Visual Exploration across Biomedical Databases

Michael D. Lieberman, Sima Taheri, Huimin Guo, Fatemeh Mirrashed, Inbal Yahav, Aleks Aris, and Ben Shneiderman

Abstract—Though biomedical research often draws on knowledge from a wide variety of fields, few visualization methods for biomedical data incorporate meaningful cross-database exploration. A new approach is offered for visualizing and exploring a query-based subset of multiple heterogeneous biomedical databases. Databases are modeled as an entity-relation graph containing nodes (database records) and links (relationships between records). Users specify a keyword search string to retrieve an initial set of nodes, and then explore intra- and interdatabase links. Results are visualized with user-defined semantic substrates to take advantage of the rich set of attributes usually present in biomedical data. Comments from domain experts indicate that this visualization method is potentially advantageous for biomedical knowledge exploration.

Index Terms—Data exploration and discovery, bioinformatics, information visualization.

1 Introduction

The amount of publicly available biomedical data has ballooned in the past several years with ever-improving technology and computational methods. In addition to increased digitization of biomedical publications, improved text mining and natural language processing techniques allow for the extraction of thousands of unique relationships from biomedical text collections. This vast quantity of biomedical data presents a unique domain-specific challenge for interface designers: what are appropriate design choices that will help knowledge discovery and exploration within the biomedical domain?

To ground our discussion, we focus on one of the most important and largest collections of biomedical data freely available on the Internet, that of the National Center for Biotechnology Information (NCBI). The NCBI maintains over 30 public databases containing biomedical information of various types, such as published medical documents (PubMed), gene listings (Entrez Gene), protein listings (Entrez Protein), and DNA sequence information (Entrez Sequence). It also stores and manages pairwise associations between records in the databases according to the various types of content. For example, a particular document d listed in PubMed might be associated with all genes G from Entrez Gene that are mentioned in d. d may also have associations with other PubMed documents that cite d as a reference, as well as associations to the PubMed documents that d itself cites. Furthermore, each gene $g \in G$ could have associations with the proteins for which g codes, or the DNA sequences in which g's code appears. Usually, the various types of records in these databases also have many attributes associated with them. For example, PubMed documents might be annotated with the date of publication, authors, and general topics, while gene records could be annotated with the relevant species, location on chromosome, or function. This rich space of record attributes is key in aiding understanding of the data.

Given the huge amount of data at NCBI, and the large number of databases, myriad variations of these associations are possible. To organize these data in a way useful for knowledge exploration, note that NCBI's multiple databases can be abstracted as a massive *entity-relation graph*. In this graph, nodes correspond to individual knowledge points or database records, such as documents, genes, proteins, and other object types. Associations between database objects can then be modeled as directed or undirected links in the graph, connecting related nodes. The entity-graph model has already been applied to various document collections, including some in the biomedical domain, and much research has dealt with providing a broad overview of research publications and trends by visualizing the graph, typically using a *force-directed* node layout scheme [19], or other schemes such as circular [13], matrix-based [7], hierarchical [17], [37], [39], or layered [12], [35], [46] node layouts. These types of top-down visualizations simplify the identification of concepts like *research fronts* [3], [15].

However, our motivation lies not in discovering overall trends, but rather in accomplishing the everyday technical tasks of knowledge exploration and discovery undertaken by biomedical scientists and researchers. Scientists researching a particular gene, protein, or topic want to find specific and relevant information that will aid in their research. As a

- *M.D. Lieberman, S. Taheri, H. Guo, F. Mirrashed and B. Shneiderman are with the Institute for Advanced Computer Studies, Department of Computer Science, A.V. Williams Building, University of Maryland, College Park, MD 20742.*
  E-mail: {codepoet, taheri, hmguo, fatemeh, ben}@cs.umd.edu.
- *I. Yahav is with the Department of Decision, Operations and Information Technologies, R.H. Smith School of Business, University of Maryland, College Park, MD 20742. E-mail: iyahav@rhsmith.umd.edu.*
- *A. Aris is with the Department of Computer Science, University of Maryland, College Park, MD 20742. E-mail: aris@cs.umd.edu.*

Manuscript received 15 Oct. 2008; revised 25 July 2009; accepted 15 Sept. 2009; published online 7 Jan. 2010.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-10-0181.
Digital Object Identifier no. 10.1109/TCBB.2010.1.

1545-5963/11/$26.00 © 2011 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

result, when using NCBI's databases, they begin with a specific query or set of queries, and explore outward from the initial query result. They might also cross-reference records from multiple databases. Our visualization tools are designed to aid this query-specific exploration.

Even though the NCBI databases form an implicit entity-relation graph, the NCBI's current web interfaces offer no option to explore multiple areas of the graph simultaneously. Researchers explore the NCBI databases by retrieving a single page of information at a time, essentially limiting them to viewing a single node at a time. They must continuously click forward and backward to retrieve additional information from other NCBI databases. However, based on our interactions with biomedical domain experts and the kinds of exploratory tasks they undertake, we believe that explicitly viewing and exploring multiple nodes in parallel will lead to improved performance in exploration and discovery tasks. We provide a data collector for the NCBI databases that enables this exploration by initially retrieving a query-based subset of nodes from one or multiple NCBI databases. Users then specify a *query tree* that defines a data retrieval path between databases. For example, users interested in the role of genetics in alcoholism could use the data collector to perform a document keyword search for "alcoholism" within medical literature and disease databases, and retrieve links from the resulting heterogeneous set of nodes to gene records in another database.

To make use of the typically rich attribute space of biomedical data, we display graph nodes and links using a visualization technique known as *user-defined semantic substrates* [2], [42], as implemented in the Network Visualization by Semantic Substrates (NVSS) tool. Fig. 1a shows one such visualization of a query about "cervical cancer" across three NCBI databases. Unlike force-directed layouts, semantic substrates rely mainly on node attributes for meaningful and regular node placement into user-defined regions. These regions allow users to attach specific semantics to node positions on-screen, and enable a simple filtering paradigm based on node position. In contrast to other methods, semantic substrates offer expert users fine-grained control over the placement of nodes and their spatial meaning. Furthermore, different node layouts can be effected by simply designing additional substrates, allowing various views of the same data, and possibly leading to different insights about the data. Semantic substrates are thus especially suited to the display of biomedical data, and represent a drastic improvement over force-directed layouts used in similar situations, such as that in Fig. 1b, as well as other node layout strategies. We used the NVSS tool to visualize the results of several queries in NCBI's biomedical databases. On reviewing our work, domain experts indicated that it shows a strong potential toward biomedical visualization applications.

The paper proceeds as follows: Section 2 contains a survey of related work in network visualization and its applications within the biomedical domain. Next, Section 3 describes the data collection process, including our data model and data collector design. In Section 4, the concept of semantic substrates is further elaborated upon, and the controls and methodology of designing and visualizing semantic substrates are introduced. Section 5 provides several visualizations of sample queries that demonstrate the power and breadth of semantic substrates in cross-database exploration, while Section 6 contains an evaluation of our visualization methods by domain experts. Finally, Section 7 outlines further avenues of improvement and concluding remarks.

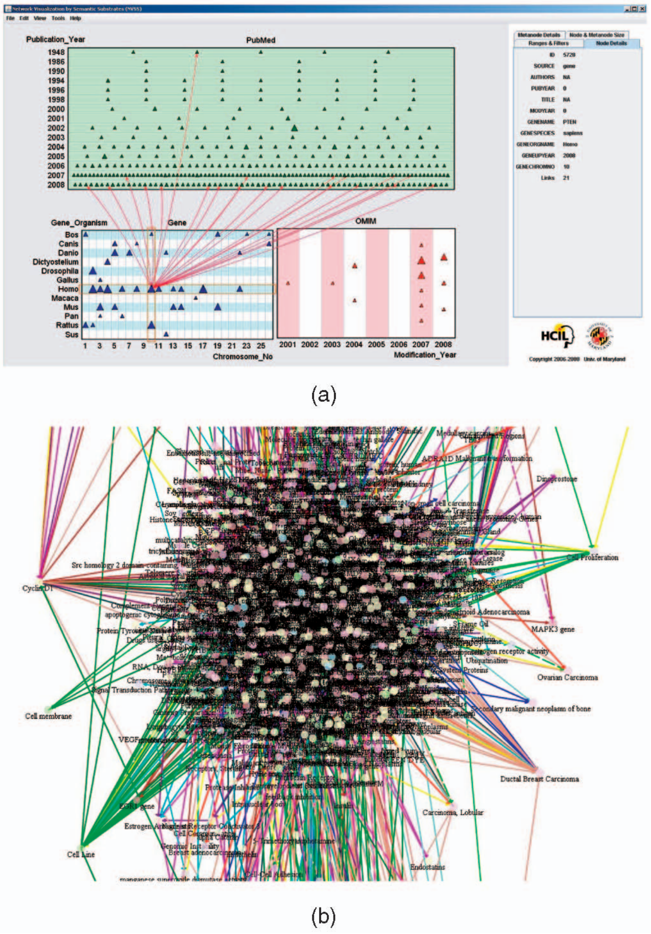


Fig. 1. For large cross-database exploration, (a) semantic substrates provide a more useful node layout when compared to a (b) force-directed layout. In a semantic substrate, nodes are placed into separate substrate regions based on attribute values (here, source database) and are further organized within each region using additional attributes. Link filters allow fine-grained exploration. (a) Semantic substrates. (b) Force-directed.

2 Related Work

Network visualization has a long and distinguished research history. In this section, we provide a brief survey of work in network visualization, and some of its many applications within the biomedical domain. For broader overviews, refer to di Battista et al. [16], Herman et al. [26], and Suderman and Hallett [45].

2.1 Network Visualization Methods

The vast majority of network visualizations make use of force-directed layouts [19]. The basic idea behind the force-directed layout is to model links as mechanical springs or attractive forces between nodes, and nodes as exhibiting repulsive forces. It thus tends to draw connected nodes together while separating unlinked nodes. This layout is favored because it tends to reduce the number of link overlaps, and reveals clusters that are not necessarily known by users. Popular alternatives to the force-directed layout include circular and radial [13], hierarchical [17], [37], [39], layered [12], [35], [46], and matrix-based [7] layouts. Circular layouts generally place nodes around central pivot nodes, which allow for a simple one-dimensional ordering of nodes around the circle. Hierarchical layouts group nodes into clusters, usually based on link strength. Layered layouts, often used with temporal placement strategies and citation networks, create levels of nodes with each level's nodes sharing a common attribute, and create an arrangement of nodes within each level to satisfy various graph drawing aesthetics [46], such as minimizing link crossings. Matrix-based layouts offer a visual representation of an adjacency matrix, which avoids the node and link occlusion problems of node-link methods.

However, the main drawback of these methods when compared against semantic substrates is that they take a very limited or nonexistent consideration of node attributes, which are generally prominent in biomedical data sets, and therefore do not impart any meaning to nodes' spatial positions. In particular, the force-directed, hierarchical, and matrix-based layouts are overly dependent on link attributes to determine node positioning. Circular and layered visualizations have been used in limited ways to augment nodes' spatial positions with some meaning, such as ordering or grouping nodes in terms of time, and so share some similarity with semantic substrates. However, they lack placement methods in terms of multiple node attributes. In other words, they are limited to one-dimensional node ordering. On the other hand, semantic substrates can impart meaningful significance to nodes' spatial positions across multiple dimensions.

Another drawback of these visualization methods is that as the number of nodes and links grows, the resulting layouts grow cluttered and difficult to understand. To illustrate, Fig. 1b shows a force-directed layout of concepts related to breast carcinoma. Even though the concepts and relationships depicted in the figure vary widely, the large number of links between nodes causes them to be tightly grouped and unreadable. It is difficult if not impossible to identify and explore interesting relationships or patterns in the data. In contrast, semantic substrates provide fixed node positions based on node attributes, which makes interesting nodes easy to identify by their attribute values. Also, powerful and interactive methods of node and link filtering provide a simple means of making sense of larger data sets.

Of course, it maybe meaningful to incorporate some of these visualizations into semantic substrates to allow more compelling exploration of particular biomedical data sets. See Section 7 for a discussion of some of these potential additions.

2.2 Biomedical Visualizations

Because many biomedical concepts and relationships can be abstracted as networks (e.g., molecular interactions, metabolic pathways, regulatory networks, and disease correlations), many visualization systems have been developed that cater to exploration of specific knowledge domains or biological networks. These visualizations abstract some knowledge domain as a network representation, and then use a corresponding network visualization method to display the data. Generally, the layout method is chosen based on the general network topology of the underlying knowledge domain.

Many systems have been developed for visualizing molecular interactions and pathways within and across data sets, including Cytoscape [40], Pathway Studio [38], Osprey [13], WebInterViewer [24], ProViz [33], VisANT [30], PathBank [28], BiologicalNetworks [4], and most recently, ProteoLens [31]. Most of these systems' visualizations are based on specialized forms of the force-directed node-link layout [6], [34], though most offer alternative network views such as circular, radial, hierarchical, or layered layouts. Because they are designed for exploration, many of these systems offer querying capabilities based on statistical attributes of the network, or local topology, to draw attention to interesting parts of the network. Some (e.g., Osprey [13], and VisANT [30]) also allow for selective expansion of network nodes, rather than displaying the entire network, and many are extensible via user plug-ins.

There are also several visualization tools developed to aid analysis of interspecies relationships, based on genomic or phylogenetic data. Fung et al. [20] evaluates the effects of using two visualizations, based on matrices and bipartite graphs, on DNA microarray data analysis. Shaw [41] presents another analysis-based visualization technique where the similarity of gene order across species is displayed as a node-link diagram using a force-directed layout. Also, a number of systems visualize phylogenetic networks, which represent species as nodes and ancestral relationships as edges. Huson [32] presents one popular method for layout of phylogenetic networks called SplitsTree, and Gambette and Huson [21] describe a number of algorithms for drawing split networks.

To visually explore relationships and connections between diseases and their associated genes, Goh et al. [23] generated two complementary network projections that they term as the human disease network (HDN) and the disease gene network (DGN). In the HDN, nodes are disorders and they are connected if they share a disease-causing gene, while in the DGN nodes are genes and they are linked if they are associated with the same disease. They use a force-directed layout to generate these two projections, and use color and size coding to impart information about the diseases and genes in question. They analyze the graph noting interesting statistical and topological properties such as apparent clusters of diseases or genes. Based on this work, Muhammed et al. [36] create and analyze a drug—target network, which visualizes associations between drugs and the proteins that they target or affect.

A number of tools were developed specifically for visualizing large biological networks. Pajek [5] is a popular software package for visualizing biomedical data sets such as DNA interactions and genealogies, as well as a variety of other large networks including citation networks. Adai et al. [1] present a large graph layout algorithm that computes a Minimum Spanning Tree (MST) of the network, and then uses the MST for node layout based on an iterative force-directed algorithm. Using this algorithm, they created and explored a large protein homology network. Also, the knowledge visualization tool VxInsight [11] displays large networks of information such as documents or genomic data as a 3D mountainous landscape, where peaks correspond to important data points. Boyack et al. [10] used VxInsight to study how genes, protein, and papers related to melanoma are interconnected via co-occurrence patterns of Medical Subject Heading (MeSH) terms. They generated a manually annotated Paper-Gene-Protein map (papers from MEDLINE, genes from the Entrez Gene database, and proteins from UniProt) using a force-directed layout algorithm to see prominent co-occurrence relationships. Their visualization was used to find "bursty" genes related to melanoma research, indicating possible correlations. Pathway Studio [38], Osprey [13], ProViz [33], and VisANT [30] also specialize in large networks. Additional commonly used tools for biomedical network visualization include the Prefuse toolkit [25] and GraphViz [22].

Several visual tools have been created for exploring databases of biomedical literature and the mined semantic relationships found within. Arrowsmith [44] is a textual web-based tool that supports the discovery of relationships between two sets of literature in MEDLINE, a database of life science and biomedical information. Arrowsmith lets users look for items or concepts that maybe common between two distinct sets of articles, but the presentation of query results is limited to its text interface. In a similar vein, the iHOP system [29] attempts to mine gene and protein co-occurrences through manual user specification of sentences of interest from PubMed abstracts. Aggregated results are displayed as an entity-relation graph of genes, but again with no regard to each gene's attribute values. In other words, node layout is purely link-based and ignores the intrinsic qualities of each gene, which if used would provide additional exploratory value. CiteSpace [14] is a system for detecting and visualizing trends and changes in scientific disciplines and their corresponding literature over time, based on clusters of important keywords. To explore the literature, two complementary visualizations are presented that are based on cluster views and temporal views. CiteSpace uses a force-directed layout to provide a high-level overview of important or seminal works. Bodenreider and McCray [9] examine several network visualization methods to explore semantic relationships from the Unified Medical Language System, a database of medical concepts and terminology gathered from many medical vocabularies.

3 Data Retrieval

To facilitate exploration of the NCBI databases, we designed a data collector to retrieve a small subset of the entire data collection, based on an initial keyword query and coupled with subsequent node and link expansion. In this section, we describe our entity-relation graph model in more detail, as well as the web services available through NCBI's website, both of which influenced our final data collector design.

3.1 Database Model

Fig. 2 shows the largest of NCBI's databases, along with interdatabase associations. As mentioned previously, rather than keeping these databases distinct within our data collection and visualizations, we abstract NCBI's databases into an *entity-relation graph*. In this graph, NCBI database items, such as documents, genes, and proteins correspond to nodes of the graph. Nodes have unique identifiers as well as many other attribute values corresponding to each node's associated information. Also, the set of attributes varies according to the type of node. For example, PubMed document nodes include attributes for the document's title, authors, year of publication, and keywords, while Gene nodes have attributes for gene name, genus and species name, and chromosome location, among others. To ensure meaningful node placement in semantic substrate regions, several node attributes that represent semantic information should be selected from each node. Fortunately, each node type in the NCBI graph has many attributes and it was therefore easy to settle on appropriate attributes for node layout.

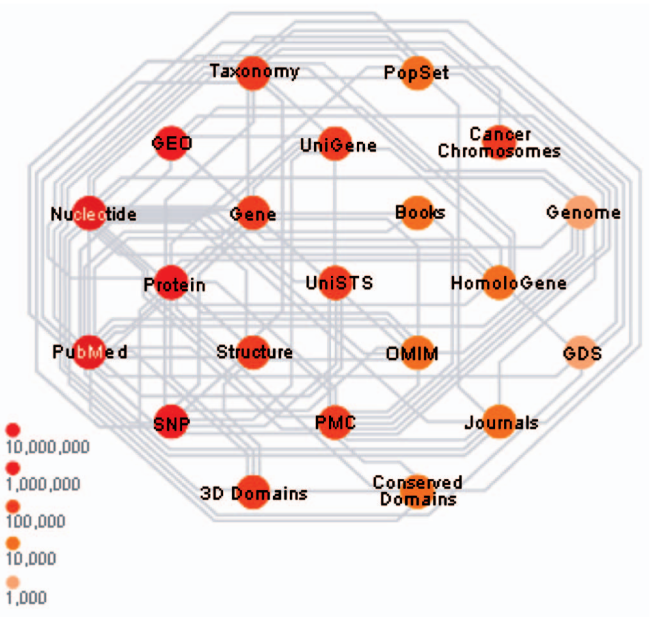
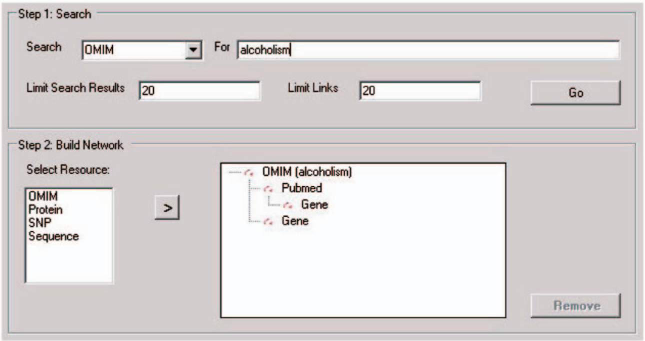


Fig. 2. The largest of NCBI's databases, where each node corresponds to a database, and node color representing the database's size. Links between databases are also shown. Associations between databases are numerous, and being able to explore these relationships is key to understanding data from these databases (from http://www.ncbi.nlm.nih.gov/Database/).

While database items correspond to nodes of the entity-relation graph, database associations correspond to links of the graph. Each link has pointers to the two nodes that it joins, as well as a type classification, such as "document citation" or "content similarity." In addition, the graph has several link properties that make it more difficult to visualize:

1. Nodes may be connected by multiple links.
2. Links may be weighted or unweighted.
3. Links may be directed or undirected.

For example, suppose two medical report documents d_1, d_2 concerning breast cancer appear in the PubMed database. d_1 may cite d_2, so after mapping to the graph, a "document citation" link would point from d_1 to d_2. Likewise, because d_1 and d_2 have similar content, an undirected "content similarity" link, with weight of 0.9, might join d_1 and d_2. It is therefore vital to incorporate a method of distinguishing or filtering links based on link attributes.



Step 1: Search

Search: OMIM For: alcoholism

Limit Search Results: 20 Limit Links: 20

Step 2: Build Network

Select Resource: OMIM, Protein, SNP, Sequence

- OMIM (alcoholism)
  - Pubmed
    - Gene
  - Gene

Fig. 3. The main interface of our data collector. Users specify a keyword query in the search entry field, and a query tree representing the path of data retrieval. The data collector then traverses the query tree, collecting nodes and links along the retrieval path. Here, "alcoholism" will be the initial query in the OMIM database, and links into the Gene and PubMed databases will be collected.

3.2 Data Collection

To retrieve data, we use the Entrez Programming Utilities (eUtils).[1] eUtils is a programming interface to the Entrez Global Query Cross-Database Search System outside of NCBI's regular web browser-based query interface. We use the following eUtils services in our data collector:

1. *eSearch*, which executes a keyword search query in a specified database, returning a set of matching record ids and relevance scores.
2. *eLink*, which retrieves links from a given set of record ids to record ids from another database.
3. *eFetch*, which retrieves all record attribute values for a given record id.

Responses are retrieved in an XML format, and thus are easy to parse with any standard software. In addition to eUtils, NCBI provides a web service that offers access to the Entrez Utilities via the Simple Object Access Protocol (SOAP). We developed our data collector in C# .NET using this service.

3.3 Data Collector Design

As we want our tools to be used by as wide an audience as possible, we designed our data collection tools to work with any NCBI database that users might want to query. Fig. 3 depicts our data collector's user interface. It enables user queries of NCBI databases by making use of a keyword *search query* coupled with a *query tree*. The search query is a collection of keywords and an initial database (e.g., in Fig. 3, OMIM) in which to search. The query tree is a specification of the requested links between the search results and entities in other databases. This tree represents the path of data retrieval that will be taken by the data collector. Each node in the tree corresponds to one of NCBI's databases, with the root node corresponding to the initial search query's database. Links between nodes in the query tree represent interdatabase links that will be collected between records from different databases. To retrieve intradatabase links, users can link a database to itself. For example, in Fig. 3, the user has specified "alcoholism" as the initial query string and OMIM as the initial database. The corresponding query tree specifies that links from OMIM to the PubMed and Gene databases should next be retrieved, as well as links from the resulting set of PubMed documents into the Gene database. This query tree will cause the data collector to retrieve links into the Gene database from multiple sources, allowing for potentially interesting visualizations that make it easy to find the genes most linked with alcoholism.

The data collector proceeds by executing an eSearch query in the database corresponding to the root node of the query tree. It then gathers data links and nodes by traversing the tree and executing corresponding eLink queries. Finally, node attributes for all collected nodes are retrieved using the eFetch utility. The data collector generates two tab-delimited text files as output corresponding to node data and link data.

Because our visualization method (see Section 4) relies heavily on node attributes for node placement, we designed our data collector to allow users a flexible definition of required attributes without interacting with the data collector's code. We store a list of attributes to be retrieved outside of the software in an external XML attribute description file. Each attribute in the XML file is composed of attribute name, attribute type, optional filter string, an indicator of whether the attribute can have multiple values (e.g., author names), and optional conversion rules for specific attribute values. These conversion rules were sometimes necessary to resolve Entrez database field inconsistencies. For example, Gene records corresponding to human genes have attributes for the chromosome on which the gene is found, which can be either numeric (e.g., 9) or nominal (e.g., X and Y). Using a conversion rule, we mapped X and Y to chromosome numbers 23 and 24, respectively, to allow for more meaningful Gene node positioning.

3.4 Data Retrieval Limitations

NCBI enforces rate limits for programs using the XML eUtils interface. Programs using the interface are limited to a single request every three seconds. In addition, the system imposes limits to avoid, particularly, time-consuming queries. If a query takes longer than 30 seconds to complete, the query is canceled and no results are returned. These limits create a challenge for interactive retrieval of query result nodes and links, as the data collector must obey these limits to ensure a full set of query results. They also prevent the timely retrieval of potentially interesting nodes and links between the results of independent keyword queries, or independent clusters of nodes.

Our initial versions of the data collector experienced time-outs and service disconnects due to these limits, resulting in incomplete or missing query results. To avoid these problems in later versions, we used a combination of query batching, attribute filtering, and result prefetching. We also executed several simpler queries using the same query text, but searching different node attributes in each query, such as the title, body text, or clinical synopsis attributes. Thus, we ensured that each individual query was completed within the required time limit, and still collected enough data to be useful. Due to these multiple independent queries, we often retrieved redundant node records, which were removed from the final result.

An alternative to working within NCBI's rate limits is to download a copy of NCBI's public databases and simply

[1] http://eutils.ncbi.nlm.nih.gov/.

query the local copy. However, as the NCBI's biomedical databases continue to grow and be augmented with additional semantic information, the feasibility of storing and querying a local copy rapidly diminishes, due to size and synchronization issues.

4 Semantic Substrates

Typical graph visualizations within the biomedical domain use *force-directed* node layouts [19]. However, as stated earlier, force-directed layouts make little or no use of node attributes for node positioning, and thus overlook an important dimension of semantic information. Furthermore, it can be difficult to visually distinguish nodes of different types using force-directed layouts, as demonstrated by the "hairball" visualization in Fig. 1b. Even though node color, shape, or size can be used to differentiate node attributes (in Fig. 1b, color is used), the emphasis on placement using links causes a cluttered and confusing display, even for some small to moderately sized networks. This clutter extends to the network's links in that it is difficult to follow links from source to destination. As a result, force-directed layouts tend to hamper the type of high-dimensional, cross-database exploration that we seek.

Therefore, instead of using link strength (as force-directed layouts do), we position nodes within *semantic substrates*. A semantic substrate consists of a collection of nonoverlapping regions within which nodes are placed and positioned based on node type and other node attributes. To create a semantic substrate, users create a set of regions, and select node attributes and values that determine into which region each node is placed. For example, a natural way to segregate nodes into regions would be to place all PubMed nodes in a single region, all Gene nodes in another region, and so on for each database under consideration. Next, for each substrate region, users select additional node attributes and values to determine how nodes are positioned within the region, such as positioning PubMed nodes based on their publication date, with older publication dates in the left portion of the region, and newer dates to the right.

Rather than relying on links between nodes, semantic substrates provide a consistent node layout, mostly independent of link data. While this may result in more link overlaps, semantic substrates preserve relationships among nodes of the same type. Thus, if users already have ideas or expectations of the types of patterns in their data (as is usually the case with biomedical data), placing nodes based on known attribute values provides a useful visual grounding for further data exploration. For example, the PubMed region layout mentioned above, where PubMed nodes are positioned by publication date, allows users to quickly find the most recent articles covering their topic of interest, rather than having to hunt for them on the screen. As another example, Gene nodes might be positioned within a gridded substrate region according to their genetic locus, with the X-position corresponding to the gene's chromosome, and the Y-position corresponding to the gene's chromosome band.

Semantic substrates are also useful for cross-database exploration because they provide a natural way to group nodes of the same type together. Each database can be represented in a substrate by its own region (e.g., having separate regions for PubMed, Gene, Protein, OMIM, and Taxonomy). As a result, it is easy to distinguish intradatabase links from interdatabase links by visual inspection. Segregating nodes into distinct regions also simplifies the user interaction necessary to filter nodes and links to a selected subset of interest, which is important when visualizing large databases with many relationships between nodes. Links can be identified using the attribute values of the nodes that they connect, which can be determined easily using the nodes' positions, and can be filtered based on both node and link attribute values. For example, users might find a cluster of interesting links between certain PubMed nodes and Gene nodes, where the PubMed nodes' publication date was after 1980 and the Gene locus was on chromosomes 12 and 13.

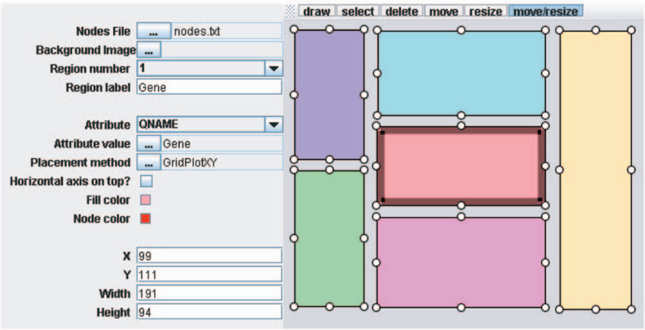
An additional benefit of using semantic substrates is that they provide a natural and powerful way of creating multiple views of the same data set. To do so, users can simply create another semantic substrate by choosing a different region layout, or selecting different node attributes and values to use for region placement and positioning. Having multiple views of the same data is especially useful for visualizing the large, high-dimensional data sets used by the biomedical community, where new and interesting visualizations can be obtained by using a different subset of attributes and values. These different visualizations can afford different insights into the data set under consideration. We used the implementation of semantic substrates called *Network Visualization by Semantic Substrates*.

We now provide an overview of the semantic substrate design process, and the visualization controls available in NVSS.

4.1 Designing a Semantic Substrate

Designing a substrate in NVSS amounts to deciding the number of substrate regions, their positions on the display, and which attributes to use for node placement into and within regions. In general, the process of designing satisfying and useful substrates is an iterative procedure. Often, it is hard to tell how useful a given substrate will be for exploration prior to loading and exploring the data. In addition, the node placement method (i.e., along the regions' X-axis, Y-axis, or both) may affect the visualization's usefulness.

Fortunately, NVSS simplifies the creation of multiple semantic substrates using a built-in substrate designer, fully described by Aris and Shneiderman [2] and shown in Fig. 4. Users draw substrate regions in the right pane of the designer, and then set properties for each region in the left pane. For each region, the most important settings are those which determine the nodes that will be placed in the region, set using the "Attribute" and "Attribute value" fields. In Fig. 4, the user has created six regions, corresponding to nodes of type Gene (the central selected region), Homologene, OMIM, PubMed, Protein, and Taxonomy. For each region, the node positioning within the region is further set using the "Placement method" field, which opens another dialog box with node positioning options. In the figure, for the central Gene region, the "GridPlot XY" method was chosen, with the gene chromosome number used along the



Nodes File ... nodes.txt
Background Image ...
Region number 1
Region label Gene
Attribute QNAME
Attribute value ... Gene
Placement method ... GridPlotXY
Horizontal axis on top? ☐
Fill color 🟥
Node color 🟥
X 99
Y 111
Width 191
Height 94

Fig. 4. The NVSS substrate designer. Users draw regions in the right pane. Each region's node grouping and display properties can be set in the left pane.

X-axis and the chromosome band used along the Y-axis. Apart from node placement and positioning, various display properties for each region can be set, including region labels, node colors, and region fill colors.

To illustrate the substrate design process and show the dramatic difference when using different node placement and positioning attributes, we now describe the process we followed to create a substrate region containing PubMed nodes. These nodes have several attributes which could be useful for node placement, such as authors, publication dates, and publication types. They were also of particular interest and a challenge to visualize due to the relatively large number of PubMed results in our queries and their somewhat skewed distribution of publication dates.

Fig. 5 contains three variants of a substrate region containing the same data, namely PubMed nodes from a query about "cervical cancer." Fig. 5a is our initial visualization of these nodes, using the publication year attribute for layout along the X-axis and with uniform attribute binning. As can be seen, the right portion of the region is overcrowded with nodes, indicating that most documents in the query result were published within the past 10 years. Another problem with the initial layout, of which we were not aware before visualizing the data, is the large gap in PubMed results between 1948 and 1980. This layout wastes screen space and apart from distinguishing the 1948 publication, provides no useful information about the visualized data.

To remedy these problems, we used a different layout, shown in Fig. 5b. In particular, we used the Y-axis for node placement and used custom (i.e., nonuniform) bin sizes to group nodes. This layout causes nodes to be spaced more evenly, and allow users to more easily distinguish individual nodes, which is important for useful data exploration especially in combination with link visualizations. We also size-coded each node to indicate the node's indegree, to impart some measure of the node's importance to the query as a whole. Note that choosing appropriate custom bin sizes requires prior knowledge about the distribution of attribute values, so proper bin lengths can best be set after an initial visualization. These difficulties can also be somewhat mitigated by integrating additional statistical displays, such as attribute value histograms, into the substrate designer, as well as incorporating scrolling or zooming features within the visualization itself to view compact regions more closely. We plan to extend NVSS to address some of these limitations.

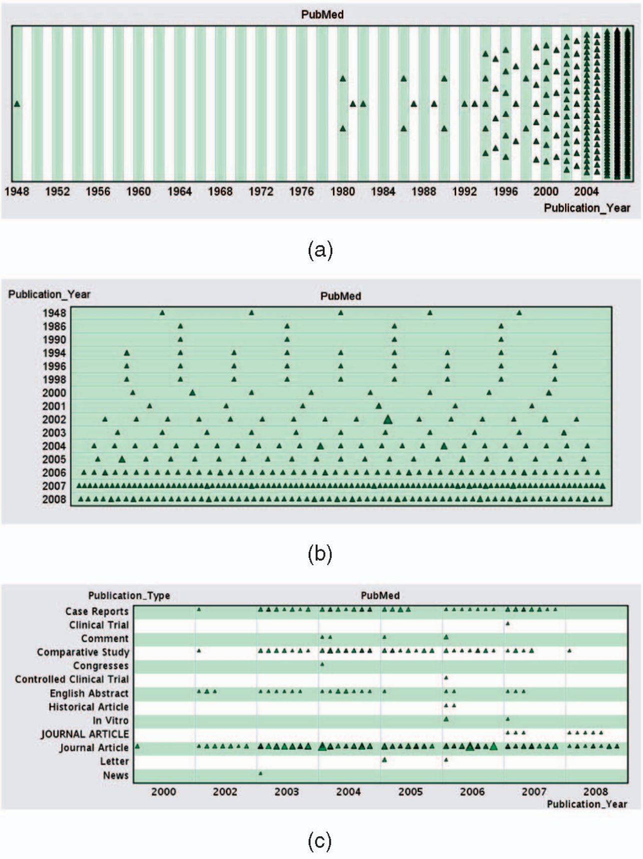


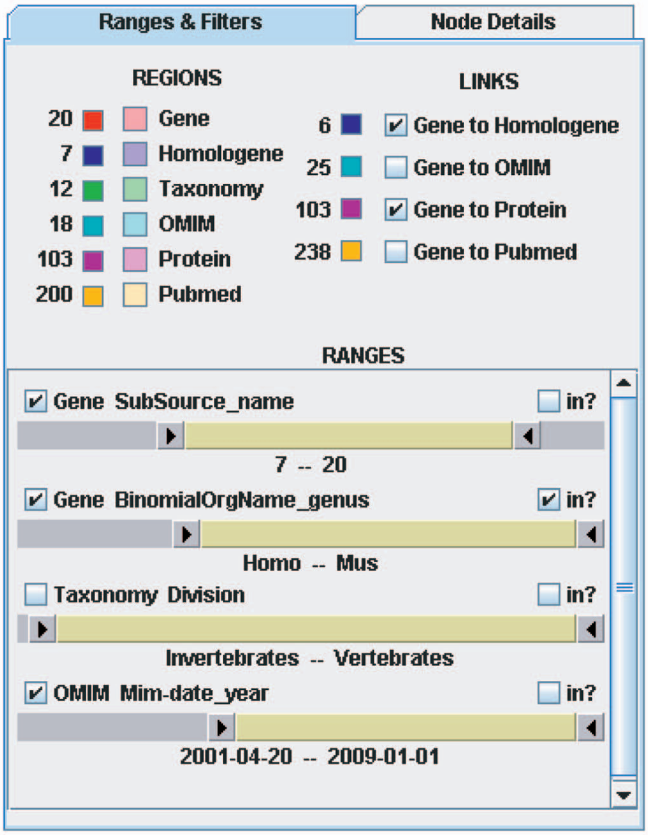
Fig. 5. Three variants of a region with PubMed nodes using different node positioning attributes: (a) Publication year along X-axis with uniform binning. (b) Publication year along Y-axis with custom binning and size coding for node indegree. (c) Publication year along X-axis and publication type along Y-axis.

For a third example layout, shown in Fig. 5c, we created a 2D layout using two attributes: publication year along the X-axis, and publication type along the Y-axis. This layout allows users to find interesting groups of publications by both type and year simultaneously, and provides a quick overview of the types of publications relevant to the query of interest. It also demonstrates how using a different semantic substrate can provide a different means of exploring the same data.

4.2 Visualization Controls

After designing substrates, users proceed to visualize their data using NVSS's visualization module, which has a variety of additional controls. Here, we describe these controls; we will provide complete visualization examples in subsequent sections. Fig. 6 shows the control panel of NVSS's visualization module. In the top portion, users can customize node, region, and link colors. The numbers next to each region and link type represent the number of nodes in regions and links between regions, respectively. In addition, users can control the visibility of links using the link check boxes. In the figure, the user has chosen to show the six links from Gene nodes to Homologene nodes, and the 103 links from Gene nodes to Protein nodes.

The lower portion contains additional link filters based on source and destination node attributes. These filters are



REGIONS

20	🟥	Gene
7	🟦	Homologene
12	🟩	Taxonomy
18	🟦	OMIM
103	🟪	Protein
200	🟧	Pubmed

LINKS

6	🟦	☑ Gene to Homologene
25	🟦	□ Gene to OMIM
103	🟪	☑ Gene to Protein
238	🟧	□ Gene to Pubmed

RANGES

☑ Gene SubSource_name □ in?
7 -- 20

☑ Gene BinomialOrgName_genus ☑ in?
Homo -- Mus

□ Taxonomy Division □ in?
Invertebrates -- Vertebrates

☑ OMIM Mim-date_year □ in?
2001-04-20 -- 2009-01-01

Fig. 6. Visualization and filtering controls available in NVSS. Above, users select colors for substrate regions, nodes, and links. Check boxes also allow users to selectively display subsets of links based on source and destination region. Below, users perform additional link filtering based on additional source and destination node attributes, using checkboxes to enable filters, and two-way sliders to select attribute value ranges.

vital when exploring very large databases with many node relationships, as are often present in the biomedical domain, to ensure a meaningful and useful visualization. In the figure, three attribute value filters are activated, including links from those Gene nodes with chromosome number between 7-20 and genus Homo or Mus, as well as a filter for OMIM nodes with modification dates between 2001 and 2009. In addition to the above controls, NVSS provides node details when nodes are selected in the display, accessible by clicking the "Node Details" tab at the top. Furthermore, by clicking on a node in the visualization, users can open a web browser to display a URL associated with the node, which for our data sets and queries, was the NCBI webpage corresponding to that node. This feature was especially important to our users, as our data collector was unable to retrieve all the attributes in which biomedical researchers were interested due to database access limitations.

5 Sample Visualizations

Based on our interviews with domain experts (see Section 6), we created several visualizations of sample queries that might interest typical biomedical researchers, based around diseases being actively researched at NCBI, the National Institutes of Health (NIH), and other biomedical centers. Our queries encompassed six NCBI Entrez databases, namely PubMed, OMIM, Gene, Protein, Homologene, and Taxonomy.

In collecting data from these databases, we found that the set of attributes available through the Entrez system was rather limited in size and breadth. As a result, the node placement and positioning attributes we used for demonstration purposes would be of somewhat limited use for the highly specific queries of biomedical research. NCBI's internal databases, hidden from the web, contain a much richer set of attributes, and it is these attributes that would make for even more interesting visualizations using semantic substrates, which thrive on rich attribute spaces. See Section 6 for a description of some of these attributes and their potential use.

The queries and their visualizations are detailed below.

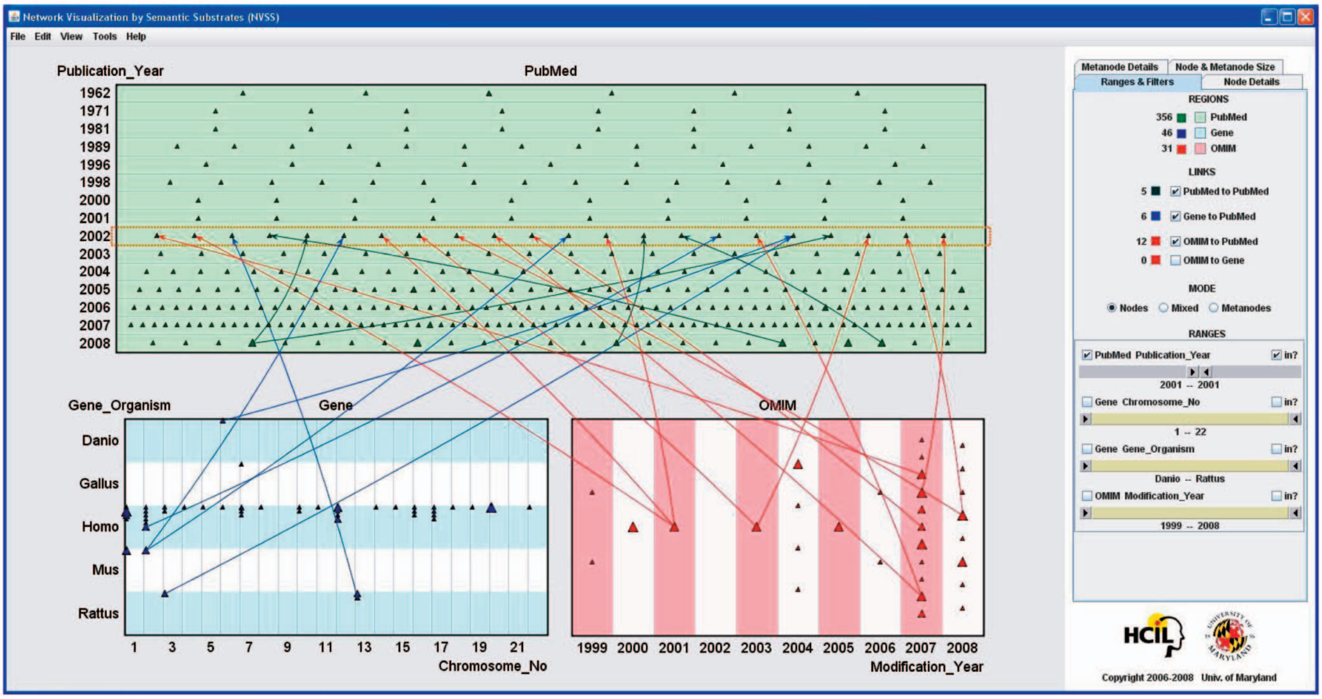
5.1 Hypertension

Our first query was of the general form: "What are the most significant publications, genes, and diseases related to hypertension?" To execute this query, we performed a keyword search in the OMIM database for "hypertension," and retrieved links from the resulting set of OMIM nodes to the Gene and PubMed databases. We also retrieved links from the Gene nodes to PubMed nodes, as well as some similarity links between PubMed records. For the hypertension query, we collected a total of 433 nodes, including 357 PubMed records, 45 Gene records, and 31 OMIM entries, in addition to 440 links.

Fig. 7 shows one visualization of the query results, using a substrate with separate regions for each database. To position nodes within regions, we ordered PubMed and OMIM nodes by publication year and modification year, respectively, while for Gene nodes, we used a 2D layout using each node's Genus and Chromosome Number attributes. In the figure, the links have been filtered to only those PubMed documents published in 2002, using NVSS's slider bar filters. Notice that the collection of PubMed documents within that time range are linked from all three databases, and most are linked from a single source only. The figure exemplifies the need for visual cross-database exploration, in that this phenomenon likely resulted from our not knowing the correct keywords to use to return all relevant results. This is a typical problem with strict keyword searches, even those performed by domain experts. To retrieve the equivalent set of results in a traditional textual exploration interface would require repeating the query several times in multiple databases, and manually merge the results. Using semantic substrates, we can easily perform these cross-database searches and visually display query results in an intuitive and comprehensible manner.

5.2 Mental Disorders

Our next query was gene-centric and involved finding genes implicated in several mental disorders, namely anxiety, depression, addiction, and schizophrenia, as well as publications related to these genes. For this query, we used the same databases as used in our hypertension query (i.e., OMIM, Gene, and PubMed). However, we performed four separate keyword searches in OMIM, and retrieved links from each separate search to the Gene database. We then retrieved links to PubMed from the Gene result nodes.



REGIONS

356 🟩 PubMed
46 🟦 Gene
31 🟥 OMIM

LINKS

5 ☑ PubMed to PubMed
6 ☑ Gene to PubMed
12 ☑ OMIM to PubMed
0 □ OMIM to Gene

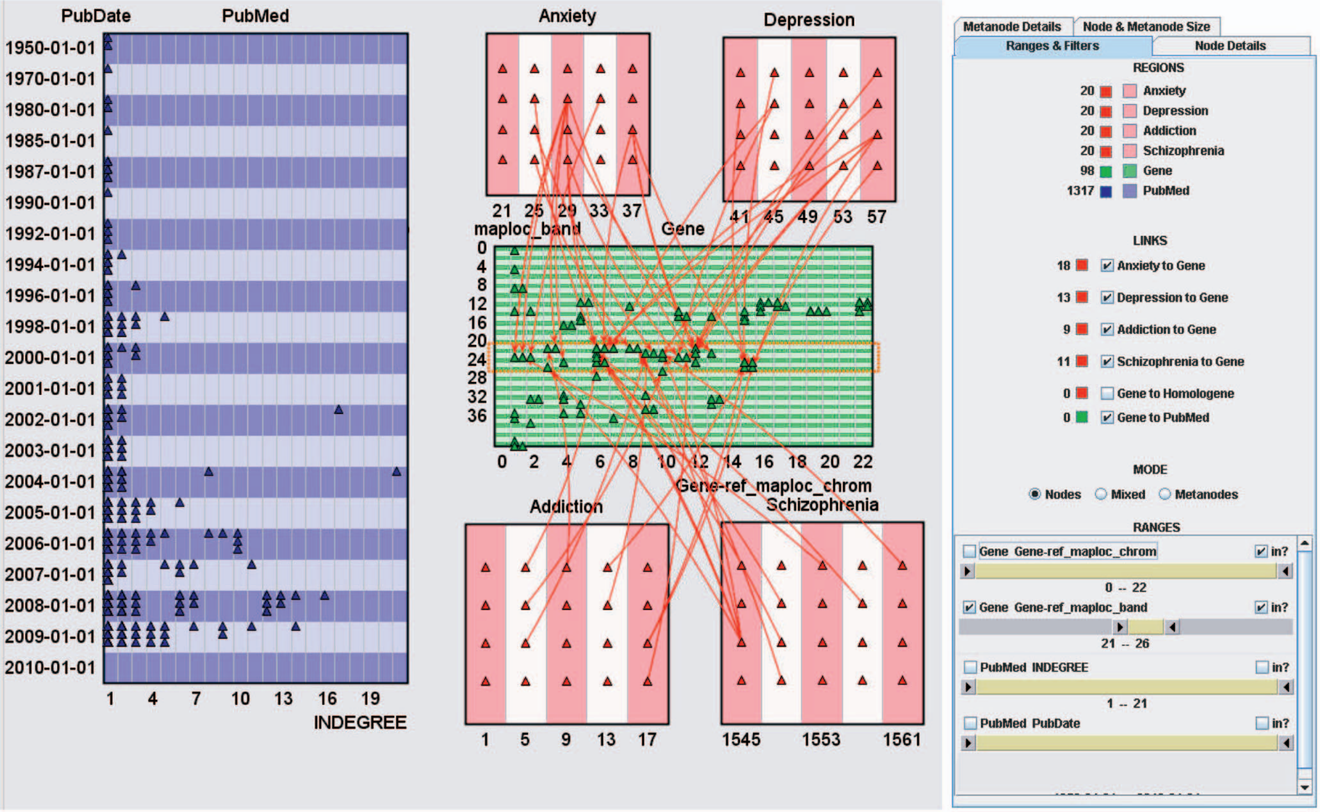
MODE

● Nodes ○ Mixed ○ Metanodes

RANGES

☑ PubMed Publication_Year ☑ in?
2001 -- 2001
□ Gene Chromosome_No □ in?
1 -- 22
□ Gene Gene_Organism □ in?
Danio -- Rattus
□ OMIM Modification_Year □ in?
1999 -- 2008

Fig. 7. Results of a cross-database query about hypertension. Gene-PubMed, OMIM-PubMed, and PubMed-PubMed links are shown where the link's target publication was published in 2002. Notice that links to the relevant publications are scattered across multiple sources and would be difficult to find using nonvisual cross-database exploration methods.



REGIONS

20 🟥 Anxiety
20 🟥 Depression
20 🟥 Addiction
20 🟥 Schizophrenia
98 🟩 Gene
1317 🟦 PubMed

LINKS

18 ☑ Anxiety to Gene
13 ☑ Depression to Gene
9 ☑ Addiction to Gene
11 ☑ Schizophrenia to Gene
0 □ Gene to Homologene
0 ☑ Gene to PubMed

MODE

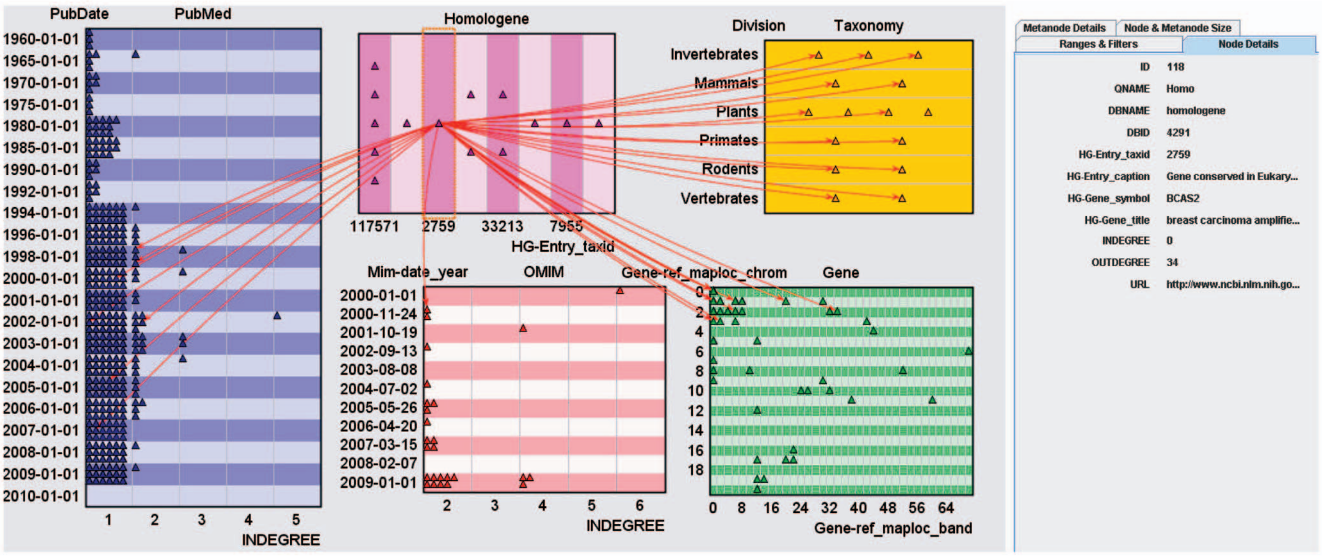
● Nodes ○ Mixed ○ Metanodes

RANGES

□ Gene Gene-ref_maploc_chrom ☑ in?
0 -- 22
☑ Gene Gene-ref_maploc_band ☑ in?
21 -- 26
□ PubMed INDEGREE □ in?
1 -- 21
□ PubMed PubDate □ in?

Fig. 8. Results from a query about genes implicated in several mental disorders. Multiple OMIM keyword searches allow a visual query intersection of Gene nodes. Also, link statistics from Gene nodes to PubMed nodes are used to find the most prevalent publications related to these genes.

Fig. 8 shows our visualization. Even though we used the same databases as before, our substrate is substantially different, demonstrating the ease of creating multiple views of the same data. The four OMIM keyword search results are placed into different regions, labeled Anxiety, Depression, Addiction, and Schizophrenia, respectively, which provide a simple means of visually distinguishing OMIM nodes from the different keyword searches. Gene and



ID	118
QNAME	Homo
DBNAME	homologene
DBID	4291
HG-Entry_taxid	2759
HG-Entry_caption	Gene conserved in Eukary...
HG-Gene_symbol	BCAS2
HG-Gene_title	breast carcinoma amplifie...
INDEGREE	0
OUTDEGREE	34
URL	http://www.ncbi.nlm.nih.go...

Fig. 9. Results from our query on breast carcinoma. The highlighted Homologene entry, corresponding to the BCAS2 conserved gene, contains links to a wide variety of Gene and Taxonomy nodes, indicating its importance to breast carcinoma.

PubMed nodes again are given their own substrate regions. For OMIM node positioning, we used database identifiers, while Gene nodes are positioned using chromosome number (Y-axis) and chromosome band (X-axis). PubMed nodes are positioned using publication date (Y-axis) and node indegree (X-axis), which allow visual determination of PubMed nodes' recency of publication and importance in terms of citation by Gene nodes. We collected a total of 80 OMIM nodes (20 per keyword search), 98 Gene nodes, and 1,317 PubMed nodes, and a total of 2,034 links.

We can see that several clusters of genes collected through links from OMIM can be found in the Gene region, which maybe difficult to find using nonvisual search methods. Also, when applying filtering to show links from OMIM nodes to centrosomal Gene nodes (i.e., those genes with middle band numbers), we observe that these genes have links from a large number of OMIM nodes related to all four queried mental disorders. This may indicate that the genes under consideration have strong ties to all these disorders. Upon examining the PubMed nodes, obtained through links from the Gene nodes (but not shown), we see that research into genes related to the queried mental disorders has a rich history. In addition, the PubMed nodes with large indegree (i.e., those nodes with a large influence on the Gene nodes as measured by number of links from Gene nodes) readily stand out, allowing quick determination of the most relevant publications related to the genes of interest. As this example shows, NVSS's interactive filters allow the exploration of data in a variety of useful ways.

5.3 Breast Carcinoma

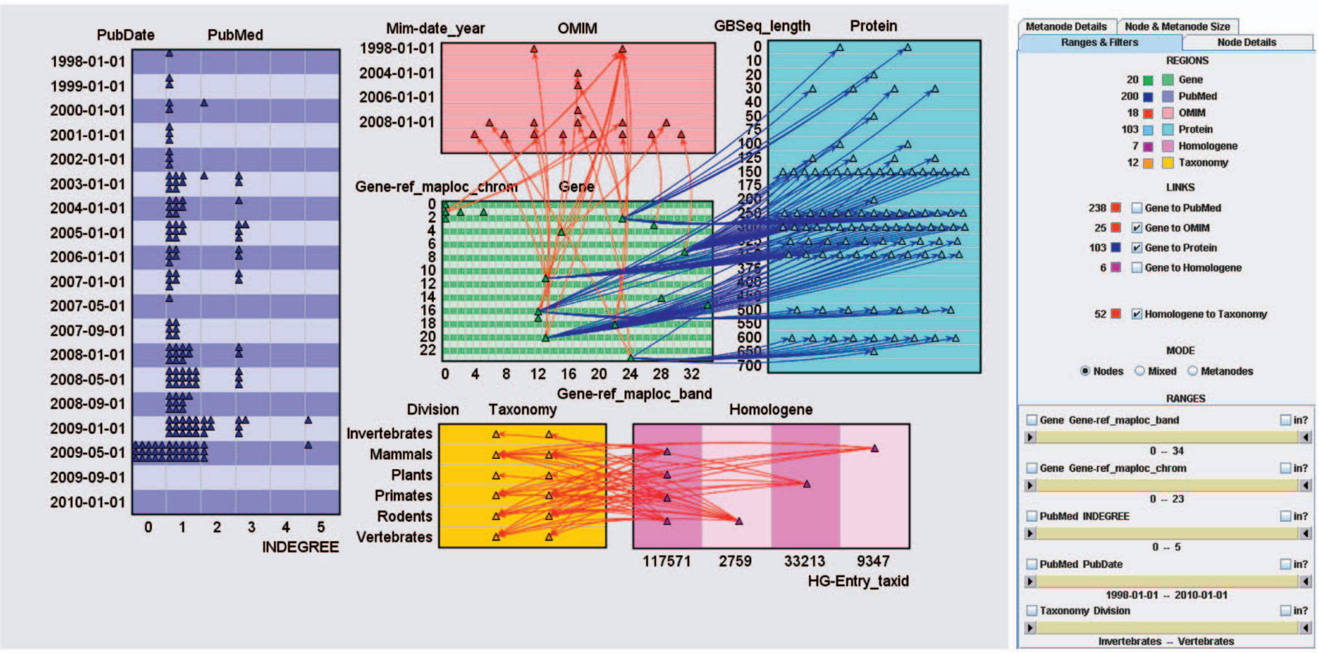
For our third query, we searched for cross-species genetic information and publications related to breast carcinoma. An initial keyword search for "breast carcinoma" was performed in the Homologene database to retrieve cross-species information. Links to the Gene, OMIM, Taxonomy, and PubMed databases were then retrieved. We also retrieved additional links from the returned OMIM nodes to PubMed nodes. In total, we retrieved 14 Homologene, 15 Taxonomy, 26 OMIM, 129 Gene, and 587 PubMed nodes, along with 908 links.

Fig. 9 shows our query results in a substrate with five regions, with each region corresponding to a different database accessed in our query. For Homologene node positioning, we chose a taxonomy ID associated with the node, which indicated its primary species association. Of course, each node had several such associations, as evidenced by the links from each Homologene node to multiple Taxonomy nodes. We divided the Taxonomy region based on the species division as specified in the Taxonomy database. For the OMIM region, both the modification date (Y-axis) and indegree (X-axis) were used for node positioning. Finally, for the PubMed and Gene regions, we used the same node layouts as those in our previous query about mental disorders.

By interactively filtering links, we quickly found a Homologene node with links to many Taxonomy nodes (highlighted in Fig. 9). In other words, we found a gene with many cross-species links that was especially relevant to breast carcinoma. The node's details are shown in NVSS's right panel, which indicates the underlying gene symbol as BCAS2 and corresponding title "breast carcinoma amplified sequence 2." Also note that the Homologene node in question has links to a tight cluster of Gene nodes, which may indicate the disease's approximate genetic locus. As before, finding relevant OMIM entries and PubMed documents becomes simple when using the indegree for node layout in their respective regions. All these visual indications can allow domain experts to find useful starting points for more in-depth exploration.

5.4 Obesity

Our final example query for relevant entries about "obesity" also used the Protein database to find relevant proteins, in addition to the previously used PubMed, Gene, OMIM, Homologene, and Taxonomy databases. We began with two keyword searches for "obesity" in the Gene and PubMed databases. Next, we retrieved links



REGIONS

20	Gene
200	PubMed
18	OMIM
103	Protein
7	Homologene
12	Taxonomy

LINKS

238	□ Gene to PubMed
25	☑ Gene to OMIM
103	☑ Gene to Protein
6	□ Gene to Homologene
52	☑ Homologene to Taxonomy

MODE

● Nodes ○ Mixed ○ Metanodes

RANGES

□ Gene Gene-ref_maploc_band □ in?
0 -- 34
□ Gene Gene-ref_maploc_chrom □ in?
0 -- 23
□ PubMed INDEGREE □ in?
0 -- 5
□ PubMed PubDate □ in?
1998-01-01 -- 2010-01-01
□ Taxonomy Division □ in?
Invertebrates -- Vertebrates

Fig. 10. A substrate showing results from our obesity query. Several interesting patterns emerge, such as several Gene nodes with links to multiple OMIM nodes or multiple Protein node clusters.

from the Gene node results to nodes in the OMIM, Homologene, Protein, and PubMed databases. Finally, we retrieved links from the Homologene results to Taxonomy database nodes. A total of 200 PubMed, 27 Gene, 18 OMIM, 7 Homologene, 12 Taxonomy, and 103 Protein nodes were retrieved, and 425 links.

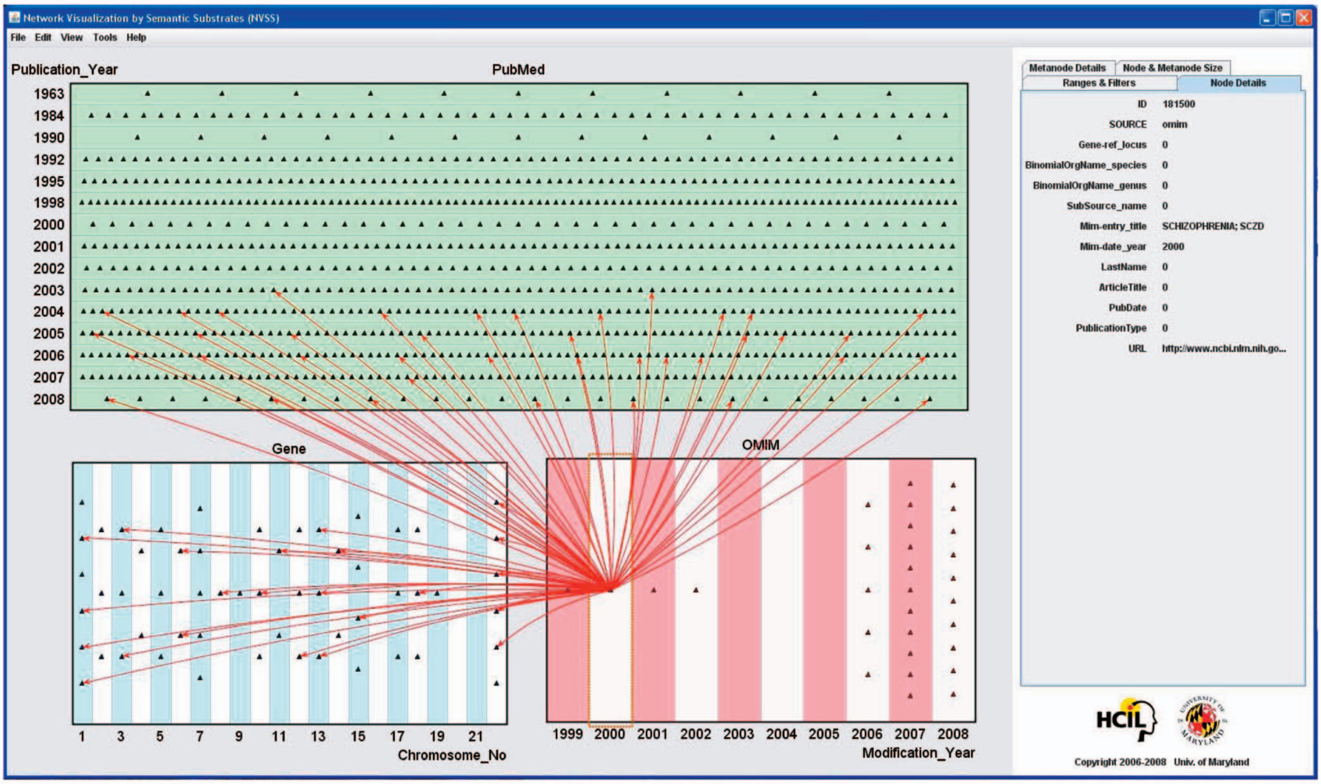
Fig. 10 contains our obesity query visualization. For node positioning in the Protein region, we used the corresponding protein's length, which served as a rough clustering measure for the protein nodes. For the remaining regions we used the same attributes as in the previous visualization. In the figure, we filtered the links to show only those from Gene nodes to Protein and OMIM nodes, as well as links from Homologene nodes to Taxonomy nodes. In doing so, we observe that several Gene nodes have links to multiple OMIM nodes, indicating their possible connection with several diseases or medical conditions related to obesity. In addition, some Gene nodes have links to multiple clusters of Protein nodes, which may indicate their importance to the query result. With different node filtering, domain experts can explore the query results to discover additional details useful in their research.

6 Expert Evaluation

To judge the effectiveness of our visualization methods using semantic substrates, we met with 10 bioinformatics specialists from the National Library of Medicine. These researchers have expertise in a variety of areas, including biomedical informatics, biomedical ontologies, machine learning, and text analysis. Most also hold medical degrees and the PhDs in medical informatics and computer science, and have on average 15 years of experience in their respective fields.

These researchers mainly used the PubMed, Gene, and OMIM databases for their work, in addition to NCBI's various other databases. In general, they were dissatisfied with the current state of affairs in bioinformatics visualizations, especially related to visualization of manually or automatically extracted semantic relationships among PubMed documents, as well as the hierarchical relationships of the MeSH and GO ontologies. For example, text mining methods on a large collection of PubMed documents, Gene records, and OMIM articles might yield relationships such as "gene X is correlated with disease Y." They had tried using off-the-shelf tools such as Prefuse [25] and GraphViz [22] to visualize these relationships, but found them to be inadequate for exploratory purposes, mainly due to "insufficient flexibility" of the data visualization and their "limited navigation paradigms." The sheer number of semantic relationships extracted from PubMed documents—in the hundreds of thousands—was also a limiting factor, as most visualizations lost their effectiveness when the number of visualized relationships exceeded the hundreds. Also, these tools generally did not allow integration of data from multiple sources, which severely limited their utility. Unlike the existing visualizations, semantic substrates' powerful filtering capabilities are better-suited for showing interesting subsets of large, complicated networks.

We arranged a 1.5 hour combined presentation and focus group discussion with the team of experts. A half-hour was dedicated to a presentation of our exploration methodology using semantic substrates, after which we asked for comments and feedback from the experts for the remaining hour. We asked the experts how biologists seeking information from the NLM or NCBI databases would normally explore their vast collections of data. They commented that detailed literature and topical surveys are normally carried out by the NLM's expert librarians, who maintain their own private indexing systems, separate from the public interfaces available through the Internet. They



ID	181500
SOURCE	omim
Gene-ref_locus	0
BinomialOrgName_species	0
BinomialOrgName_genus	0
SubSource_name	0
Mim-entry_title	SCHIZOPHRENIA; SCZD
Mim-date_year	2000
LastName	0
ArticleTitle	0
PubDate	0
PublicationType	0
URL	http://www.ncbi.nlm.nih.go...

Fig. 11. A visualization of results from the query: "What centrosomal genes are implicated in diseases of brain development?" Links from an OMIM record about schizophrenia are shown to relevant genes and PubMed documents.

also mentioned that typical keyword searches using the NCBI's web interfaces would not return an exhaustive collection of relevant literature and information, as query results are heavily dependent on the exact terms used, and do not adequately take synonyms and other relationships into account. They were excited that our approach augmented an initial keyword search with link information that effectively expands the results of a given query in possibly interesting ways. In other words, they believed that it is not strictly necessary to know all the synonyms or related terms for a given keyword query, as these synonyms are implicit in the link relationships found among the query results. They further commented that semantic substrates offer a "useful visual metaphor" for exploring ever-expanding collections of semantic relationships in a scalable way. The researchers also mentioned that "pulling in multiple databases for cross-searching" was the correct way to explore large collections of biomedical data.

To evaluate our methods in the context of specific technical queries, we used our data collector to prepare sample query results involving the PubMed, Gene, and OMIM databases. We retrieved data based on several queries from the TREC 2007 Genomics Protocol [27], three of which were as follows:

1. What centrosomal genes are implicated in diseases of brain development?
2. What is the genetic component of alcoholism?
3. What mutations in apolipoprotein genes are associated with disease?

We visualized the query results in NVSS, using one region for each of the PubMed, Gene, and OMIM data. Fig. 11 is one such visualization of the brain development query. The figure shows one particular OMIM node, corresponding to schizophrenia, and all the Gene and PubMed records that it references. The Gene records are organized by chromosome number, PubMed documents are ordered by year of publication, and OMIM entries by the last date of modification.

The NLM team commented that the referenced genes were likely implicated in or related to schizophrenia. They liked that they were able to see, at a glance, what the most important genes and documents related to schizophrenia were. They also suggested that using more attributes of each node type would make it easier to answer the query. In particular, "centrosomal genes" refer to those genes with a central physical location on their respective chromosomes—in other words, with a middle band number. As evidenced by Figs. 8, 9, and 10, gene chromosome and band numbers serve as natural and useful attributes for node positioning within semantic substrate regions. To improve the visualization of this query's results, we could position Gene nodes using chromosome number and chromosome band number, as we had done earlier. This modified layout would allow users to quickly find centrosomal genes at a glance, by examining the nodes' spatial positions within the region.

To improve the PubMed and OMIM regions, the researchers suggested additional node attributes to use. For PubMed nodes, genotypic and phenotypic associations might make for interesting visual classifications. Also, for OMIM nodes, rather than using modification date for node positioning, they suggested using the class of disease connected with the OMIM entry. The researchers commented that this richer set of node attributes would greatly enhance the visualization and make it immediately useful for answering a variety of queries. Unfortunately, these attributes, while present in internal NCBI databases, were not accessible through the NCBI's web interfaces. However, if available, these attributes could be easily integrated into semantic substrate designs and would be useful for exploring query results.

The NLM researchers offered several suggestions for making cross-database exploration in NVSS more dynamic. In particular, they wanted ways to refine their initial query based on additional keywords found in the set of results, or selectively filter or expand subsets of the graph. Also, the NLM team suggested that it would be useful to dynamically add more substrate regions, and reposition regions if the current substrate layout was not found to be useful. We plan to integrate these improvements into future versions of NVSS.

7 Conclusion

Parting from textual query result lists like those at NCBI's website, semantic substrates offer a novel way to browse and explore biomedical data across multiple databases. This browsing would be further enhanced by incorporating dynamic query retrieval of nodes and links and the subsequent visualization of results within appropriate substrate regions. Furthermore, new methods would have to be developed for visualizing the number of results, and determining and displaying the most interesting or relevant results. Navigating through the various sets of query results, in a manner analogous to a web browser's forward and back buttons, also poses a challenge. One way to incorporate query navigation might be to navigate using a tree, in the same way that our data collector uses a query tree. However, instead of nodes corresponding to databases, nodes of this navigation tree would correspond to substrates in the navigation history, similar to the history mechanism used for VisPad [43]. When a node is clicked, the previous exploration state corresponding to that node would be loaded into the visualization.

Also, as many biomedical data sets involve ontological or hierarchical relationships (e.g., Gene Ontology, MeSH terms, and Taxonomic/Phylogenetic trees), our visualizations could be enhanced by incorporating additional visualization methods within the semantic substrate framework. In particular, the regions within semantic substrates could use a treemap [8], [18] to hierarchically organize nodes. For example, a visualization involving genes of multiple species might incorporate a treemap subdividing the region space hierarchically according to the taxonomy of genes in the data set. Node positioning within each treemap cell could be customizable depending on users' preferences. Some of the many existing alternative network visualization algorithms (e.g., layered [46]) could be incorporated as well. Another useful feature would be a means of displaying or interacting with the ontological information associated with each node, if present.

In addition, while our current visualization favors exploration of individual nodes, such as PubMed documents or genes, more sophisticated link filtering and exploration may improve our visualization tool. In particular, NVSS currently supports link filtering based on source and destination node attributes, but would benefit from additional filtering options based on other link attributes. Also, NVSS's handling of multiple link filters is currently limited to "AND" rules (e.g., show links with source in region X and destination in region Y), but does not allow "OR" rules (e.g., show links with source in region X and destination in either Y or Z). Adding better support for link filtering and manipulation would allow more expressive queries for user exploration using semantic substrates.

As the amount of semantically tagged biomedical data continues to grow, we believe that semantically relevant visualizations like semantic substrates will have increasingly important roles in exploring and understanding biomedical databases in the near future.

Acknowledgments

The authors wish to thank Sameer Antani, Olivier Bodenreider, Bruce Bray, Marcelo Fiszman, Dina Demner-Fushman, Michael Galperin, Adam Lee, Jimmy Lin, Adam Phillippy, Louiqa Raschid, Tom Rindflesch, Charles Sneiderman, George Thoma, and the anonymous reviewers for their invaluable comments and assistance in evaluating our methods. This work was supported in part by the US National Science Foundation (NSF) under grants EIA-00-91474, CCF-05-15241, and IIS-0713501, as well as the Office of Policy Development & Research of the Department of Housing and Urban Development, Microsoft Research, and NVIDIA.

References

[1] A.T. Adai, S.V. Date, S. Wieland, and E.M. Marcotte, "LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks," *Molecular Biology,* vol. 340, no. 1, pp. 179-190, June 2004.

[2] A. Aris and B. Shneiderman, "Designing Semantic Substrates for Visual Network Exploration," *Information Visualization,* vol. 6, no. 4, pp. 281-300, Nov. 2007.

[3] A. Aris, B. Shneiderman, V. Qazvinian, and D. Radev, "Visual Overviews for Discovering Key Papers and Influences across Research Fronts," *J. Am. Soc. for Information Science and Technology,* vol. 60, no. 11, pp. 2219-2228, Nov. 2009.

[4] M. Baitaluk, M. Sedova, A. Ray, and A. Gupta, "BiologicalNetworks: Visualization and Analysis Tool for Systems Biology," *Nucleic Acids Research,* vol. 34, pp. W466-W471, July 2006.

[5] V. Batagelj and A. Mrvar, "Pajek: A Program for Large Network Analysis," *Connections,* vol. 21, no. 2, pp. 47-58, 1998.

[6] M.Y. Becker and I. Rojas, "A Graph Layout Algorithm for Drawing Metabolic Pathways," *Bioinformatics,* vol. 17, no. 5, pp. 461-467, May 2001.

[7] R.A. Becker, S.G. Eick, and A.R. Wilks, "Visualizing Network Data," *IEEE Trans. Visualization and Computer Graphics,* vol. 1, no. 1, pp. 16-28, Mar. 1995.

[8] B.B. Bederson, B. Shneiderman, and M. Wattenberg, "Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies," *ACM Trans. Graphics,* vol. 21, no. 4, pp. 833-854, Oct. 2002.

[9] O. Bodenreider and A.T. McCray, "Exploring Semantic Groups through Visual Approaches," *J. Biomedical Informatics,* vol. 36, no. 6, pp. 414-432, Dec. 2003.

[10] K.W. Boyack, K. Mane, and K. Börner, "Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research," *Proc. Eighth Int'l Conf. Information Visualization (IV '04),* pp. 965-971, Jul. 2004.

[11] K.W. Boyack, B.N. Wylie, and G.S. Davidson, "Domain Visualization Using VxInsight for Science and Technology Management," *J. Am. Soc. for Information Science and Technology,* vol. 53, no. 9, pp. 764-774, Aug. 2002.

[12] U. Brandes and D. Wagner, "Visone: Analysis and Visualization of Social Networks," *Graph Drawing Software,* M. Jünger and P. Mutzel, eds., pp. 321-340, Springer-Verlag, 2004.

[13] B.-J. Breitkreutz, C. Stark, and M. Tyers, "Osprey: A Network Visualization System," *Genome Biology,* vol. 4, no. 3, article R22, Feb. 2003.

[14] C. Chen, "CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature," *J. Am. Soc. for Information Science and Technology,* vol. 57, no. 3, pp. 359-377, Feb. 2006.

[15] D.J. de Solla Price, "Networks of Scientific Papers," *Science,* vol. 149, no. 3683, pp. 510-515, July 1965.

[16] G. di Battista, P. Eades, R. Tamassia, and I.G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs,* P. Hall, ed., Prentice-Hall, 1999.

[17] P. Eades and Q.-W. Feng, "Multilevel Visualization of Clustered Graphs," *Proc. Symp. Graph Drawing (GD '96),* pp. 101-112, Sept. 1996.

[18] J.-D. Fekete, D. Wang, N. Dang, A. Aris, and C. Plaisant, "Overlaying Graph Links on Treemaps," *Proc. Ninth IEEE Symp. Information Visualization (InfoVis '03) Poster Compendium,* pp. 82-83, Oct. 2003.

[19] T.M.J. Fruchterman and E.M. Reingold, "Graph Drawing by Force-Directed Placement," *Software—Practice and Experience,* vol. 12, no. 11, pp. 1129-1164, Nov. 1991.

[20] D.C.Y. Fung, S.-H. Hong, K. Xu, and D. Hart, "Visualizing the Gene Ontology-Annotated Clusters of Co-Expressed Genes: A Two-Design Study," *Proc. Fifth Int'l Conf. BioMedical Visualization (MEDIVIS '08),* pp. 9-14, Jul. 2008.

[21] P. Gambette and D.H. Huson, "Improved Layout of Phylogenetic Networks," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 5, no. 3, pp. 472-479, July-Sept. 2008.

[22] E.R. Gansner and S.C. North, "An Open Graph Visualization System and Its Applications to Software Engineering," *Software—Practice and Experience,* vol. 30, no. 11, pp. 1203-1233, Sept. 2000.

[23] K.-I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The Human Disease Network," *Proc. Nat'l Academy of Sciences USA,* vol. 104, no. 21, pp. 8685-8690, May 2007.

[24] K. Han, B.-H. Ju, and H. Jung, "WebInterViewer: Visualizing and Analyzing Molecular Interaction Networks," *Nucleic Acids Research,* vol. 32, pp. W89-W95, July 2004.

[25] J. Heer, S.K. Card, and J.A. Landay, "Prefuse: A Toolkit for Interactive Information Visualization," *Proc. Conf. Human Factors in Computing Systems (SIGCHI '05),* pp. 421-430, Apr. 2005.

[26] I. Herman, G. Melançon, and M.S. Marshall, "Graph Visualization and Navigation in Information Visualization: A Survey," *IEEE Trans. Visualization and Computer Graphics,* vol. 6, no. 1, pp. 24-43, Mar. 2000.

[27] W.R. Hersh, A.M. Cohen, L. Ruslen, and P.M. Roberts, "TREC 2007 Genomics Track Overview," *Proc. 16th Text Retrieval Conf. (TREC '07),* Nov. 2007.

[28] J.W.K. Ho, T. Manwaring, S.-H. Hong, U. Roehm, D.C.Y. Fung, K. Xu, T. Kraska, and D. Hart, "PathBank: Web-Based Querying and Visualization of an Integrated Biological Pathway Database," *Proc. Int'l Conf. Computer Graphics, Imaging and Visualisation (CGIV '06),* pp. 84-89, Jul. 2006.

[29] R. Hoffmann and A. Valencia, "Implementing the iHOP Concept for Navigation of Biomedical Literature," *Bioinformatics,* vol. 21, pp. ii252-ii258, 2005.

[30] Z. Hu, J. Mellor, J. Wu, T. Yamada, D.T. Holloway, and C. DeLisi, "VisANT: Data-Integrating Visual Framework for Biological Networks and Modules," *Nucleic Acids Research,* vol. 33, pp. W352-W357, July 2005.

[31] T. Huan, A.Y. Sivachenko, S.H. Harrison, and J.Y. Chen, "ProteoLens: A Visual Analytic Tool for Multi-Scale Database-Driven Biological Network Data Mining," *BMC Bioinformatics,* vol. 9, no. suppl 9, article S5, Aug. 2008.

[32] D.H. Huson, "SplitsTree: Analyzing and Visualizing Evolutionary Data," *Bioinformatics,* vol. 14, no. 1, pp. 68-73, Feb. 1998.

[33] F. Iragne, M. Nikolski, B. Mathieu, D. Auber, and D.J. Sherman, "ProViz: Protein Interaction Visualization and Exploration," *Bioinformatics,* vol. 21, no. 2, pp. 272-274, Jan. 2005.

[34] P.D. Karp and S. Paley, "Automated Drawing of Metabolic Pathways," *Proc. Third Int'l Conf. Bioinformatics and Genome Research,* pp. 225-238, June 1994.

[35] C. Kosak, J. Marks, and S. Shieber, "Automating the Layout of Network Diagrams with Specified Visual Organization," *IEEE Trans. Systems, Man and Cybernetics,* vol. 24, no. 3, pp. 440-454, Mar. 1994.

[36] A.Y. Muhammed, K.-I. Goh, M.E. Cusick, A.-L. Barabási, and M. Vidal, "Drug—Target Network," *Nature Biotechnology,* vol. 25, pp. 1119-1126, Oct. 2007.

[37] B.A. Nardi, S. Whittaker, E. Isaacs, M. Creech, J. Johnson, and J. Hainsworth, "Integrating Communication and Information through ContactMap," *Comm. ACM,* vol. 45, no. 4, pp. 89-95, Apr. 2002.

[38] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway Studio—The Analysis and Navigation of Molecular Networks," *Bioinformatics,* vol. 19, no. 16, pp. 2155-2157, Nov. 2003.

[39] D. Schaffer, Z. Zuo, S. Greenberg, L. Bartram, J. Dill, S. Dubs, and M. Roseman, "Navigating Hierarchically Clustered Networks through Fisheye and Full-Zoom Methods," *ACM Trans. Computer-Human Interaction,* vol. 3, no. 2, pp. 162-188, June 1996.

[40] 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, no. 11, pp. 2498-2504, Nov. 2003.

[41] C.D. Shaw, "Genomic Spring-Synteny Visualization with IMAS," *Proc. Fifth Int'l Conf. BioMedical Visualization (MEDIVIS '08),* pp. 3-8, Jul. 2008.

[42] B. Shneiderman and A. Aris, "Network Visualization by Semantic Substrates," *IEEE Trans. Visualization and Computer Graphics,* vol. 12, no. 5, pp. 733-740, Oct. 2006.

[43] Y.B. Shrinivasan and J.J. van Wijk, "VisPad: Integrating Visualization, Navigation and Synthesis," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST '07),* pp. 209-210, Oct. 2007.

[44] N.R. Smalheiser and D.R. Swanson, "Using ARROWSMITH: A Computer-Assisted Approach to Formulating and Assessing Scientific Hypotheses," *Computer Methods and Programs in Biomedicine,* vol. 57, no. 3, pp. 149-153, Nov. 1998.

[45] M. Suderman and M.T. Hallett, "Tools for Visually Exploring Biological Networks," *Bioinformatics,* vol. 23, no. 20, pp. 2651-2659, Oct. 2007.

[46] K. Sugiyama, S. Tagawa, and M. Toda, "Methods for Visual Understanding of Hierarchical System Structures," *IEEE Trans. Systems, Man and Cybernetics,* vol. 11, no. 2, pp. 109-125, Feb. 1981.

Michael D. Lieberman received the BS degree in computer engineering and the MS degree in computer science from the University of Maryland, College Park, and is currently working toward the PhD degree at Maryland. In 2008, he was awarded the Dean's Fellowship Award for research excellence. His main research interests include geographic information extraction, data mining, and spatial databases. He is a student member of the IEEE and IEEE Computer Society.

Sima Taheri received the BSc and MSc degrees in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2003 and 2005, respectively. In 2005, she moved to Singapore to pursue research studies on biomedical image analysis. She is currently working toward the PhD degree in the Department of Computer Science at the University of Maryland, College Park. Her research interests include computer vision, multimedia analysis, and pattern recognition. She is a student member of the IEEE.

Huimin Guo received the BS and MS degrees in computer science from Beijing Normal University, China. During her master's study, she exchanged to Waseda University (Tokyo, Japan) under the 1.5 year Japanese Government (MEXT) Scholarship. She is currently a PhD student in the Computer Science Department at the University of Maryland, College Park. Her research focuses on computer vision and pattern recognition. She is a student member of the IEEE.

Fatemeh Mirrashed received the MSc degree in electrical engineering from Pennsylvania State University, and is currently a PhD student in the Computer Science Department at the University of Maryland, College Park. Her areas of research include pattern recognition and computer vision.

Inbal Yahav received the BA degree in computer science and the MSc degree in industrial engineering from Israel Institute of Technology. She is a PhD candidate in the Department of Decision, Operations & Information Technologies at the Smith School of Business, University of Maryland, College Park. Her main research interest is the interface between operations research and statistical data modeling. She has presented her work at multiple conferences and has published papers in books and journals.

Aleks Aris received the PhD degree in computer science from the University of Maryland, College Park, where he was a member of the Human-Computer Interaction Laboratory, and collaborated with Prof. Ben Shneiderman on the Treemap, TimeSearcher, and NVSS projects. In 2003, he received a Teaching Excellence Award from the department, awarded to one graduate teaching assistant per year. He was also a summer intern at Microsoft Research in 2004 and worked on the MyLifeBits project.

Ben Shneiderman is a professor in the Department of Computer Science, founding director (1983-2000) of the Human-Computer Interaction Laboratory, and member of the Institute for Advanced Computer Studies at the University of Maryland at College Park. He is coauthor with Catherine Plaisant of *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (fifth edition, 2010). He is a senior member of the IEEE and IEEE Computer Society.