 0.99 from the HAPPI database.

ACO: Ant colony optimization; HAPPI: Human annotated and predicted protein interaction.
Table 1. A summary of systems biology visualization tools and their application in drug discovery.

<table>
<thead>
<tr>
<th>Visualization category</th>
<th>Visualization tools</th>
<th>Application in drug discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional genomics of diseases and drug response</td>
<td>Cluster/Treeview, Spotfire, GeneSpring</td>
<td>Establish gene expression patterns to identify a desired drug effect for specific disease model [26]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Monitor dynamic changes in gene expressions in drug dose response studies [68]</td>
</tr>
<tr>
<td>Disease network biology modeling</td>
<td>PATIKAmad [58], VisANT [39], Caleydo [59], GeneTerrain [60], GeneSpring, BiologicalNetworks [57]</td>
<td>Display broad contextual information such as apoptosis molecular interaction network in cancer [69]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Integrate functional genomics and molecular networks specific to disease [70,71]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Study system dynamics as an influence network based on time-series functional genomics data [69]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analyze molecular interactions that become dysregulated in specific phenotypes [72]</td>
</tr>
<tr>
<td>Drug target and drug finding in molecular networks</td>
<td>Cytoscape [5], Osprey [6], Interviewer [19], ProteoLens [8], SNAVI [43], SNOW [42], JNets [41]</td>
<td>Help understand broad-range disease-disease associations [65]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Display specific contextual molecular interaction network that involve key disease drug targets [73]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify targets in tissue-specific molecular interaction networks [74]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Search for best targets and design multi-targeted drugs [74]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evaluate network topological feature for drug target proteins and develop therapeutic index [3,66]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evaluate new cocktail drug approach based on molecular networks [76]</td>
</tr>
<tr>
<td>Disease pathway biology modeling</td>
<td>Pathway Studio [7], PATIKAweb [77], PathVisio [53], Pathway Explorer [54]</td>
<td>Identify candidate targets from disease biology pathways [75]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Examine drug’s biological activity in participating biological pathways [2]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Predict drug effect by measuring its toxicological effects at the metabolic level [2]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correlate phenotypes with pathway-level molecular changes to guide targeted therapy [78]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Design drugs by modeling downstream signaling molecules inside a particular pathway [30]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Examine side effects of drugs on related pathways [79]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify pathway cross-talks based on key proteins and disease-modifying genes that span multiple pathways [80]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evaluate new multi-component drugs on metabolic changes as new therapeutic solutions [31]</td>
</tr>
<tr>
<td>Molecule docking and drug screening</td>
<td>AutoDock [81], GOLD [82], FlexX [83], DOCK [84], ICM [33], PostDock [85]</td>
<td>Dock drug to protein structures for structure-based optimization of lead compounds [86]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Screen novel ligands from compound libraries [87]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Estimate binding energies of inhibitors to target protein structures [88]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optimize multiple structure-activity relationships while maintaining drug-like properties [3]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Identify novel protein and small molecule drug binding [89]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Study off-target effects of drugs and its binding effect on other proteins [90]</td>
</tr>
</tbody>
</table>

(www.broad.harvard.edu/cancer/software/genepattern), which are commercially available, Cluster-view [32], GeneCluster 2.0 (part of GenePattern) (www.broad.harvard.edu/cancer/software/genepattern), MAExplorer (maexplorer.sourceforge.net) and Hierarchical Clustering Explorer [35], which are freely available. Most of them support visual data presentation schemes such as heatmap visualization, dendrogram visualization to view cluster structures, line graphs for dynamical profile comparisons and self-organizing maps for clustered data relationship. Commercially available tools are more comprehensive, integrating systems biology data analysis routines, spreadsheet visualizations and other types of genomics data such as genome locations, pathways and gene ontology annotations.

3.2 Disease-related molecular networks and drug discovery

Network biology has been the focus of systems biology. Biomolecular interaction networks, including physical PPI networks and gene regulation networks, are essential to the understanding of disease biology. In network biology, the model for identifying key genes or proteins suitable for drug target has shifted from isolated biochemical studies of specific
250 protein/gene functions to coordinated studies of global properties of a network of proteins/genes. New therapeutic intervention strategies call for the identification of perturbation effects of proteins/genes in disease biology networks, instead of fishing for ‘disease-causing’ gene from genes functional analysis [36]. Many studies have shown that disease-modifying genes/proteins are likely to physically interact or functionally associate with each other [37,38]. Therefore, studying disease-related proteins in their interconnected molecular network context is expected to reveal novel drug targets, which may improve pharmacological and toxicological profiles of compounds that are subsequently developed to affect the behavior of the disease-related proteins.

Suderman and Hallett [17] reviewed 35 biological network visualization tools in 2007 and several new powerful tools have been developed since that review. In Table 2, we list six new biological network visualization tools and compare their features with the popular Cytoscape [5]. Cytoscape contains all basic network graph visualization functions and a variety of customizable network display styles. It has a robust graph layout engine that allows for 13 layout styles and a customizable annotation engine for mapping data attribute values onto graph nodes/edges properties. It supports software plug-ins that can allow third-party software to extend its core functionality. More than 40 Cytoscape plug-ins are currently available. VizANT [39] offers built-in statistical functions to help users calculate network topological parameters and perform real-time network analysis. WebInter-Viewer [19] uses an ultra-fast graph-layout algorithm that can handle a large network layout of tens of thousands of nodes on a desktop computer. It also provides several network abstraction and comparison operators. A latest feature-rich database-driven network visualization tool, ProteoLens [8], supports direct database connectivity to either Oracle or PostgreSQL database tables/views, against which SQL queries may be specified to build visualizations. The robust query language embedded directly within the visualization software helps users to integrate Omics data analysis with powerful visual annotation and exploration of multi-scale biological networks. Arena3D [40] allows users to take a large data set and decompose a large network graph into a series of simpler 2D graphs to study. jNets [41] has a sub-network filtering function, which works by filtering annotations available on network nodes and edges, and subsequently helps users explore sub-network modules visually and statistically. SNOW [42] provides functions to calculate relevant network parameters and compare two networks with their corresponding network parameters. SNAVI [43] helps perform visual analytics of network node clustering, node degree of connectivity distribution calculations, network motif identifications and annotated web page linkings. GraphWeb [44] is a web server for biological network analysis and module discovery with visualization capabilities, all of which help users discover network modules in their data sets and link network modules with comprehensive systems biology database resources.

3.3 Disease pathway biology modeling

Biological pathways are abstract representations of interconnected sets of molecular reactions, interactions and regulations activated under specific conditions in a cell [45,46]. Pathways could be considered parts of larger biological networks. Here, we want to emphasize the part of pathway visualization software that differs from the software used for PPI network. Distinguished from molecular networks, pathways can encompass more than one type of biological process, and much more detail about molecular relationships. There are many biological pathway databases publicly available for disease systems biology studies today. The KEGG database includes 33,679 pathways from >100 species generated from 269 reference pathways concerning metabolism, genetic information processing, environmental response, cellular processes, human disease and drug response [47]. The BioCarta database contains >296 regularly updated pathways for multiple species (http://www.biocarta.com/). The Reactome database collects 1081 human pathways among 4076 proteins and 3669 biochemical reactions [49]. The recently published HPD database is constructed by integrating heterogeneous human pathway data from several pathway databases and contains 999 human pathways involving >59,341 human molecular entities [50].

Compared with network visualization tools, pathway visualizing tools can provide static molecular layouts, overlays of subcellular localization of components, annotation for different types of reactions, interactions and regulations, and integration capabilities to bring in literature annotation [45]. GenMAPP [51] is a freely available first-generation biological pathway visualization tool. It supports the display of changes of gene expression values on genes/proteins in displayed pathways. MAPFinder [52] is a GenMAPP built-in statistical tool that performs pathway-based gene/protein enrichment analysis. PathVisio [53] is another GenMAPP extensible software tool that allows advanced editing commands similar to those in diagramming software. PathwayExplorer [54] is a web server application that can map genes onto enzymes in metabolic pathways from various data sources. Pathway Studio [7] is commercial software that integrates text mining, molecular networks, gene-expression mapping, supporting literature curation and pathway diagramming/editing. The HPD database also provides heatmap visualization for users to compare the overlaps and differences between related pathways.

3.4 Integrated visual analytics

In disease systems biology studies, information about disease-related genes, proteins, networks and pathway consist of multi-scale and multi-dimensional attributes, making it challenging to map to 2D planes or 3D surfaces. For example, in studies that analyze functional genomic data in molecular networks and pathways, visualization not only needs to show the end result of expression level changes, but also to reorganize regions of ‘densely’ connected network modules or pathway
<table>
<thead>
<tr>
<th></th>
<th>Cytoscape</th>
<th>ProteoLens</th>
<th>GraphWeb</th>
<th>Arena3D</th>
<th>JNets</th>
<th>SNAVI</th>
<th>SNOW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Graph manipulation</strong></td>
<td>yFiles and GINY</td>
<td>yFiles</td>
<td>Graphviz and</td>
<td>Java 3D</td>
<td>TouchGraph,</td>
<td>WinGraphViz</td>
<td>TouchGraph</td>
</tr>
<tr>
<td><strong>Laying out network algorithm</strong></td>
<td>More than 13 kinds of layout styles</td>
<td>Force-directed, radial layout hierarchical, circular, orthogonal</td>
<td>Force-directed, radial layout hierarchical, circular</td>
<td>Two force-directed methods, hierarchical</td>
<td>Force-directed</td>
<td>Fixed visualization using presetting of parameters from GraphViz</td>
<td>Force-directed</td>
</tr>
<tr>
<td><strong>Drawing appearance</strong></td>
<td>Node shape/color/ border/label, edge color/style/direction/ label</td>
<td>Node shape/color/ border/label, edge color/style/direction/ label</td>
<td>Node label</td>
<td>Multiple layers</td>
<td>Node/edge color/ size/label</td>
<td>Node color/label</td>
<td>Node label</td>
</tr>
<tr>
<td><strong>Filters</strong></td>
<td>Select nodes/links according to properties (SQL-like)</td>
<td>Select nodes/links according to properties or using SQL statement for table attributes selecting directly</td>
<td>Automatic filters by edge weight and node annotation</td>
<td>Highlighted the selected nodes and their connecting sub-network</td>
<td>Automatic filters by node annotations; SQL-like filters</td>
<td>Pathway generator: select a source node and a target node and then find the paths that connect these two elements in a network</td>
<td>None</td>
</tr>
<tr>
<td><strong>Expand/collapse nodes</strong></td>
<td>Plug-in</td>
<td>Expand node neighbors</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td><strong>Database incline</strong></td>
<td>Plug-in</td>
<td>Common relational database</td>
<td>Limited by author provided Web-based</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Hide nodes</td>
</tr>
<tr>
<td><strong>System requirements</strong></td>
<td>Java applet or stand-alone</td>
<td>Java stand-alone</td>
<td>Java applet or stand-alone</td>
<td>Java stand-alone</td>
<td>Java applet or stand-alone</td>
<td>Java applet or stand-alone</td>
<td>C++ stand-alone</td>
</tr>
<tr>
<td><strong>Save</strong></td>
<td>GML, SIF</td>
<td>GML, XML session, Text, GML, XML, Oracle or PostgreSQL</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td><strong>Imports</strong></td>
<td>Text, GML, expression matrix, OB O</td>
<td>GML, XML session, Text, GML, XML, Oracle or PostgreSQL</td>
<td>Text, SIF, GML, XGMML and BioPAX</td>
<td>None</td>
<td>Text, XGMML</td>
<td>Text, XGMML</td>
<td>None</td>
</tr>
<tr>
<td><strong>Exports</strong></td>
<td>Graphical file, SVG, GML, network relations (text)</td>
<td>JPEG,BMP, GML network relations, node lists, selections node lists (text)</td>
<td>SIF, node lists, selections node lists (text)</td>
<td>JPG, 3D representations compatible to other 3D tools, VRML format and Pajek format</td>
<td>PDF, PNG</td>
<td>JPG, SVG, DOT</td>
<td>And some graph description files</td>
</tr>
<tr>
<td><strong>Comments and other features</strong></td>
<td>The importance of Cytoscape is its solid support for plug-in, growing number of which is available</td>
<td>Embedding the SQL query makes its software more flexible to suit powerful bioinformatics experts usage</td>
<td>Orthology mapping; multiple gene modules detecting methods; interprets discovered modules using GO, pathways and cis-regulatory motifs</td>
<td>Counts node types and computes connectivity distribution, characteristic path length, clustering and grid coefficient; finding motifs size 3 – 5 nodes; finding paths from source to target; generate random network</td>
<td>Nodes and edges were assigned to a group; node degree; GO calculating</td>
<td>Calculating topological parameters; degree, betweenness and so on</td>
<td></td>
</tr>
<tr>
<td><strong>Websites</strong></td>
<td><a href="http://www.cytoscape.org">www.cytoscape.org</a></td>
<td>bio.informatics.iupui.edu/proteolens/</td>
<td>bioinformatics.iupui.edu/graphweb/</td>
<td><a href="http://www.bioinf.manchester.ac.uk/nets">www.bioinf.manchester.ac.uk/nets</a></td>
<td><a href="http://www.bioinf.manchester.ac.uk/nets">www.bioinf.manchester.ac.uk/nets</a></td>
<td>code.google.com/p/snavi/</td>
<td>snow.bioinfo.cipf.es</td>
</tr>
</tbody>
</table>

*Table 2. A comparison of biological network visualization tools.*
regions in which clustered genes/proteins from statistical analysis can be presented for further visual analytic efforts.

An integrated visual analytic software tool makes it easy to integrate and interpret Omics results with molecular networks. Color Pathway [55] can help map time-series gene expression results onto KEGG pathways. It divides the enzyme nodes into columns, colors these sections according to their gene regulation values and sorts gene expression values onto a color matrix. GScope [56] attaches gene expression values to cellular networks by projecting each heatmap graph on a hyperbolic plane. Biological Networks [57] allows advanced bioinformatics users to integrate microarray data analysis with biomolecular interaction network analysis over a diverse set of databases that can be queried through powerful user interfaces. PATIKAmad [58] accepts experimental microarray data files and external database references as inputs to cluster and visualizes significant relationships between differentially expressed genes and external annotations. Caleydo [59] uses 3D views inside a cube to relate pathways, descriptive texts and gene expressions intuitively. GeneTerrain [60] renders a smooth 3D surface by adding the interpolated gene expression values for each gene in a molecular interaction network, which combines the strengths of heatmaps and ball-and-stick graph visualizations to help reveal subtle expression patterns hidden in the network. ACOR [16, 48] integrates the ant colony optimization (ACO) algorithm with matrix visualization to show how node-reordering in a biomolecular network can self-organize into clusters of nodes and fractal structures characteristic of the whole network. DrugViz [61], a Cytoscape plug-in, can import drug–target network information and help analyze drug–target relationship on the network scale, by displaying 2D structures of small molecules and providing bipartite graphs of associated drugs and targets. AMEN [62] is a unified visualization suite that provides modules for pre-processing microarray data, analyzing groups of significantly co-expressed genes, searching for enriched functional annotations within those groups, connecting them to PPI networks and annotating them with additional information such as cellular colocalization patterns.

4. Case studies

In this section, we present case studies in breast cancer to demonstrate how visualization tools help to expose drug mechanism of action and identify drug target proteins. All network visualizations performed with ProteoLens [8], Cytoscape, VisANT and so on also can be used to draw network graphs. We choose ProteoLens because of its ability to connect to and retrieve data from local systems biology database servers and its flexible visual analytics functions.

4.1 Drug target candidate prioritization

Breast cancer can be defined as a heterogeneous disease that results from the accumulation of multiple genetic and epigenetic defects. These effects lead to deregulation in many pathways such as amphiregulin/EGFR signaling, IGF, integrin, Notch, NF-κB, STAT, TGF-β and Wnt pathways [63]. Here, we show one strategy to exploit information visualization. First, we search the Online Mendelian Inheritance in Man database [64] to obtain breast cancer-related genes as seeds to construct the breast cancer molecular interaction network. These seed genes may be combined with previous knowledge of breast cancer proteins based on literature curation. Then, we build breast cancer specific protein interaction networks by expanding breast cancer associated genes/proteins in the HAPPI database [14] using a nearest-neighbor expansion method described in [65]. Each network that we build this way includes only protein interactions above a specific confidence score recorded in the HAPPI database to filter out low-confidence interactions unlikely to be true physical interactions. Next, we apply the ACOR algorithm [16, 48] to rank all the proteins in the weighted breast cancer-related protein interaction networks. The ranking results using two different confidence score thresholds (0.50 and 0.99) are visualized with the heatmap of an adjacency matrix that shows final ‘ant colony’ density distribution (Figure 2). While the network involving more medium-to-high-quality interactions (Figure 2A) shows global scale-free connectivity patterns, the high-quality interaction network (Figure 2B) shows a different pattern that is dominated by fewer number of highly clustered proteins. The top 20 proteins ranked from the high-confidence network (confidence score >0.99) are subsequently shown in Figure 3. These top-ranking proteins serve as important functional hubs relaying signals in the breast cancer disease sub-network, and, therefore, may serve as effective drug targets, based on related disease biology studies [65].

4.2 Drug–target poly-pharmacological network

The traditional view of drug action is as a ‘key’ fitting into a ‘lock’ which is oversimplified by underestimating downstream network cascade effects. The incomplete understanding of the biology of breast cancer hampers the identification of new therapeutic targets as well as the optimal use of the targets we know about. Systems biology studies in drug discovery can go further to help evaluate the global relationships among diseases, drugs and drug target proteins in the context of network pharmacology [66]. In this case study, we take advantage of visual analytic tools to show a breast cancer poly-pharmacological network that consists of candidate drug compounds, protein targets and drug–target relationships. In the breast cancer drug–target poly-pharmacological network (Figure 4), we highlighted 16 FDA approved breast cancer drugs, their 15 target proteins based on information collected from the DrugBank [12] and breast cancer disease gene-coding proteins linked to the drug target proteins by PPIs. The network visualization clearly shows the comprehensive relationships between drugs and proteins absent from those recorded in the DrugBank. We used ProteoLens to
draw Figure 4. We assigned the nodes different shapes for different molecular types and different colors to distinguish disease genes (red) and drug target proteins (orange). The round nodes represent a protein. The triangle nodes represent the FDA approved drugs used for treating breast cancer. The attributes can be mapped to nodes/edges automatically by user-defined association rules.

To study the comprehensive drug–protein relationships, we constructed a new set of networks based on information derived from biomedical text mining (Figure 5). First, we built a breast cancer specific protein interaction network, in which known breast cancer drug target proteins from the DrugBank are highlighted (Figure 5A). To obtain candidate drug compounds and their related proteins, we incorporated results from the breast cancer connectivity map ('C-Map') [67]. C-Map uses text mining to extract statistically significant associations between candidate drugs and drug targets in a specific disease context. Top 500 drug–drug target associations based on sorted p-values involving 20 compounds and 145 proteins are collected from the breast cancer C-Map. Next, we built a breast cancer candidate drug–drug association network (Figure 5B) in which two candidate drugs are connected if they share at least one protein. Finally, we built a drug target protein–protein association network (Figure 5C), in which two possible drug target proteins are connected if they share at least one drug. Here, distinct clusters of target proteins grouped in the same functional categories are clearly seen. Also, a large number of candidate drug targets are revealed, each connecting to more than one candidate drug for breast cancer. Therefore, visual exploration of the many-to-many relationship between candidate drugs and protein targets in the poly-pharmacological network can help drug developers to gain valuable insights with the help of visualizations.

5. Conclusion

There is a wide range of recently developed systems biology visualization tools suitable for drug target discovery applications. These tools can help integrate Omics data, lay out molecular networks, develop disease biology in silico models and perform limited simulation studies. Determining which tool to use is mostly a choice of personal preference, based on supported file formats and the visual analytic tasks at hand. In this review, we summarize major data representation schemes for these visualization tools, describe their application in systems biology and present two case studies in systems biology-oriented drug discovery. When applied properly, visualization can enhance statistical and bioinformatics data
analysis to reveal weak yet significant patterns hidden in functional genomics results. Essential to the success of these tools is the layout of multi-scale biological networks and pathways, which integrate disparate measurements of different biological entities into a coherent qualitative model. Although such models are static and semi-quantitative, ongoing development of increasingly sophisticated visual analytics tools will enable full-fledged simulation of drug perturbation of a biological system in the future.

6. Expert opinion

In spite of the complexity inherent in systems biology studies, proper use of information visualization tools may reveal emerging properties hidden in the data and enhance chances of success for drug discovery. As of today, there are >50 publicly available visualization tools suitable for analyzing data in systems biology. These tools help biomedical researchers unfamiliar with statistical and computational techniques find convenient ways to integrate large sets of gene expression profiles, organize Omics results in molecule interaction networks or biological pathways, and study how to best make use of emerging systems-level properties to improve the quality of drug targets and lead compounds. Here, we offer five suggestions to our readers interested in choosing and applying these tools to their research.

First, choose the right visualization format for the right data type. There is no magic bullet and do-it-all type of visualization that will work under all assumptions. Functional genomics data usually contain tens of thousands genes measured at dozens of conditions; therefore, some form of clustering or heatmap visualization will always be helpful to identify initial patterns. To identify a large set of relationships between molecular entities, ball-and-stick graph layout will work, only if the network that the relationships forms is relatively small and uncongested; otherwise, adjacency matrix scale up to large network and can show topological feature better [16]. Manual editing and addition of biological details are best accomplished by pathway visualization tools. When features to visualize exceeds six or seven that may be afforded by 2D network/pathway diagram attributes, there would be a higher demand for intelligent visualization tools that can project high-dimensional space to low-dimensional visualization space intuitively.

Second, consider visualizing large biological networks at multiple levels--global network, filtered network, sub-network
Figure 5. Breast cancer pharmacological networks. A. A breast cancer protein–protein interaction network. The network is constructed by using first-neighbor expansion methods [65] against the HAPPI database [14]. Only protein interactions with a confidence score >0.75 from the HAPPI database are retained. Seed genes are constructed from the OMIM morbid map using the MeSH query term ‘breast cancer’. B. A breast cancer candidate drug–drug association network. All data in this network are based on a filtered subset of breast cancer connectivity map (‘C-Map’). The edge width is drawn in proportion to the number of common target proteins shared between two drugs and the node size is in proportion to the number of possible target proteins the specific drug affects. C. A breast cancer drug target protein–protein association network. The edges and nodes are colored as targets of different drug categories.

and network modules. When the node or edge of a network becomes large, visualization tools generally perform poorly in laying out networks on the screen. Drug discovery is demanding to any visualization technique, which does not yet support the examination of networks that span multiple scales, from the global scale covering tens of thousands of molecular interactions to local molecular association relationship details covering specific biochemical mechanism of interaction. Therefore, researchers should apply different types of filters and constraints to manipulate biological networks, for example, filter the network by quality, type of interaction, tissue-specific expression, cellular location, relevance to a biological process and small groups of interacting molecules within the networks [91]. These network visual analysis tasks cannot be automated by visualization tools and must be rational decisions made by the analyst. As we demonstrated in case studies, we present the brief view of drug–drug target map in breast cancer according to DrugBank records. But by text mining examples, text mining of the biomedical literature can potentially be used to enable the large-scale of disease knowledge to screen out the drug discovery statues and the mapping associations of drugs and target proteins [67]. It provided so much potential drugs and target proteins information, but not all proposed targets may be druggable, and the network of drug–drug target mapping would be a mess. So we retrieved the drug–drug associations and target-target associations, in order to view the drug–drug targets in a higher level view. These views replaced direct map view and were better for analyzing complex networks.

Third, choose a visualization tool that supports extensive linking, querying and integration of external annotation data. While standalone visualization tools are useful, to gain novel insights usually requires further validation of visual data patterns revealed. External data sources such as gene ontology, protein function and genomic annotation can be tedious to access for biomedical domain experts. Over time, we find tools that support extensible plug-ins (e.g., Cytoscape), entire data analysis in one suite (e.g., Pathway Studio), powerful database query language support (e.g., ProteoLens) or online functionalities (e.g., GraphWeb [41]) have become our favorite choices for this reason.

Fourth, do not choose visualization because you want to avoid functional genomics or systems biology data analysis. Visualization and visual exploration of the data make it easy to generate new hypothesis. By linking gene expression profiles in different drug perturbation conditions to molecular networks and pathways, the dynamic effect on gene clusters organized in biological context becomes clear. However, visualization cannot quantitatively help researchers decide which genes or proteins are more sensitive, more specific, less variable and more significant in these experiments. Current biological network visualization tools cannot visualize pathway kinetics. Therefore, visualization cannot replace the hypothesis-driven data analysis and knowledge synthesis process.

Fifth and last, expect significant innovations to take place in systems biology visualizations. We are only at the beginning phase of generating new wave of Omics data. Next-generation sequencers, high-density microarrays, protein chips, tandem mass spectrometers and high-throughput screening robots are generating data at an unprecedented rate. Systems biology studies, including data analysis and visualization tool developments, are still at its infancy. Here, we propose three aspects for systems biology visualization software developers to be considering. First, systems biology visualization tools will need to go beyond simplistic graphical models and find new ways of integrating dynamic information into the network. Specific to drug discovery contexts, these tools also need to make it easier to integrate and ‘data mine’ drug compound information. Second, it needs to design some advanced tools designed for users who have powerful programming skills. Bioinformatics experts have a lot to learn from computer science, in which innovation in information visualization techniques have been fueling the growth to harness information available through mobile computing and multimedia computing. Accelerated insights and techniques from general-purpose computing domain are expected to propel significant progress of visualization tool development. Other cutting edge visualization tools, such as Gapminder (http://www.gapminder.org/), could be brought into systems biology research by taking advantage of its special features and widening the view of systems biologist. Third, it needs to design new methods to auto-mine the specific protein sub-network, according to compounds/drugs information. During our literature review process, there is only one tool, DrugViz [61], which could simply display small molecule information in protein networks. There is an improving space for drug discovery by automatically viewing the compound binding status in molecule interaction contexts and pathway contexts. To drug developers, the best such tools enabling future systems biology studies are yet to come.

Acknowledgments

The authors thank M Grobe of the Pervasive Technology Institute at Indiana University for proofreading the manuscript.

Declaration of interest

This work is partly supported by the Department of Defense (DOD) Breast Cancer Research Program (BCRP) Concept Award (W81XWH-08-1-0623) to J Chen.
Bibliography

38. Kann MG. Protein interactions and disease: computational approaches to
uncover the etiology of diseases.


BMC Syst Biol 2008;2:104


BMC Bioinformatics 2009;10:95


Nucleic Acids Res 2006;34(Database issue):D504-6

47. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes.


Nat Genet 2002;31(1):19-20


Genome Biol 2003;4(1):R7


BMC Bioinformatics 2008;9:399


Proc IEEE 2002;90(11):1793-802


Bioinformatics 2003;19(3):437-8


Nucleic Acids Res 2006;34(Web Server issue):W466-71


Proteomics 2008;8(11):2196-802


Bioinformatics 2009;25(20):2760-1


Inf Vis 2008;19(9):1-12


BMC Bioinformatics 2008;9:86


Mol Syst Biol 2008;4:169


Affiliation
Tianxiao Huan¹, Xiaogang Wu²,⁴ & Jake Y Chen²,³,⁴ PhD
¹Author for correspondence
²Shandong University, College of Life Sciences, Shandong, IN 250100, PR China
³Indiana University-Purdue University, School of Informatics, Indianapolis, IN 46202, USA
⁴Professor, Indiana University-Purdue University, School of Informatics, Indianapolis, IN 46202, USA
Tel: +1 317-274-7542; E-mail: jakechen@iupui.edu

¹Shandong University, College of Life Sciences, Shandong, IN 250100, PR China
²Indiana University-Purdue University, School of Informatics, Indianapolis, IN 46202, USA
³Professor, Indiana University-Purdue University, School of Informatics, Indianapolis, IN 46202, USA
⁴Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, IN 46202, USA