Visual Exploration of Biomedical Databases

Mike Lieberman Sima Taheri
Huimin Guo Fatemeh Mir Rashed
Institute for Advanced Computer Studies
Department of Computer Science
University of Maryland
College Park, MD 20742
{codepoet,taheri,hmguo,fatemeh}@cs.umd.edu

Inbal Yahav
Smith School of Business
Department of Decision, Operations and
Information Technologies
University of Maryland
College Park, MD 20742
iyahav@rhsmith.umd.edu

ABSTRACT

In recent years the amount of public biomedical data has increased dramatically. Improving understanding of these vast, ever-growing repositories of information through visual methods is thus increasingly important for biomedical research. A new approach is offered for visualizing a query-based subset of the National Center for Biotechnology Information's databases. These databases are modeled as an entity-relation graph, and the graph structure is exploited using a graph-based data collection tool. To collect data, users specify a query search string and a data collection path. The resulting data is then visualized using an implementation of user-defined semantic substrates. Comments from domain experts indicate that this visualization method is potentially advantageous for biomedical knowledge exploration.

1. INTRODUCTION

The amount of publicly available biomedical data, and especially of genomic and proteomic data, has ballooned in the past several years with ever-improving technology and computational methods. The development of methods like shotgun genome sequencing in the late 1990s vastly decreased the time required to sequence a complete genome, and hence allowed for the generation of large amounts of sequence data in a relatively short time. In addition, improved computational resources allow for the modeling, simulation, and study of many aspects of computational biology, such as sequence alignment, gene finding, protein structure prediction, and molecular dynamics. 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?

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 the various 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. Given the huge amount of data at NCBI, and the large number of databases, myriad variations of these associations are possible.

To organize this 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, individual knowledge points, such as documents, genes, proteins, and other object types correspond to nodes of the graph. 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. This type of top-down visualization simplifies the identification of concepts like *research fronts* [9].

However, our motivation lies not in discovering overall trends, but rather in accomplishing the everyday technical tasks of knowledge exploration faced 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 result, when using NCBI's databases, they begin with a specific query or set of queries, and explore outward from the initial query result. We seek to design a visualization tool that aids 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, we believe that explicitly viewing and exploring multiple nodes in parallel will lead to improved performance in exploration and discovery tasks. We can enable this exploration by starting with a query-based subset of nodes from the NCBI databases, and dynamically exploring and expanding links with other database entries, in order of query relevance or user preference.

We display graph nodes and links in *user-defined semantic substrates* [16] by leveraging on the capabilities of the graph visualization tool known as *network visualization by semantic substrates* (NVSS) [3,16]. We also extend the capabilities of NVSS by providing a data collector that dynamically retrieves query results using NCBI's web application services. The data collector models user exploration as a *query tree*, where each node of the tree corresponds to the results of a particular search query.

The rest of the paper is organized as follows. In section 2, we provide a literature survey, covering a range of biomedical discovery methods, data collection and visualization tools. Section 3 describes the domain experts that gave us direction in exploring the biomedical databases. Section 4 introduces our data model and data collector. We illustrate our approach using semantic substrates with several example visualizations in section 5, and describe comments from our experts in section 5.3. Finally, we offer concluding remarks in section 6. In addition, a list of features we would like to see added to NVSS is presented in section A.

2. RELATED WORK

2.1 Exploratory Tools for Biomedical Data

There exists a substantial body of work in the literature to help scientists with the daunting task of knowledge discovery in very large biomedical databases. Many different tools and approaches have been used, including text mining, text summarization, semantic annotation, and clustering. *ArrowSmith* [17, 18], a web-based tool that supports the discovery of relationship between two sets of literature in Medline database, is an example of such systems. This system enables users to look for items or concepts that may be common between two distinct sets of articles. Though it may be useful for discovery of meaningful but indirect links between disjoint research works, it remains limited to publication databases.

Given the complexity and huge diversity of biomedical resources, using information visualization techniques has been the next natural step. *Osprey* [6] is a biological network visualization system. It is a domain-specific tool which enables biologists to combine data sets and provides node filtering based on attributes. Osprey not only represents interactions in a flexible and rapidly expandable graphical format, but also provides options for functional comparisons between datasets. Although Osprey plays well in viewing interaction networks in a graphing application, it focuses on genes only, that is, it does not support analysis for multiple node types.

Boyack et al. [4] study how genes, protein and papers are interconnected via co-occurrence patterns in the melanoma research field. This system generates a Paper-Gene-Protein map (papers from Medline, genes from the Entrez Gene database, and proteins from UniProt) to see their co-occurrence relationships. The association network is visualized and explored by the calculation of pairwise similarities between records and the result is spatially laid out in a landscape view using a force-directed algorithm. This work achieves its goal of giving researchers a global view of the structure and dynamics of a research domain. However it is not extendable to our mission which is helping with the everyday technical tasks of scientists who are researching a particular gene, protein, or topic.

2.2 Network Visualization Tools

Many interactive visualization systems have been developed to help users explore and analyze sets of related documents or communities. *Vizster* [12] presents social networks using a simple network node-link representation, where nodes represent members of the system and links represent the articulated "friendship" links between them. It is a sophisticated social network data visualization system that end-users of social networking services can use to facilitate discovery and increased awareness of their online community. *SocialAction* [15] is an interactive exploratory tool for social network analysis. It filters nodes and edges using rankings by statistical measures such as *betweeness-centrality*. It can show clustered groups of nodes, called *communities*, that are determined by using a structural clustering algorithm with user-controlled parameters. *DualNet* [14] shows how using multiple coordinated views improves navigation and provides insight into large networks with multiple node and link types in network data visualization. The tool enhances display of attributes at one time and allows comparisons of different subsets. The knowledge visualization tool *VxInsight* [5] transforms information such as documents, patents, or even genomic data into an intuitive visual format that is easy to interpret. It presents information as a landscape which allows very large datasets to be represented. *CiteSpace* [7] is a system for detecting and visualizing trends and changes in scientific disciplines over time. Two complementary visualization views are designed and implemented in this system: cluster views and time-zone views.

Most existing network visualization tools are based on the *force-directed* layout algorithm [11]. The basic idea behind this approach is to model links as mechanical springs, with each spring's strength proportional to the link's weight. It is favored because it provides reasonable node visibility by spreading out network nodes. It also tends to reduce the number of link overlaps by moving nodes to their natural position, and is useful for visualizing relationships of uniform node types, such as social networks. However, we argue that the force-directed approach is unsuitable for the NCBI entity-relation graph, for several reasons. We want to give nodes some spatial significance, e.g. by grouping nodes of the same type together. For example, it can be difficult or impossible to place nodes with the same type together, as this layout relies only on links for node placement. The semantic information associated with these groups of nodes (i.e. they are of the same type) is thus lost. To work in a force-directed context, we would therefore need to introduce unwieldy additions to our network, such as "virtual" links between nodes of the same type that function to group these nodes together.

Network visualization by semantic substrates (NVSS) [3, 16] is another method which works for our network analysis task. It supports layouts based on node attributes. Placing nodes in this way helps make data more comprehensible to users. Semantic substrates require two conceptual steps to organize nodes. First, nodes are grouped into regions according to one of their attributes. Second, nodes are positioned within each region according to one or more other attribute values. Once the nodes are organized, user control of link visibility according to their source and destination regions reduces the cluttered displays that exist in many implementations. NVSS also includes a tool called the Substrate Designer that allows users to visually define all parameters of the substrate, including region size and position, attributes for grouping nodes into regions, and attributes for positioning them within regions.

2.3 Data Collection Tools

Besides network layout strategies, an efficient data collector of biological resources (e.g., NCBI recources) is also required. In many visualization systems, interaction networks are viewed within a graphing application, but data is collected and manipulated in other contexts, often manually. Hasan Davulcu et al. [8] proposed an approach that combines two software tools, WinAgent and dbXML, for biological data integration. The interface can retrieve information from different Web sites and store the extracted data in a standardized format for later use. They give an example of navigating between two biomedical databases by retrieving bibliographical references related to a genetic disorder. Previous research [10] has shown that scientists cannot be expected to use and query XML databases, as described in [8]. A more scientist-friendly querying interface that enables scientists to express their complex data collection protocols would be better.

3. DOMAIN EXPERTS AND TASKS

Millions of biomedical domain scientists search NCBI's various databases every month. These experts come from a wide array of fields and have a diverse set of query needs, involving different databases. For example, a doctor searching for information about cases of a particular disease might want to query PubMed, NCBI's database of published medical documents, while a genome researcher would be more interested in exploring NCBI's gene and sequence databases.

To determine a broad set of specific tasks for our visualization design, we interviewed several biomedical domain experts who provided design suggestions and insight into how they most often used NCBI's online databases. These experts work in a wide variety of fields, and as such, they tend to use different NCBI databases in their work. Their comments and feedback helped us focus on particular representative queries with which we tested our visualization interface.

3.1 Tracking Research By Topic

We first interviewed Michael Galperin, a Ph.D. research scientist working at the NCBI who is involved in microbial genomics research. Galperin uses a large portion of the available NCBI databases extensively for his work, which involves a wide variety of research topics and areas. He deals extensively with common protein sequences shared among various related organisms, known as *clusters of orthologous groups*, and therefore tends to make most use of NCBI's gene and protein databases.

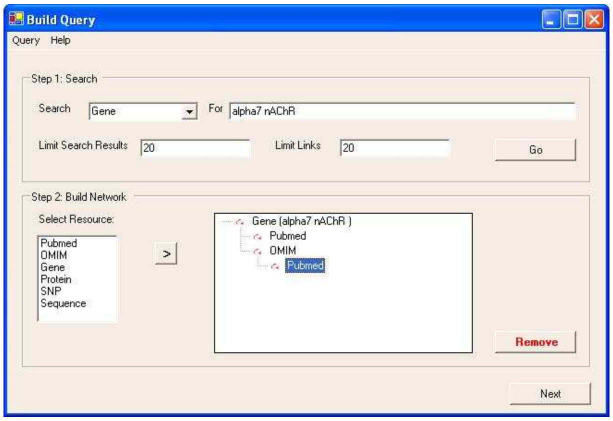
In his work, Galperin executes many queries on a particular topic or area to get a better understanding of the existing work in that area, and what possible contributions could be made. He suggested that a good start would be to execute a topical query in the Online Mendelian Inheritance in Man (OMIM) database, as each entry contains links to the various kinds of information available about a particular topic. For example, an OMIM entry about "cervical cancer" would offer links to various publications, genes, proteins, and observed DNA sequences related to cervical cancer. This type of OMIM-centric exploration would give a user a general understanding of the queried topic. Galperin also suggested gene-centric queries beginning with the Entrez Gene databases, as the majority of NCBI users all deal with genes in some way.

In addition, Galperin indicated his interest in finding documents related to an existing document, using a text-similarity search engine called eTBlast [13]. He also suggested a visualization showing the path of scientific discovery, by displaying the temporal sequence of documents, genes, and other related materials created relative to a particular biomedical topic.

3.2 Medical Term Mining

We also met with Louiqa Raschid, a professor affiliated with the University of Maryland's Center for Bioinformatics and Computational Biology (CBCB), and Adam Lee, one of her students involved in computational biology. Their current research projects revolve around new ways of visualization and exploring biological resources, with particular emphasis on searching for unknown patterns in biological data. Raschid and Lee have been investigating resource and link patterns in various interrelated databases, including the Gene Ontology (GO) and NCBI's PubMed, Medical Subject Headings (MeSH), and Entrez Gene repositories.

Raschid suggested that the MeSH and GO databases were of most interest to her. She was sympathetic to our approach of starting with a query-based subset to visualize, and expanding outward from this initial subset, indicating that it would be useful to her in exploring relationships between datasets. However, she was doubtful whether the majority of existing users of NCBI's databases would adapt well to our visualization and use it for exploration (i.e. displaying query results as explicit nodes and links, rather than webpages). She therefore suggested that we implement a details-on-demand feature to allow users to display a selected node at NCBI's website in a web browser. This would give users a more familiar interface to lean on if they were unable to interpret results correctly in our interface.



Step 1: Search

Search Gene For alpha7 nAChR

Limit Search Results 20 Limit Links 20

Step 2: Build Network

Select Resource:

Pubmed
OMIM
Gene
Protein
SNP
Sequence

- Gene (alpha7 nAChR)
  - Pubmed
  - OMIM
    - Pubmed

Figure 1: A screenshot of our data collector. The user specifies a keyword query in the text box at the top, and specifies node and link limits. The query tree is then specified at bottom.

3.3 Genome Analysis

Our third domain expert was Adam Phillippy, a Ph.D. student also working in the CBCB at the University of Maryland. Unlike Raschid and Lee, Phillippy works mainly with genome sequence analysis and alignment, so the databases of most interest to him are those containing sequences and proteins for particular species. He also uses NCBI's Basic Local Alignment Search Tool (BLAST) [2] extensively to perform sequence alignment and pattern searching. BLAST takes a DNA sequence as input, and searches its massive database of sequence data to return all other sequences that share much of the same nucleotide sequence (i.e. other sequences that *align* with the query sequence), as well as a score for each sequence indicating the amount of alignment. He also uses NCBI's species taxonomy tree to find groups of related organisms and parts of their DNA sequences.

Phillippy indicated his dissatisfaction with several aspects of NCBI's current query interface that required too many mouse pointer clicks and scrolling. He suggested that BLAST query results could be presented in an easier to understand manner by mapping each sequence to a graph node and creating weighted links between sequences that align well, where the weight corresponds to the alignment score. In particular, because each result sequence might come from different organisms, the BLAST results could be integrated with NCBI's taxonomy tree browser, to discover relationships between all organism sequences found in the result.

Phillippy also mentioned other visualization potentially of interest to NCBI users. He was interested in an easy-to-understand way to learn what genome sequencing projects were currently underway at the various sequencing centers around the world, such as the Broad Institute. Phillippy also wanted an easy way to browse NCBI's taxonomy tree, rather than having to scroll through a large page of text.

4. DATA MODEL AND COLLECTOR

4.1 Data Model

To arrive at an understanding of our design choices, it is necessary to describe our model of NCBI's various databases in more detail. As mentioned previously, 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 attributes corresponding to the types of information contained within. Furthermore, attributes vary according to the type of node. For example, PubMed document nodes have attributes for the document's title, authors, year of publication, and keywords, while Entrez Gene nodes have attributes for gene name, genus and species name, and chromosome location. To ensure proper and meaningful node placement in semantic substrates, several numeric node attributes should be extracted from each node that have some underlying semantic meaning. Fortunately, each node type in the NCBI graph had many attributes and it was therefore easy to settle on appropriate attributes for node layout.

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 generalize:

1. links may be weighted or unweighted;
2. links may be directed or undirected, i.e., the graph is *mixed*; and
3. two nodes may be connected by multiple links, i.e., the graph is not *simple*.

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.

Many associations in the NCBI databases are added manually as new records get added to the database. However, other links are computed by the system. It may therefore be useful to distinguish these computed links from manually added links for a compelling visualization.

4.2 Data Collector

Entrez Programming Utilities (eUtils) are tools that provide access to the Entrez Global Query Cross-Database Search System outside of NCBI's regular web query interface. eUtils are composed of the following seven server-side programs that access the Entrez system through a fixed URL syntax:

eInfo Provides information about the set of databases, their terms and their attributes.

eSearch Searches and retrieves primary IDs that correspond

eSummary Retrieves document summaries from a list of primary IDs

eFetch Retrieves records (in a requested format) from a list of one or more primary IDs

eLink Retrieves primary IDs and relevancy scores for links to Entrez databases or Related Articles

eGQuery Provides Entrez database counts in XML for a single search using Global Query (cross database search)

eSpell Retrieves spelling suggestions

The Entrez system includes 23 databases, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature. 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 collection tool in C#.NET using this web service.

The data collector generates two tab-delimited text files as output, corresponding to node data and link data. Nodes are required to have at least two attributes: one that determines which substrate region the node will be placed, and another that determines its placement within the region. In practice, nodes have many attributes, some of which are more useful than others for visualization. For example, document nodes we collected had a variety of attributes, including document title, authors, and publication date. To group all document nodes into a single substrate region, we added another attribute, NODETYPE, whose value was "document" for the document nodes. We ordered document nodes within the document region according to the publication year.

4.3 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 a user would want to query. Figure 1 is a screenshot of our data collector's user interface. It enables the user to access the NCBI databases by defining a *search query* and a *query tree*. The search query is a collection of keywords and an initial database (e.g., OMIM) to search in. 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 exploration taken by the data collector. Each node in the tree corresponds to one of NCBI's databases, and links between query tree nodes represent types of links that will be collected between databases. 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 are retrieved using the eFetch utility.

To allow users a flexible definition of required attributes without interacting with the data collector's code, we store the list of attributes outside the software, in an XML document. 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.

4.4 Data Retrieval Challenges

NCBI enforces rate limits for programs using the XML eUtils interface. Programs using the interface are limited to a single request every 3 seconds. In addition, the system imposes limits for 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.

Our initial versions of the data collector experienced timeouts 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 for each

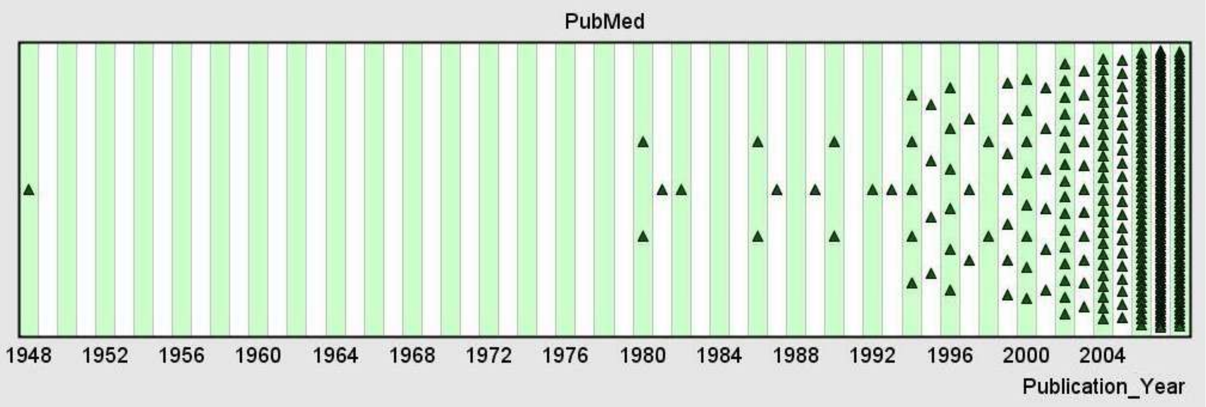
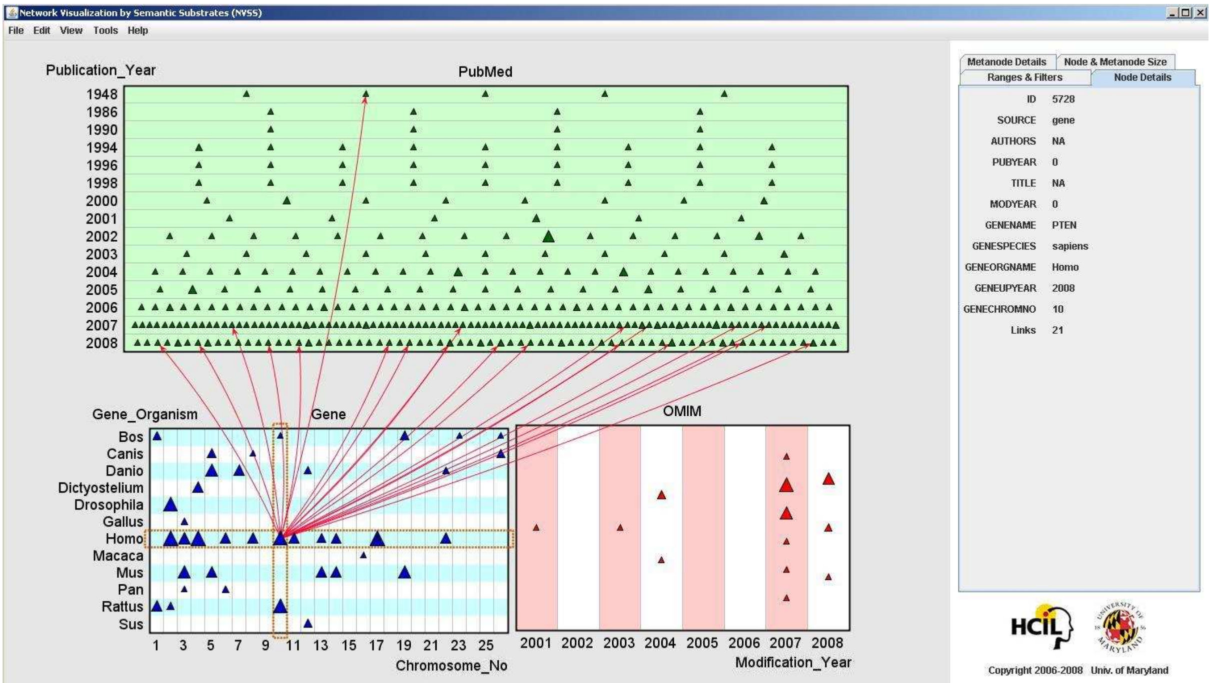


Figure 2: The PubMed region from our initial query visualization. This region appears cluttered, due to a large number of publications in recent years. In addition, the outlier paper in 1948 creates a large gap in the region.



ID 5728
SOURCE gene
AUTHORS NA
PUBYEAR 0
TITLE NA
MODYEAR 0
GENENAME PTEN
GENESPECIES sapiens
GENEORGNAME Homo
GENEUPYEAR 2008
GENECHROMNO 10
Links 21

Copyright 2006-2008 Univ. of Maryland

Figure 3: A visualization of the cervical cancer query results. In the PubMed region, adaptive bin sizing is used, and the size of each node is proportional to its degree. Also, the gene PTEN has links to a group of recent publications, in addition to the older paper in 1948. It can be a sign of reviving an idea.

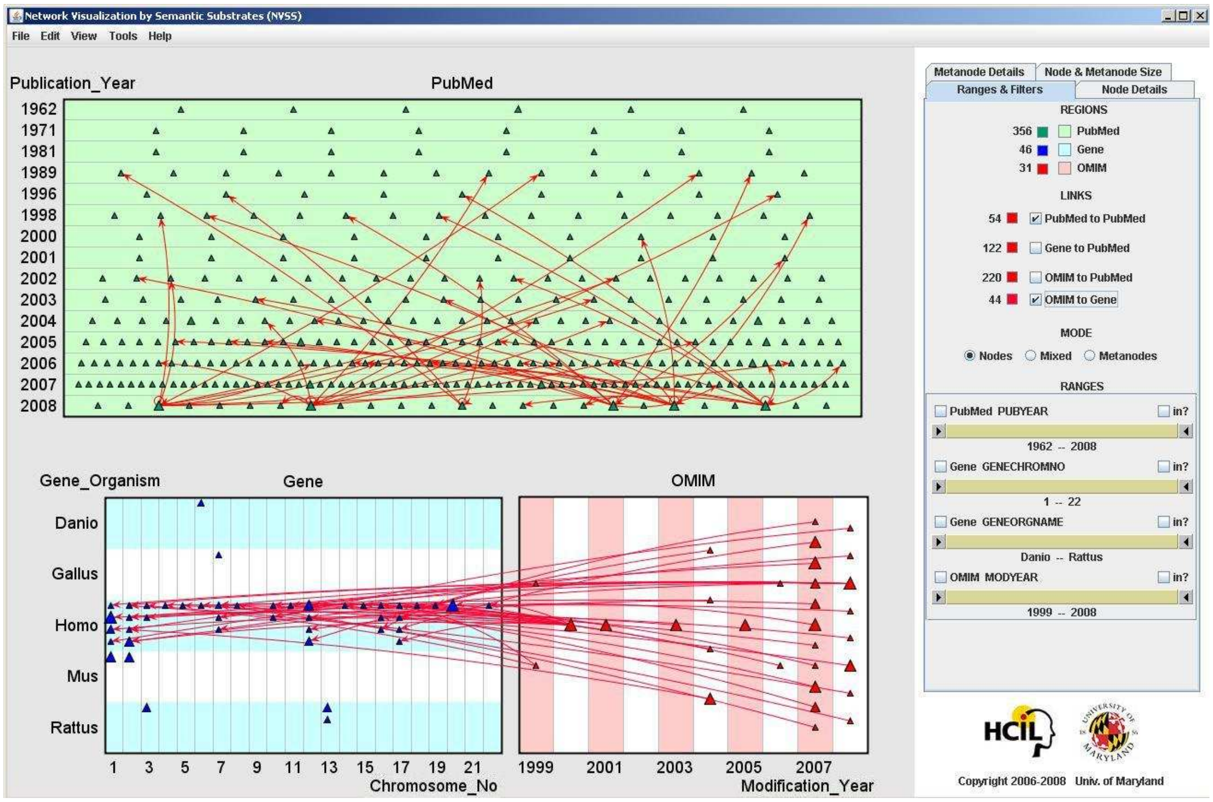
query, such as title, body text, and clinical synopsis. We thus ensured that each individual query was completed within the required time limit, while still collecting enough data to be useful. Due to these multiple independent queries, we often retrieved redundant node entries, which were removed from the final result.

Another challenge we faced is the inconsistency between the database attribute listing, as retrieved from eInfo, and the actual attributes to use in eFetch. The inconsistency manifests itself both in the attribute alias and in the actual list of attributes. To overcome this problem, we generated our own attribute lists, based on the attributes obtained through eFetch.

5. VISUALIZATION WITH NVSS

We wish to enable dynamic exploration of a subset of the NCBI entity-relation graph. To accomplish this, we need to ensure that several important features of the network are not lost in a visualization. In particular, the various nodes of our network have different types — e.g., documents, genes, and proteins — and we want users to be able to easily distinguish between these types, so that users can decide to further explore only nodes of a particular type. Moreover, the visualization should allow users to focus on a single node or small set of nodes and see all links going to or from this set.

To accomplish these requirements, we position nodes in a *semantic substrate* based on node type and other node attributes. This layout method groups nodes into explicit substrate regions based on node type. For example, all document nodes, gene nodes, and protein nodes will be placed in their own regions, allowing users to spatially understand what types of data they are viewing with-



REGIONS

356 🟩 □ PubMed

46 🟦 □ Gene

31 🟥 □ OMIM

LINKS

54 🟥 ☑ PubMed to PubMed

122 🟥 □ Gene to PubMed

220 🟥 □ OMIM to PubMed

44 🟥 ☑ OMIM to Gene

MODE

● Nodes ○ Mixed ○ Metanodes

RANGES

□ PubMed PUBYEAR □ in?

1962 -- 2008

□ Gene GENECHROMNO □ in?

1 -- 22

□ Gene GENEORGNAME □ in?

Danio -- Rattus

□ OMIM MODYEAR □ in?

1999 -- 2008

Copyright 2006-2008 Univ. of Maryland

Figure 4: Visualization of the hypertension query results. The OMIM-Gene and PubMed-PubMed links are shown.

out the need for scanning nodes on screen. In addition, nodes are positioned within substrate regions according to attribute values, allowing further semantic meaning to be spatially extracted. As a further benefit, multiple links between nodes pose less of a challenge for the semantic substrate layout, as it relies purely on node attributes rather than link strength. There is also no problem with placing nodes without any links.

We used NVSS as an implementation of visualization by semantic substrates. NVSS allows the creation of user-defined semantic substrates using a substrate editor. For visualization, NVSS allows the user to filter nodes within substrate regions using slider bars, and also allows selective link filtering based on user preference.

5.1 Sample Datasets

Based on our interviews with domain experts, we created several visualizations of sample NCBI queries that might interest typical users of the NCBI databases. We extracted query-specific node and link data from three of NCBI's most-queried databases:

1. *Online Mendelian Inheritance in Man* (OMIM), a catalog of human genes and genetic disorders, with links to a wide variety of relevant information, including literature references, sequence records, maps, and related databases;

2. *PubMed*, a database of published medical literature and index terms;

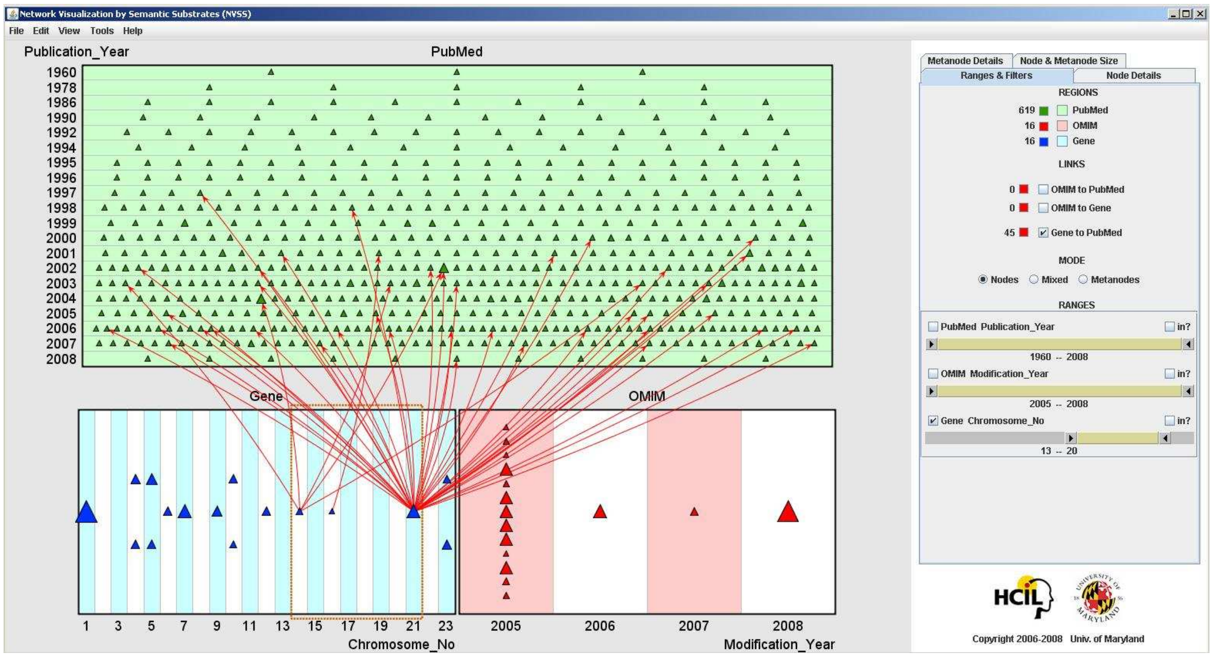
3. *Entrez Gene*, a collection of gene-centered information, including cross-references to relevant genomes and sequences.

To collect data, we chose a database to begin the exploration, and used a keyword search to retrieve an initial set of nodes and all associated attributes. We used the OMIM database for our initial queries, as each OMIM entry contains links to many other types of information spread throughout NCBI's various databases. Then, we collected links to nodes in PubMed and Entrez Gene, as well as the set of attributes and values for each additional node.

One important step for visualizing the dataset was to select attributes to use for grouping nodes into substrate regions, as well as positioning nodes within the regions. Grouping nodes into regions was fairly straightforward. We created a meta-attribute called NODETYPE for each node that indicated the database from which the node was retrieved, and used NODETYPE for node grouping. Next we had to decide how to position nodes within substrate regions by selecting meaningful attributes for each node type. We used different attributes depending on the type of node, as attributes generally were not shared between node types. Positioning document nodes was fairly simple, as a logical choice was the year of publication. For OMIM entries, which are not formally published, we settled on the date of last modification, as this indicates how recent the entry's information is. To position gene nodes, we used the gene's position on the chromosome due to the fact that genes whose sequences appear close together in an organism's genome have similar functions. Note that a fuller investigation is needed to determine appropriate attributes to use for positioning the many types of biomedical data in NCBI's databases, as the selection of different attributes can lead to different discoveries and insights.

5.2 Sample Queries

Our initial queries were related to cervical cancer, which is a model disease with a genetic component. Using the data collector, we retrieved related entries to cervical cancer in the OMIM database, as well as those entries in the PubMed and Entrez Gene



REGIONS

619 🟩 □ PubMed
16 🟥 □ OMIM
16 🟦 □ Gene

LINKS

0 🟥 □ OMIM to PubMed
0 🟥 □ OMIM to Gene
45 🟥 ☑ Gene to PubMed

MODE

● Nodes ○ Mixed ○ Metanodes

RANGES

□ PubMed Publication_Year □ in?
1960 -- 2008

□ OMIM Modification_Year □ in?
2005 -- 2008

☑ Gene Chromosome_No □ in?
13 -- 20

Figure 5: Visualization of the nudix query results. The node layout using semantic substrates allows for exploration of multiple gene nodes in parallel, which may lead to knowledge discovery.

databases to which the OMIM results have links. As previously mentioned, the OMIM database only provides links to human genes. But one interesting exploration could be to find the relevant genes in other species, such as rat, mouse, and monkey. To this end, we also directly executed the query in the Entrez Gene database, and collected relevant genes from different species, and their links to PubMed entries. In total, we retrieved 13 OMIM entries, 343 publication records in PubMed published between 1948 and 2008, and 42 Entrez Gene records from 12 different species. Moreover, the data collector returned 412 OMIM-PubMed, OMIM-Entrez Gene, and Entrez Gene-PubMed links.

To visualize a dataset using NVSS, we first compile the node and link files that contain the network data. NVSS is sensitive to errors in these input files, as it does not accept files with duplicate nodes or links. In addition, NVSS expects all numerical attribute values to be greater than one. We therefore modify or remove the nodes, links, and attributes that do not meet NVSS's data constraints.

We next use the substrate designer to create a substrate. The process of designing a satisfying substrate, which can be considered a trial and error task, involves several challenges, such as finding appropriate positions and sizes for each substrate region, defining the number and size of bins in different regions to avoid a cluttered display, and choosing proper node placement methods (i.e., based on the nodes' x- or y-coordinates). Among the three different sets of nodes in our dataset, the most difficult and problematic one for visualization was publication nodes. As figure 2 shows, the large number of publications causes several cluttered areas in the PubMed region. Another problem in this figure, which we were not aware of before observing the first visualization, is the large number of empty bins between 1948 and 1980. Such a large gap is detrimental to information visualization, as it wastes the space without providing useful information. As can be seen in figure 3, changing the node placement method to spread the nodes along the horizontal axis and also using adaptive bin sizing remedied these problems. Adaptive bin sizing assigns different lengths to each bin to achieve an almost uniform node distribution within a substrate region. Unfortunately, it requires prior knowledge about the distribution of nodes, so proper bin lengths can be set only after an initial visualization.

Figure 3 shows our final cervical cancer query visualization. The gene nodes are laid out in the substrate using two attributes, namely the chromosome number and the organism name. This xy-coordinate placement method enables users to focus on a particular gene, using slider filters, and explore its corresponding links. Moreover, they can find the node details in the right panel by clicking on a particular node. We also used size-coding for the nodes based on the number of incoming and outgoing links from them. This coding provides additional visual cues about a node's importance, relative to the initial query. The figure also shows an interesting observation we made from this visualization. All the outgoing links from the gene PTEN point to the recent publications from 2007 to 2008, except one of them, which points to an older publication from 1948. Observations like this might be of interest to researchers, as it can be a sign of reviving an idea.

We repeated the same visualization process for other queries related to medical conditions, including "hypertension" and "nudix". For the hypertension query, a total of 357 publications, 45 gene records, 31 OMIM entries, and 440 Links were collected. Figure 4 shows the hypertension query results. For this query, some similarity links between publications were also retrieved.

The nudix query shown in figure 5 resulted in 619 publications, 16 human gene records, 16 OMIM entries, and 713 links among the nodes. Both figures 4 and 5 exhibit the parallel exploration that our visualization makes possible. Using these visualizations, researchers can explore multiple nodes, and examine the links between them, in a single display. For example, one observation we can make about the nudix query in figure 5 is that the three highlighted genes have completely different numbers of links to PubMed documents, possibly indicating their relative importance to the query. We believe that these observations can lead to knowledge discovery and unforseen insights.

5.3 Expert Comments

Unfortunately, one of our domain experts, Michael Galperin, could not continue his participation in our project due to his busy schedule. However, we received comments from another domain expert, Alberto Labarga, who was seeking a way to visualize DNA microarray data. Labarga appreciated our idea of using semantic substrates over competing approaches, such as force-directed layouts, and was positive in general about our approach. He believes that semantic substrates and NVSS are powerful tools for visualizing various types of biomedical data. However, he suggested that the attributes we used for gene and PubMed placement were not so interesting. Labarga suggested using biological enrichment tools, such as FatiGO and FatiScan [1], to divide Entrez Gene substrate regions. He also mentioned that PubMed documents could be clustered according to content, instead of by publication year, which might make for a more enlightening visualization.

6. FUTURE WORK AND CONCLUSIONS

There are several possible future directions for this project. In the current data collector, users are required to pre-define a query search, query tree and a list of attributes they wish to retrieve. In future versions, we plan to extend the query interface to use a query DAG, rather than the query tree, and to allow dynamic addition of new nodes or links to the query.

We believe that hierarchical organization of nodes can be very insightful for exploratory tasks. We would therefore like to break substrate regions into *subregions* based on node attributes. For example, a useful visualization would group all genes within a substrate region, but would group all genes from particular species' within individual *subregions*. Human genes could be placed to the left of a region, and all mouse genes to the right. Node positioning within each subregion could remain the same as in the original region.

We also want to enable displaying and exploring tree data (e.g., Gene ontology, MeSH terms), and visualizing connections to graph data. For example, NCBI features a taxonomy tree with relationships between related organisms, so it might be used for a visualization involving genes shared by Homo sapiens and Rattus norvegicus. We also intend to investigate other attributes for laying out biological data of various types.

In this paper we presented a new approach for exploring and knowledge discovery in biomedical databases. Parting from traditional search results lists like those available at the NCBI website, we illustrated a powerful graph-based search using an NCBI data collector, and visualization of query results using user-defined semantic substrates. We also evaluated NVSS as an application of semantic substrates for biomedical data. We believe that graph visualizations like semantic substrates will have increasingly important roles in understanding biomedical data in the near future.

7. REFERENCES

[1] F. Al-Shahrour, P. Minguez, J. M. Vaquerizas, L. Conde, and J. Dopazo. Babelomics: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Research*, 33:460–464, July 2005.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990.

[3] A. Aris and B. Shneiderman. Designing semantic substrates for visual network exploration. *Inf Visualization*, 6(4):281–300, Nov 2007.

[4] K. W. Boyack, K. Mane, and K. Borner. Mapping medline papers, genes, and proteins related to melanoma research. *Proceedings of IV2004 Conference*, 00:965–971, 2004.

[5] K. W. Boyack, B. N. Wylie, and G. S. Davidson. Domain visualization using vxinsight for science and technology management. *Journal of American Society for Information Science and Technology*, 53:764–774, 2002.

[6] B. J. Breitkreutz, C. Stark, and M. Tyers. Osprey: a network visualization system. *Genome Biology*, 4:R22, 2003.

[7] C. Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology.*, 57(3):693–700, 2006.

[8] H. Davulcu, Z. Lacroix, K. Parekh, I. V. Ramakrishnan, and N. Julasana. Exploiting agent and database technologies for biological data collection designing. *Proceedings of DEXA Workshops'2004*, pages 376–381, 2004.

[9] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[10] A. B. Eckman, K. Deutsch, M. Janer, Z. Lacroix, and L. Raschid. A query language to support scientific discovery. *The Second International IEEE Computer Society Computational Systems Bioinformatics Conference*, pages 388–390, 2003.

[11] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software Practice and Experience*, 12(11):1129–1164, 1991.

[12] J. Heer and danah boyd. Vizster: Visualizing online social networks. *IEEE Symposium on Information Visualization (InfoVis)*, page 5, 2005.

[13] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. R. Garner. Text similarity: an alternative way to search medline. *Bioinformatics*, 22(18):2298–2304, September 2006.

[14] G. Namata, B. Staats, L. Getoor, and B. Shneiderman. A dual-view approach to interactive network visualization. *International Conference on Information and Knowledge Management(CIKM2007)*, pages 939–942, 2007.

[15] A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12:693–700, 2006.

[16] B. Shneiderman and A. Aris. Network visualization by semantic substrates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):733–740, 2006.

[17] N. R. Smalheiser and D. R. Swanson. Arrowsmith: Identify meaningful links between two sets of medline articles. `http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html`. Accessed May 7, 2008.

[18] 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*, 57:149–153, 1998.

APPENDIX

A. NVSS WISHLIST

NVSS is a great tool with many interesting features. However, working with NVSS, we found ourselves often wanting an additional feature that would enable more informative information exploration, both in the biomedical domain and in general. We now present a list of some of the more compelling features we would like to see in later versions of NVSS.

- Showing links both going out from and in to a particular node when the filters are active, i.e. "OR" semantics in addition to "AND" semantics
- Visualizing links differently depending on their attributes, e.g. similarity versus citation links
- Having a customizable details-on-demand feature, which our case means opening the a node's corresponding page in the NCBI website when the user selects that node
- Optionally sorting nodes in increasing or decreasing order when choosing region layout
- Supporting both directed and undirected links
- Enabling users to select a few nodes and filter by those nodes, rather than using slider filters
- Optionally color-coding nodes according to attribute values in addition to size-coding
- Allowing changes to the data model file without recreating the entire substrate from scratch
- Zooming and panning features, allowing a user to focus on a piece of the entire data space
- Allow dynamic changes to bin limits and positioning attributes to facilitate exploration of data