

## AN ACTIVE VISUAL SEARCH INTERFACE FOR MEDLINE

Weijian Xuan<sup>1</sup>, Manhong Dai<sup>1</sup>, Barbara Mirel<sup>2</sup>, Justin Wilson<sup>1</sup>, Brian Athey<sup>2</sup>, Stanley J Watson<sup>1</sup>, Fan Meng<sup>1,2\*</sup>

*1 Molecular and Behavioral Neuroscience Institute and Department of Psychiatry,*

*2 National Center for Integrative Biomedical Informatics, University of Michigan*

*Ann Arbor, MI 48109, USA*

*\*Email: mengf@umich.edu*

Searching the Medline database is almost a daily necessity for many biomedical researchers. However, available Medline search solutions are mainly designed for the quick retrieval of a small set of most relevant documents. Because of this search model, they are not suitable for the large-scale exploration of literature and the underlying biomedical conceptual relationships, which are common tasks in the age of high throughput experimental data analysis and cross-discipline research. We try to develop a new Medline exploration approach by incorporating interactive visualization together with powerful grouping, summary, sorting and active external content retrieval functions. Our solution, PubViz, is based on the FLEX platform designed for interactive web applications and its prototype is publicly available at: <http://brainarray.mbni.med.umich.edu/Brainarray/DataMining/PubViz>.

### 1. INTRODUCTION

Understanding the biomedical significance of high throughput data, such as those from microarray gene expression analysis, genome-wide SNP genotyping and biomedical images from in situ or MRI, is a major challenge in the postgenomic era. Researchers often have to examine large bodies of literature in unfamiliar fields for new insights in their area of interest. Unfortunately, prevailing Medline search approaches were largely designed for the efficient retrieval of a small number of records rather than an in-depth exploration of a large body of literature. The inherent limitations in the prevailing search engines such as Entrez and Google Scholar prevent them from being effective large-scale literature exploration tools.

Firstly, the widely-used Medline search methods rely heavily on a step-wise narrowing of search scope but such an approach does not work well for exploring new territories. This is because employing sensible filtering criteria to investigate all potentially relevant topics often requires good background knowledge. For example, in microarray gene expression analysis, researchers frequently have to deal with lists of genes that are not known to be associated with biological processes in which they are interested. Researchers have to utilize other intermediate concepts to establish the indirect link between gene lists and specific biological processes. Identifying these intermediate concepts through literature searches with existing methods, however, is very difficult. Medline searches using a list of such gene names often lead to hundreds or even

thousands of Medline records. Few options are available for identifying potentially relevant topics or novel conceptual relationships other than going through the retrieved records one-by-one in the prevailing Medline search systems.

It will be ideal to have a flexible overview function that can summarize the search results based on criteria from different biomedical concept categories, such as the protein (gene product) interaction, pathway or cellular processes, anatomical location, known high level biological process or disease, etc. Besides increasing search efficiency, the grouping and summarization of search results using different criteria will provide many additional biomedical concepts that can potentially link unfamiliar gene names to the targeted pathophysiological processes. This ability to view the summaries of a large record set from different angles has several benefits. It exceeds even the current advance of using single grouping criteria, such as the MeSH term-based grouping, something that we implemented in our GeneInfoMiner<sup>20</sup> and in doing so significantly improves the Medline search efficiency. Additionally, examining search results from different viewpoints will stimulate new ideas. Systematic mapping of search results to different concept categories, such as interacting proteins, pathways and anatomical locations, is also likely to be more comprehensive than what a researcher can think of at a given moment, prompting him/her to examine a problem from more different aspects of under many situations.

Secondly, it will be a boon to researchers if they could see contextual similarity relationships among different records instead of the linear lists presented by PubMed or Google Scholar. Yet even when retrieved records are mapped to MeSH terms and presented in tabular format as we do in our GeneInfoMiner<sup>20</sup>, information about the similarity of record sets associated with different MeSH terms is not shown. Such similarity information between different groups of records is very useful for revealing novel conceptual relationships as well as for increasing the accuracy of document retrieval. Some applications have tackled the difficulty of apprehending the pairwise similarity relationship between different records or record sets through various visualization techniques<sup>4,7,11,13,15,17,21</sup>. The desktop application RefViz is a noteworthy example. It uses a “galaxy” view for exploring large Medline record sets after similarity-based clustering<sup>19</sup>. Yet web implementations of such similarity-based Medline record overviews have encountered obstacles due to the CPU-intensive nature of clustering Medline record sets. As we describe later, modern video hardware can be used to speedup similarity and clustering significantly.

Some new generation Medline search solutions such as ALIBABA and botXminer have begun to overcome these obstacles by using network graphs to display the biomedical conceptual relationships extracted from different Medline records<sup>2,7,13,15</sup>. They aim to enable researchers to grasp the complex biomedical conceptual relationships in search results at a glance. While these pioneering works produce impressive graphics and potentially increase literature exploration efficiency, the usefulness of these tools is severely constrained by the poor performance of conceptual relationship extraction by existing techniques. While there are a number of reasonable solutions for identifying name entities in specific categories such as gene and protein names in biomedical literature<sup>1,18</sup>, none of the existing conceptual relationship extraction methods can deal with content from the full Medline database in a satisfactory manner. As a result, the reliability of such conceptual relationships in such networks, particularly those involving indirect relationships, is questionable under many situations<sup>8,16</sup>. In addition, the apprehension of conceptual networks quickly becomes very difficult as the number of concepts increases beyond 50 or so. A

conceptual network with more than several dozen elements and whose membership and relationship among its members alters upon every new query do not encourage confidence in their typical users.

Another challenge is that merely utilizing information within the boundary of the Medline database in Medline explorations is far from sufficient. Data and knowledge residing outside of the Medline database are critical to an understanding of the full implication of search results as well as to the development of new ideas for subsequent searches. Because of this, some of the existing Medline search engines, such as Entrez, PubGene<sup>7</sup> as well as the ALIBABA<sup>15</sup> and botXminer<sup>13</sup> mentioned previously, add hyperlinks to biomedical concepts in the search results to facilitate the further exploration and understanding of the related concepts. This hyperlink approach significantly improves Medline search results exploration. But solely relying on concept associated hyperlinks has four shortcomings: 1) Low retrieval efficiency: users have to click hyperlinks one-by-one in order to investigate related external information. It will be ideal to have an automatic mechanism to grab the related information automatically from external database and present them together with Medline search results, 2) Separation of related information: because the related information can only be retrieved by clicking a hyperlink underlying a concept in the search results, it is hard to investigate similar external information together in the development of new ideas. For example, if three types of hyperlinks, Entrez Gene, Allen Brain Map and dbSNP are provided to each gene name in the Medline search results, it will be better to present external information in the same category or group hyperlinks for the same type of external data together for a given set of Medline search results, 3) Inability of mapping search results to external data for effective overview. Hyperlink only provides point-to-point association, not summary information about all the search results with regard to an external data source. For example, although it is fairly straightforward to add pathway links to individual gene names in search results, it will be more useful to map the search results to known pathways and to present an overview of search results based on pathways. This way a researcher can easily learn how each pathway is related to the search results based on the number of gene or small molecular associated with Medline records in individual



relevant requires not just that it be available on a screen but that the content is laid out and arranged in different categories that are meaningful to users. Each category of data, moreover, should also be presented in an appropriate mode and style, such as visualized networks arranged and perceptually encoded to highlight contextual similarity among a reasonable number of Medline records. Other modes of presentation that match users' needs include pathway overlay diagrams to show overlays of Medline records with different elements in a pathway, an expandable ontological tree for exploring functional or structurally related terms in the vicinity of retrieved Medline records, and links to *in situ* hybridization image. Additionally, an effective interface should also allow users to switch efficiently to different views of the same set of retrieved Medline records, with graphic views being clear enough to be remembered without image overload. In this way, researchers can easily examine the search results from different perspectives together with different types of external information. Most importantly since the display must support dynamic inquiry and not just a retrieved "fact answer", intuitive visual data exploration functions such as select/unselect, summary, forming new queries, saving results, etc., must be incorporated for the effective mining of the data set. The history of data exploration processes also is very useful since Medline exploration usually is more complicated than simple Medline record retrieval and users often need to go back to previous steps for additional exploration.

Based on the rationales presented above, we started to develop a new Medline search interface that aims at facilitating the interactive exploration of Medline utilizing information from external databases such as Michigan Molecular Interaction Database (MiMI) <sup>6</sup>, KEGG <sup>9</sup> and Allen Brain Atlas <sup>12</sup>. Different from classical search engines designed for most efficient and accurate record retrieval, our solution is mainly targeted at the understanding of high throughput biomedical data, where researchers often need to venture into unfamiliar territories for new insights in specific pathophysiological processes. Our prototype, PubViz is still a work in progress but the prototype with 5000 bipolar-related Medline records as the test data set is accessible at:

<http://brainarray.mbni.med.umich.edu/Brainarray/DataMining/PubViz>.

In this manuscript, we will first present technical aspects of our solution in the Material and Methods section. The Result section will focus on some of the novel data display and interactive search functions in PubViz using real world examples. Issues encountered in our prototype and functions we hope to include soon are described in the Discussion section.

## 2. MATERIAL AND METHODS

### 2.1. System Design Overview

PubViz consist of four major components: 1) A search component: enables users to search for biomedical literature using a series of flexible criteria, e.g. gene ID, MeSH concept, keyword and their combination. 2) A process component: retrieves pre-annotated literature (e.g. gene/protein name tagger, UMLS concept matcher we developed). It will also filter or expand the result set. The PubViz search interface communication with backend process functions is based on extensive web services. 3) An exploration component: presents processed search results in an intuitive and interactive fashion. For example, in the citation view, gene view, and MeSH view, related literatures are presented respectively in network graphs. Each node in the graph represents an entity or a concept. The connections between pairwise nodes are calculated using the similarity algorithms we describe below. Essentially this component generates overviews on the data set from different perspectives. 4) The analysis component: integrates various scaffolds to help users understand the literature set better. These analytic supports include, for example, topic grouping/sorting functions, visual exploration capabilities, extensive external links, data visualization and dynamic filtering functions.

PubViz is developed on Adobe's latest Flex 2.0 platform. It provides efficient development tools and components that allow us to build highly interactive user interface with high efficiency (<http://www.adobe.com/products/flex>).

### 2.2. PubViz Web Services

Since utilizing external information in Medline exploration is a critical design goal of PubViz but most



nature. Fortunately, Graphics processing units (GPUs) provide an inherently parallel platform suited for various distance calculation and clustering problems. The release of the Compute Unified Device Architecture (CUDA) platform for the NVIDIA GeForce 8XXX graphics cards eases the task of implementing distance and clustering algorithms by presenting the graphics card as multi-threaded co-processor<sup>14</sup>. Our initial tests show the above MeSH term-based similarity and clustering calculation of 200-1000 Medline records can be reduced to several seconds just by using one NVIDIA GeForce 8800 GTS card, which only costs around \$550. As a result, interactive visualization of similarity calculation results in web applications for decent Medline record size is now a reality with a low cost computer cluster.

### 3. RESULTS

Since PubViz is still a project in progress, the current fully functional web prototype uses a small corpus for

more efficient prototyping. The corpus contains 5000 bipolar-related Medline abstracts that are tagged by our own gene/protein name, genetic marker, and UMLS concept taggers. The full Medline record access is expected to be ready in July, 2007. Here we describe the main features of PubViz and some functions that scaffolds users' literature search. More functions are being incorporated into PubViz.

#### 3.1. PubViz User Interface

PubViz aims to bring the richness and usability of good desktop applications to the web-based environment (Fig. 1). PubViz is a purely online application, which does not require any installation nor any local upgrade or maintenance. Since PubViz is run on Flash virtual machine, it does not have any compatibility issues across different browsers.

PubViz uses tabbed layouts across the top (view tabs), right hand side (panel tabs) and bottom (data tabs) to enable access to diverse data and panels.

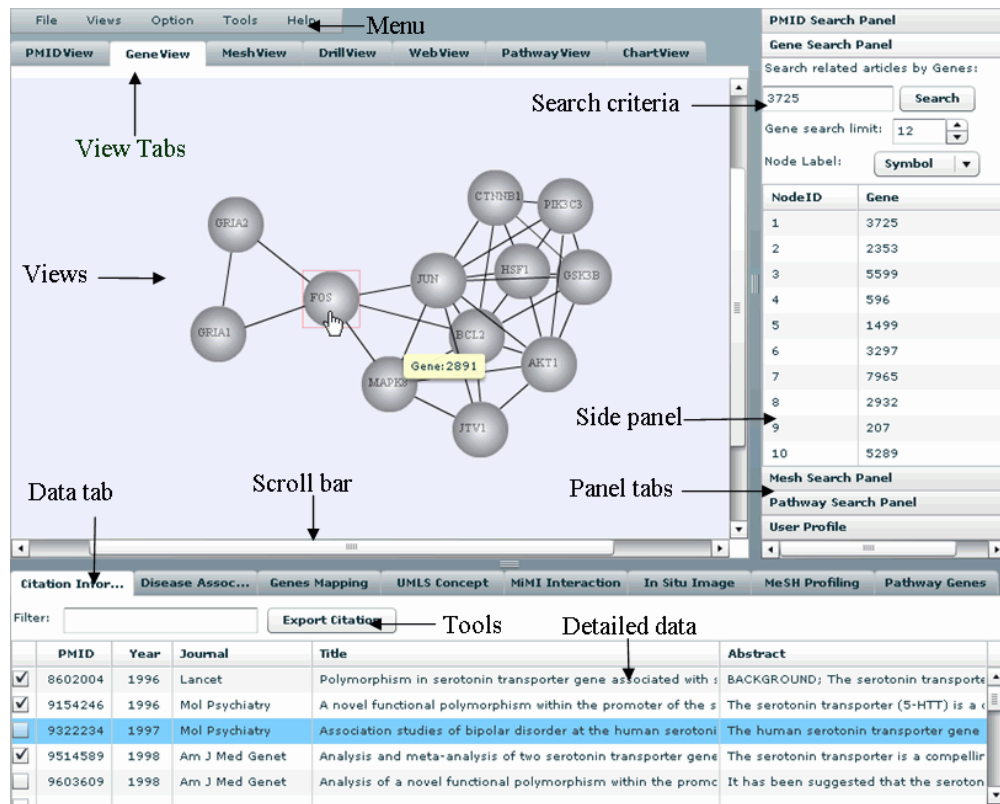


Fig. 1. PubViz interface layout



**Timeline view:** In the first example in Fig. 3 a user wants to see articles related to bipolar disease published between 1966 and 2006 the and clicks on the top “ChartView” tab to begin the search. The user enters the keyword “bipolar” into the search field, sets the time range accordingly, and clicks “Search”. PubViz returns a list of diseases, under the time range slider. The user

clicks on one of these – “Bipolar Affective Disorder” - and a line graph is displayed, showing the number of articles published (y-axis) each year (x-axis). Users can issue another search by just click on a data point on the graph, and the research results can be reflected in other data and view tabs, as well).

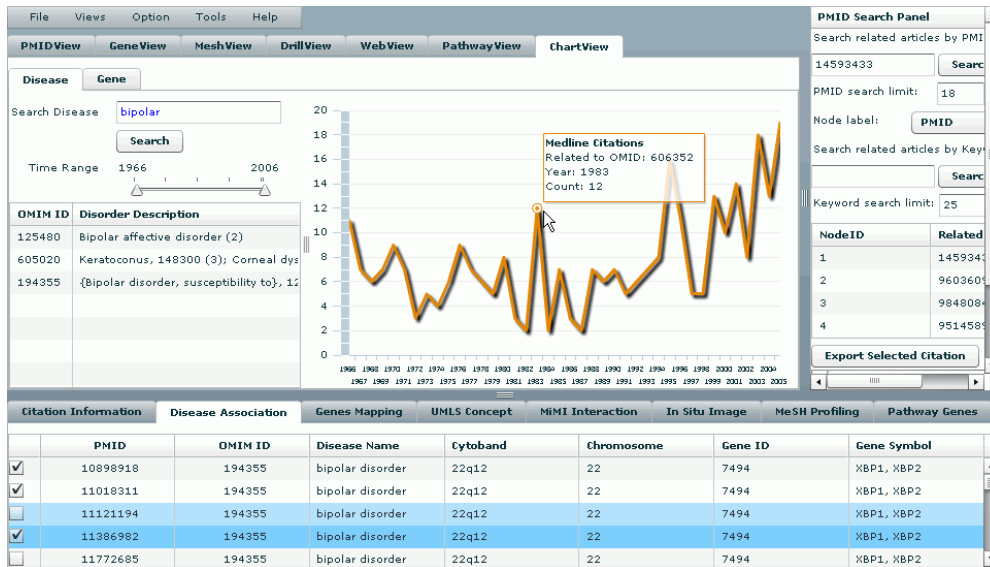


Fig. 3. Timeline View

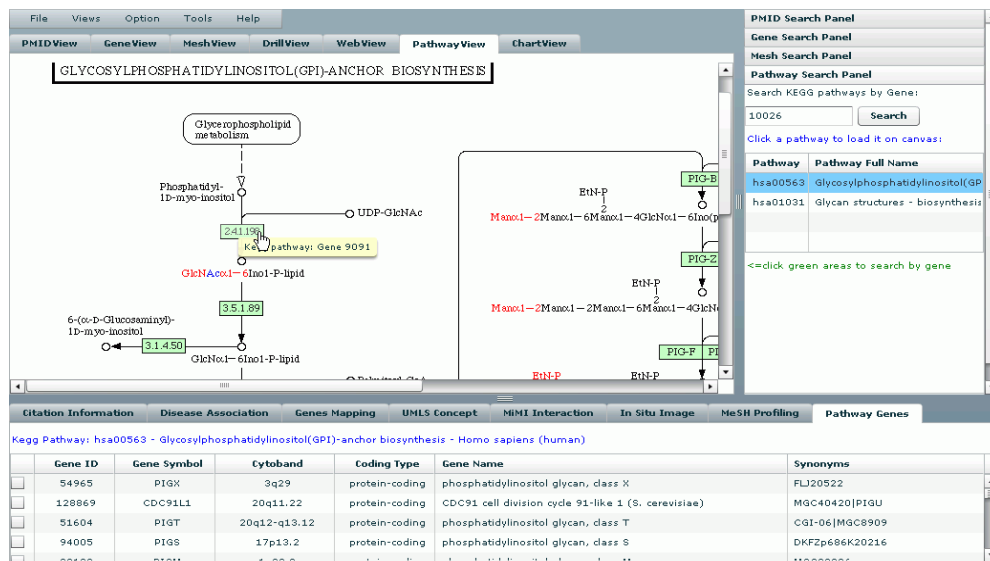


Fig. 4. Pathway View

**Pathway view:** Fig. 4 shows the results that occur when a user opens the pathway search panel tab in order to explore functional aspects of genes of interest. On the side panel, the user enters the Gene ID of the target gene and PubViz maps the Medline search results onto the KEGG pathway database. PubViz retrieves pathways from KEGG containing genes related to user query, and, as with retrieved results in the previous example, it lists them under the search Gene ID term. The user simply needs to click on the pathway of interest, and the KEGG pathway displays in the Pathway View window. By being able to see other genes in the pathway during this literature search without having to go out of the exploration environment, the user can immediately implement new ideas for additional Medline exploration.

**MeSH tree view:** This final example draws on the pre-calculated MeSH concept similarity we described in section 2.3. Working now from the MeSH side panel tab (depicted in Fig. 5 separately from the rest of the screen only to show it more clearly here) the user enters a MeSH term, PubViz returns the top N (specified by the user) related concepts as well as underlying Medline citations, which are not shown here but detailed in the lower MeSH Profiling data tab. They are sorted by year or other user chosen criteria. Additionally, MeSH connections are drawn on a MeSH View canvas, also not depicted here but placed in the center where views for all tabbed displays reside. This example shows how

users get multiple views alongside each other – a tree hierarchy, detailed data on citations, and networks based on similarity relationships. When users search for MeSH terms or click on a particular MeSH term in Mesh search panel, PubViz shows the MeSH topic in a hierarchical MeSH tree view (Fig. 5) for users to review or issue further search request.

### 3.4. Search History Tracking

Interactive search empowers user to search using a combination of criteria. Meanwhile, it also raises the question of how users can track their exploration history. From the side panel User Profile tab, PubViz allows users to set up individual accounts, and it automatically records their use history including the parameter settings (Fig.6). Therefore, users can quickly replicate or continue their precious analysis or download/upload data or literature sets.

## 4. DISCUSSION

In summary, PubViz is designed to be an efficient Medline literature exploration interface for understanding the biological implications of high throughput data. While some of the existing Medline search solutions provide gene or gene-ontology centered graphic layout<sup>3,5,7,10,13,15</sup>, few of them provide powerful interactive visual Medline exploration.

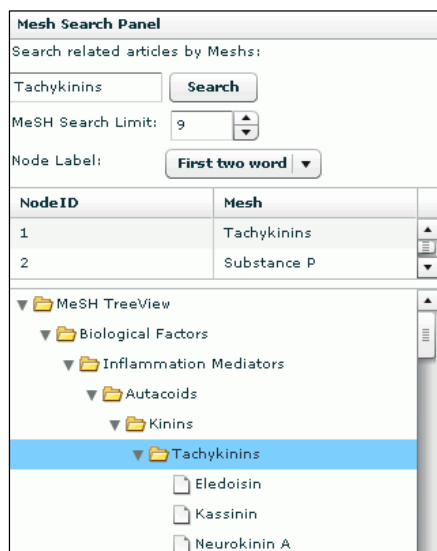


Fig. 5. MeSH search panel

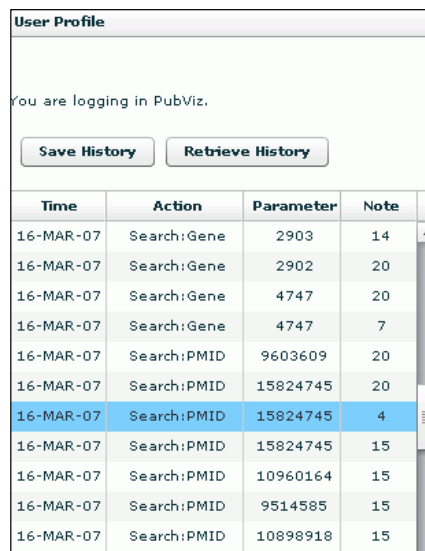


Fig. 6. User history

PubViz is distinct by providing the combination of multiple topic-centered graph layout and powerful data charting tools for efficient and flexible visual Medline exploration. In PubViz, switching between charts and underlying tabulated data just requires a single mouse click. These visualization tools enable users to grasp “big picture” of their query results, something they cannot do from results displayed in tabular form. Moreover, by combining graphic displays with capabilities for interactively selecting and filtering data, drilling down to details, and sorting, these visual data exploration tools enable users to quickly recognize and uncover hidden patterns and data of interest in Medline database. We hope the ability to translate data patterns into insights makes PubViz a highly effective literature exploration tool.

The FLEX 2.0 technology offers the possibility of rich internet application with performance approaching typical desktop programs. In traditional HTML-based web applications, in contrast, when a user click on one responsive element on a web page, usually the whole page will be resent from the server and then refresh at the client side, which usually delays the response. In PubViz, however, all web services are invoked using asynchronous calls, and most interactions on retrieved data are handled by the PubViz on the client side. It enables us to write functions to handle user interactions on complex graphs and issue more dynamic search requests to help users to drill deeper, navigate faster, and understand better.

Without doubt, our prototype only provides a framework that can be greatly improved on. Besides various new functions, a key issue we need to work on is the usability of PubViz. At this moment, the organization and presentation of data and function is certainly not optimal. We plan to conduct systematic usability studies to improve the PubViz interface to make it an efficient tool for literature exploration.

A key capability we want to add into PubViz is the use of external knowledge and information to improve Medline record retrieval. Essentially, external knowledge can be used to modify the semantic distance between different concepts thus change the similarity among different documents as well as the similarity between query terms and Medline records. For example, the use of external protein-protein interaction data can make the similarity of gene/protein names dependent on how different protein interact with each other rather

than treat each gene/protein name as independent of each other. If we assign high similarity to gene/proteins that have direct interaction with each other and incorporate such similarity information in the Medline record retrieval process, we will be able to obtain Medline records not only containing the query gene/protein names, but also those containing gene/protein names that directly interact query gene/protein names. Besides protein interaction information, we hope to include different areas of knowledge and experimental data, such as linkage disequilibrium relationship among cytochrome, SNP, STS/microsatellite marker and genes, co-regulated genes from microarray study, neuroanatomical circuits described in textbooks, etc., in the PubViz system. As a result, PubViz can greatly increase the efficiency of Medline data exploration in unfamiliar territories.

Since it is impossible for a single group to incorporate all potentially useful external knowledge sources or display functions in PubViz, we plan to build standard interfaces to allow interested researchers to add their own concept similarity matrix, web services for collecting external information and visualization tools. In the long run, we hope PubViz to become a highly extensible system for exploring Medline and other free text databases.

## Acknowledgements

W. Xuan, M. Dai, S. J. Watson and F. Meng are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. This work is also partly supported by the National Center for Integrated Biomedical Informatics through NIH grant 1U54DA021519-01A1 to University of Michigan.

## References

1. Chang, J.T. *et al.* (2004) GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics* 20 (2), 216-225
2. Chen, H. and Sharp, B.M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 5, 147
3. Doms, A. and Schroeder, M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* 33 (Web Server issue), W783-786

4. Hoffmann, R. and Valencia, A. (2004) A gene network for navigating the literature. *Nat Genet* 36 (7), 664
5. Homayouni, R. *et al.* (2005) Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics* 21 (1), 104-115
6. Jayapandian, M. *et al.* (2007) Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. *Nucleic Acids Res* 35 (Database issue), D566-571
7. Jenssen, T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 28 (1), 21-28
8. Johnson, H.L. *et al.* (2005) Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. In *Proceedings of the Pacific Symposium on Biocomputing (PSB) 2006*
9. Kanehisa, M. (2002) The KEGG database. *Novartis Found Symp* 247, 91-101.
10. Landauer, T.K. *et al.* (2004) From paragraph to graph: latent semantic analysis for information visualization. *Proc Natl Acad Sci U S A* 101 Suppl 1, 5214-5219
11. Lin, S.M. *et al.* (2004) MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics* 20 (18), 3659-3661
12. McCarthy, M. (2006) Allen Brain Atlas maps 21,000 genes of the mouse brain. *Lancet Neurol* 5 (11), 907-908
13. Mudunuri, U. *et al.* (2006) botXminer: mining biomedical literature with a new web-based application. *Nucleic Acids Res* 34 (Web Server issue), W748-752
14. Nvidia. (2007) Nvidia CUDA: Compute Unified Device Architecture. [http://developer.download.nvidia.com/compute/cuda/0.8/NVIDIA\\_CUDA\\_Programming\\_Guide\\_0.8.pdf](http://developer.download.nvidia.com/compute/cuda/0.8/NVIDIA_CUDA_Programming_Guide_0.8.pdf)
15. Plake, C. *et al.* (2006) ALIBABA: PubMed as a graph. *Bioinformatics* 22 (19), 2444-2445
16. Rinaldi, F. *et al.* (2007) Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine* 39 (2), 127-136
17. Sharma, P. *et al.* (2006) Mining literature for a comprehensive pathway analysis: a case study for retrieval of homocysteine related genes for genetic and epigenetic studies. *Lipids Health Dis* 5, 1
18. Tanabe, L. and Wilbur, W.J. (2002) Tagging Gene and Protein Names in Biomedical Text. *Bioinformatics* 18 (8), 1124-1132
19. ThomsonResearchSoft. (2005) RefViz. <http://www.refviz.com/rvinfo.asp>
20. Xuan, W. *et al.* (2005) GeneInfoMiner--a web server for exploring biomedical literature using batch sequence ID. *Bioinformatics* 21 (16), 3452-3453
21. Yuryev, A. *et al.* (2006) Automatic pathway building in biological association networks. *BMC Bioinformatics* 7, 171