SRS Browser: A Visual Interface to the Sequence Retrieval System

Ketan K. Mane [a] and Katy Börner[1 a]

[a]School of Library & Information Science, Indiana University 10th Street and Jordan Avenue
Bloomington, IN 47405, USA.

ABSTRACT

This paper presents a novel approach to the visual exploration and navigation of complex association networks of biological data sets, e.g., published papers, gene or protein information. The generic approach was implemented in the SRS Browser as an alternative visual interface to the highly used Sequence Retrieval System (SRS) [1]. SRS supports keyword-based search of about 400 biomedical databases. While the SRS presents search results as rank-ordered lists of matching entities, the SRS Browser displays entities and their relations for interactive exploration. A formal usability study was conducted to examine the SRS Browser interface's capabilities to support knowledge discovery and management.

Keywords: Information visualization, visual interfaces to digital libraries, Sequence Retrieval System, SRS

1. INTRODUCTION

Today, humanity's biomedical knowledge is stored in diverse formats (e.g., xml, flat file, relational format) in various biological, biochemical and biomedical databases (e.g. NCBI, SwissProt, DDBJ, Uniprot) around the globe. Search tools such as the Sequence Retrieval System (SRS) [1] or the NCBI Data Mining Tools (e.g., Map Viewer, LocusLink, etc.) support the retrieval of entities based on type (e.g., paper, gene, protein), disorders (e.g., melanoma), genome sequence, structures, taxonomy, etc. The search tools utilize a repository of customized parsers for querying a multitude of databases. Results are typically presented in a tabulated, rank-ordered format that resembles the display of Google search results.

This paper describes a system named 'SRS Browser' that acts as an alternative visual interface [2] to SRS. The SRS Browser aims to support its users in answering questions such as: Which proteins are associated with a particular disease? What relations exist among genes that are known to cause a certain disease? What is the most researched gene in a certain research area? What papers refer to what genes and/or proteins related to a certain disease? What papers complement the information available in a paper of interest? Thus, it supports the so-far little addressed need to understand the intrinsic relationships among biological entities.

Users of the SRS Browser submit a SRS query via the standard SRS interface. Instead of presenting the user with a rank-ordered list of query results (cf. Figure 1), the SRS Browser analyzes the query result for existing relationships and presents the user with different association networks (cf. Figure 3). The networks can be interactively explored and diverse details can be retrieved.

The remainder of this paper is organized as follows: Section 2 discusses related work. Section 3 introduces the SRS. Section 4 details the SRS Browser's system architecture, data flow, data parsing, association identification and xml file processing for visualization. Section 5 discusses sample usage scenarios in which the SRS Browser is employed to find co-occurring genes, papers that deal with similar genes and proteins, and highly interlinked proteins. Section 6 presents the results of a usability study that aimed to evaluate the interface design and functionality of the SRS Browser. Section 7 concludes the paper with a discussion of the results and future work.

---

[1] katy@indiana.edu; phone 1 812 856-3256; fax 1 812 855-6166; Indiana University

Mane, Ketan K. and Börner, Katy. (2006) SRS Browser: A Visual Interface to the Sequence Retrieval System.
SPIE Conference on Visualization and Data Analysis,
San Jose, CA, Jan 15-19, Vol. 6060, 606010: pp. 1-11.

2. RELATED WORK

Different approaches exist to derive meaningful concepts and associations from datasets. Early research focused on using natural language processing (NLP) [3, 4] to extract biological concepts from papers, typically from title, abstract and full-text data. Among others, NLP techniques make use of feature terms (e.g., receptors), core terms (e.g., U3-55k), functional dependency grammar [5], Hidden Markov Models [6], or Bayesian Classifiers [7] to build a local dictionary. The dictionary terms are then used as regular expressions in the identification of associations. Associations can also be discovered using clustering techniques. For example, the Library of User-Oriented Concepts for Access Services (LUCAS) [8] extends the NLP technique by applying inverse document frequency to rank protein names. Subsequently, latent semantic analysis [9, 10] is applied to discover protein document associations. Further clustering of related proteins is achieved through a maximum-distance algorithm [11]. The resulting protein association networks are visualized using a force directed layout algorithm. However, the NLP-based techniques are computationally intensive and unsuitable for large data sets.

Another line of research applies statistical techniques to match, e.g., Medline paper (titles and abstracts) and user-defined keywords to preserve and draw networks of functionally related genes [12]. Statistical distributions like binomial ones have been used to identify strong correlations between genes and diseases based on research reported in papers [13].

Association visualization aims to communicate the results of the association discovery process. Today, there exist a number of tools that support data-mining and association discovery among biological datasets. The LION Target Engine by LION bioscience Inc. supports text-mining, display of genome information, the ability to find related protein structures, etc. in a tabulated form. ArrowSmith [14] is a tool that establishes associations among Medline papers based on the co-occurrence of title words. The result is displayed as a list of similar papers. Tools like MedlineR [15] can be used to retrieve data from Medline and to extract co-occurrence information based on terms occurring in retrieved documents. Pajek [16] can be used to generate static visualizations of paper association networks. Chilibot[2] makes use of the visualization toolkit aiSee[3] to depict the number of retrieved search results.

The EMBL-EBI bioinformatics research lab designed the visualization tool BioLayout [17] that parses BLAST output files and visualizes the network using a modified form of the Fruchterman Rheingold algorithm [18].

GeneNet [19, 20] is a local database that provides information on gene networks. Networks can be visualized using the GeneNet Viewer. Color is used to indicate the functional state of the genes. Detailed information on a gene of interest can be acquired by querying the GeneNet database. However, the gene count in this database is restricted to entries added by the GeneNet group.

Recent work in the area of mapping knowledge domains [21] aims to visualize ecologies of different biological entities. However, the large scale visualization of the diverse relationships, e.g., paper-gene, gene-gene, gene-protein, etc., are based on canned data rather than interactively retrieved data.

To our knowledge, there exists no system that enables users to cross-search major databases, automatically mine the search results and automatically identify associations among multiple entity types, e.g., papers, genes or proteins, and that further communicates the results via highly interactive visualizations of the derived network ecologies.

3. SEQUENCE RETRIEVAL SYSTEM (SRS)

The SRS system is a software package, currently distributed by LION Bioscience Inc. It is a widely used tool for cross-searching different biological databases (e.g., NCBI, SwissProt, Uniprot) and support diverse data formats, e.g., xml, flat file or relational format. About 400 different databases are supported by SRS.

Note that usage of the SRS requires the creation of a local copy of the databases. At Indiana University, Bloomington, we use the IUBio-SRS[4] system which provides access to about 80 biological, biochemical, and biomedical databases.

Users of the SRS system enter a search term and search parameters. Matching entities, e.g., proteins are displayed as a rank ordered list.
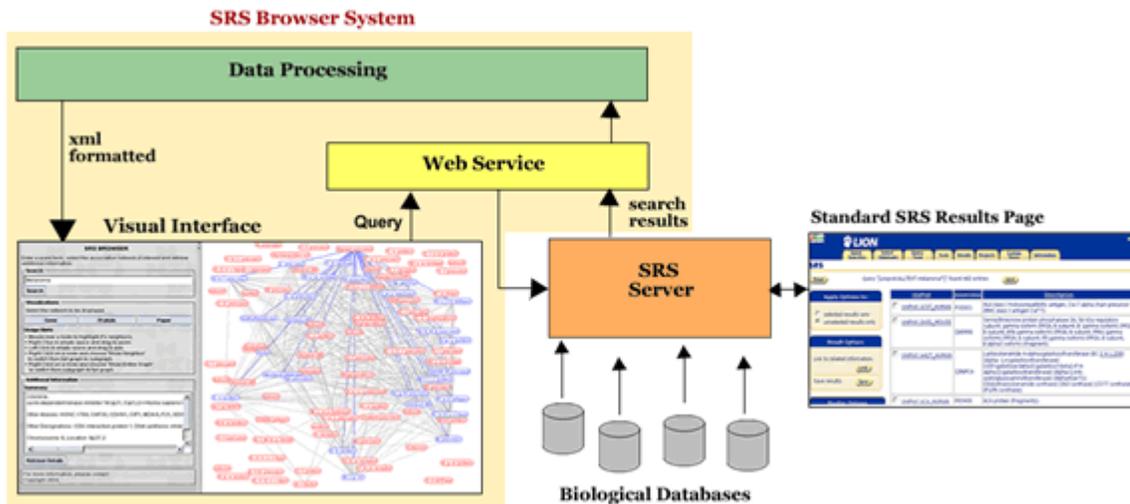
---

[2] Chilibot is accessible at http://www.chilibot.net/

[3] aiSee is accessible at http://www.aisee.com/

[4] Access IUBio-SRS at – http://iubio.bio.indiana.edu/srsbin/cgi-bin/wgetz?-page+srsq2+-noSession

SRS Browser System

Data Processing

xml formatted

Web Service

Visual Interface

Query

search results

SRS Server

Standard SRS Results Page

Biological Databases

Mane, Ketan K. and Börner, Katy. (2006) SRS Browser: A Visual Interface to the Sequence Retrieval System.
SPIE Conference on Visualization and Data Analysis,
San Jose, CA, Jan 15-19, Vol. 6060, 606010: pp. 1-11.

The Java programming language was used to develop the visual interface and to query SRS using web-services. Perl's regular expression matching is employed for parsing search results obtained using SRS. Subsequently, we explain the web service, data processing, and visual interface components in more detail.

4.1. Web service

The Genome Directory System (GDS)[5] is an interface used to access the SRS system via SOAP web services. GDS was developed as part of the Generic Model Organism Database (GMOD)[6] project. It includes Apache Axis[7] to support Java SOAP implementations and Apache Tomcat[8] as an application server to run the Apache Axis code base.

GDS queries can be submitted by any client that supports SOAP-based web services. Query results are obtained in xml or FASTA format. All data transfer between client and SRS is done using standard HTTP connections.

Currently, the SRS Browser is customized to query the indexed Uniprot database available within IUBio-SRS. Query match is performed against specified fields within SRS. Data on genes, proteins and paper PubMed-ids are retrieved for all entities that match a given query. Only the first 100 result entries are displayed to ensure legible visualizations.

4.2. Data processing

The xml-formatted results from the web-service program are converted into delimited text format and parsed for existing relationships between data elements, e.g., gene names, protein names, and papers (represented by PubMed-ids). In the current implementation, three types of relations are established: (1) gene-gene (2) protein-protein and (3) paper-paper. Two genes are assumed to be related if they are mentioned in the same paper. Analogously, two proteins are assumed to be related if they are mentioned in the same paper. As for papers, it is assumed that they are interrelated if they refer to the same gene or protein. Alternative measures could be easily implemented. The set of resulting entities and their relations are then converted to an xml-format as input for visualization.

4.3. Visual interface

The client-side applet interface of the fully functional SRS Browser is shown in Figure 3. The interface was developed using Java Swing API. It consists of a query panel (left) and a visualization panel (right). The query panel comprises a text box to submit queries to SRS, a selection menu to quickly switch between association graphs of gene-gene, protein-protein and paper-paper association networks, a 'Retrieve Details' button to request more detailed information on data entities, as well as instructions on how to interact with the visualization.

Users can enter a query term via the query panel to start the fully automatic download, parsing, association identification and the visual display of a data set of interest. They can select the association graph that is most relevant for them and acquire additional information on data entities.

The visualization panel uses the random layout algorithm implemented in the Prefuse API to layout small- and medium-sized association networks in a two-dimensional space. If a user requests a gene-gene association map, the nodes represent genes. Two genes are connected by an edge if they were mentioned together in one and the same paper.

The Prefuse API was also used to facilitate zooming and panning, node selection and the highlighting of neighboring nodes. Tooltips were implemented to display the full name of a data entity in case of densely packed networks that cause partially truncated node names.

The zooming functionality helps users to acquire an overview of the retrieved query result. Panning, active whenever the left mouse button is held down, enables users to access different sections within the visualization.

To ease the examination of potentially highly connected networks, we implemented a novel way to show the immediate neighbors (nodes connected to a selected node) and their relations only, see Figure 4. A user simply right clicks a node of interest to bring up a selection panel and selects 'Show Subgraph' to switch from full-graph to a sub-graph. Multiple nodes can be selected this way and the networks will be pruned to show those nodes, their neighbors, and relationships exclusively.

Clicking on a node displays additional information on this data entity in the left query panel. The 'Retrieve Details' button enables easy access to the original web sites. For example, retrieving details on a paper node opens the respective

---

[5] Additional information on GDS is available at http://www.gmod.org/gds/gds_summary.shtml

[6] Access GMOD website at http://www.gmod.org

[7] Access Apache-Axis website at http://ws.apache.org/axis/

[8] Access Apache-Tomcat website at http://jakarta.apache.org/tomcat/

Additional Information

Retrieve Details

Zoom

Highlight Neighbors

Query for Melanoma

Association among melanoma related genes

PRAME related genes are highlighted in blue. Missing gene CDKN1 is highlighted in red

Genes related to CDKN1A gene are highlighted in red
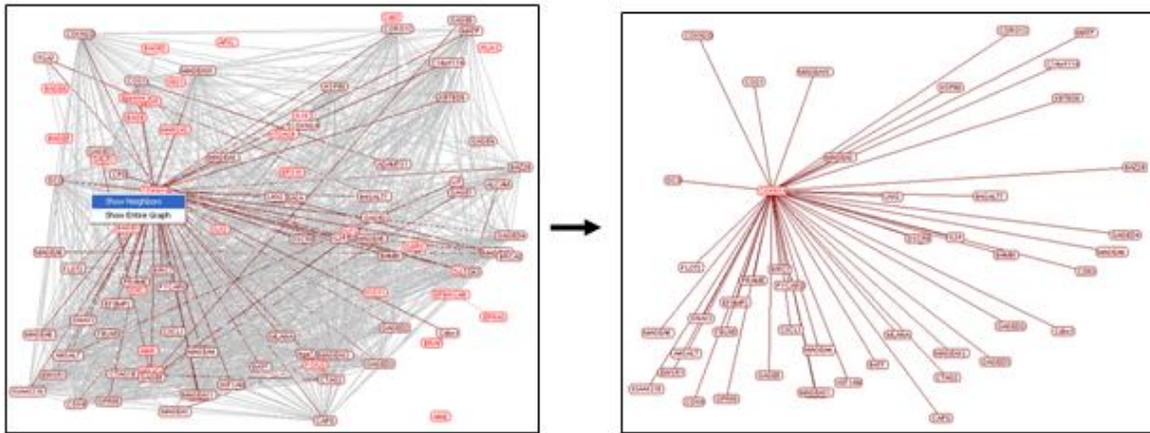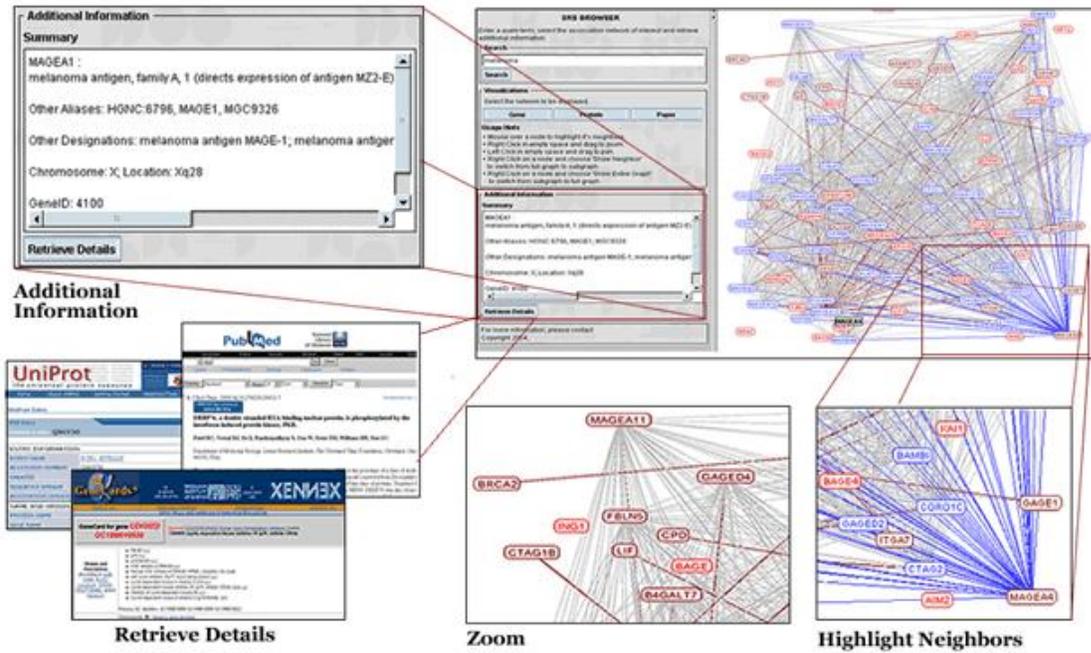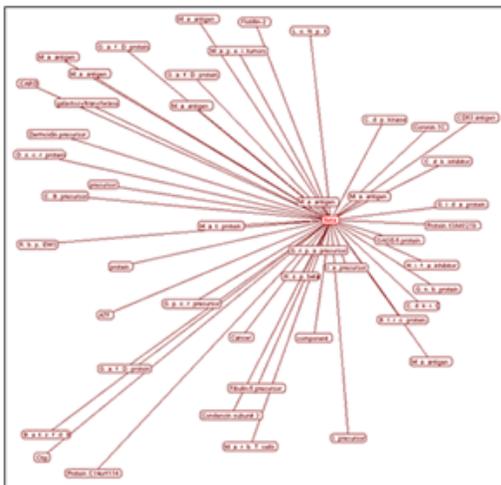
Cluster for 'Beta-1, 4-galactosyltransferase 7' protein.

Cluster for 'Integrin alpha-7 precursor' protein

neighboring nodes or to display the entire network, to access gene-summaries and to retrieve details from different databases.

Subsequently, each participant was asked to use the SRS Browser to complete the steps below and to answer the listed questions (given in bold).

1. In the search box, type the term 'Melanoma' and hit 'Search'.
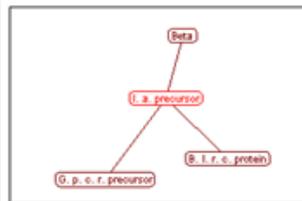
2. Display the gene, protein and paper network visualizations.

3. In the gene network visualization

   a. **Identify and list all y genes related to GAGE6.**

   b. **Filter the gene network visualization. List any five immediate neighbors of GAGE5.**

4. In the protein network visualization, find GAGE-5 protein and retrieve its details from Uniprot. **Write down the 'Accession Number' available from the Uniprot website.**

5. Select any paper from the paper network: **Paper Number 1: ……………..**

6. Filter the network to show immediate neighbors of selected paper (number1) and select one neighbor. **Paper Number 2: ……………..**

7. Retrieve information from PubMed for those two papers, **how are these two papers related? …………………………………………………………………………………..**

After the task list was completed, a post test questionnaire was handed to the participants. Participants were asked to respond to questions 1-5 using a Likert scale of 1-5, where 1 represents strongly disagree, while 5 is used to represent strongly agree. The questions were designed to acquire information about the efficiency, accuracy and speed of the browser. In addition, we were interested to obtain feedback on the positive and negative features of the SRS Browser. The list of post test questions is given below.

1. The SRS Browser functionality is easy to determine

2. The interface is easy to use

3. SRS Browser provides fast results

4. SRS Browser provides accurate results

5. It was easy to find the information that I needed to complete the task

6. What did you like most about the SRS Browser?

7. What did you like least about the SRS Browser?

8. Do you have any recommendations to improve the SRS Browser?

All participants received a $5 gift coupon upon completion of the task.

6.2. Results

Analysis of pre-test questionnaire revealed the following: 7 men and 5 women participated in the usability test. All participants were in age-group 21-30 years except one which was 18-20 years. Out of the total count of 12, five participants have higher degrees (MS, Ph.D.) in biology and hence are regarded as biology experts. One of the experts works as an administrator for the SRS system available on the IUBio server in the biology department at Indiana University. The remaining seven participants were students from the Biology department and the Bioinformatics track of the School of Informatics at Indiana University. The question: 'Have you used any tool that involves information visualization?' was answered with 'No' by 7 participants. The other participants listed tools such as CATPA, SpotFire, Cytoscape but also general network analysis and visualization tools such as Pajek and UCINet. The question: 'Have you

used any software which can search online biological database' was answered with 'Yes' by all participants. Systems listed comprise NCBI tools, Swissprot, Unigene, etc.

In the usability study, all participants performed all tasks successfully (average time = 8 min) except for one which was not able to find the protein and hence could not continue to find the Uniprot accession number required to answer question 4 from the task list.

The post-test questionnaire was analyzed to gain insight into the interactive experience with the SRS Browser (questions 1-5). The average scores were as follows: Q1 (4.42); Q2 (4.25); Q3 (4.00); Q4 (3.92); Q5 (4.33). The low average for Q3 is mainly due to a very low score by one of the participants. S/he experienced a delay in receiving results from the IUBio server (see section 4.1). A closer examination of the individual scores for Q4 revealed that many students reported a neutral (score = 3) while experts gave higher scores.

Asked about what they liked most about the SRS Browser, all participants had positive feedback on the integration of the diverse features within the SRS Browser system. Most users reported that the tool was easy to use and provided ability to perceive different information simultaneously. One of the experts described it as a "useful tool as it consolidates important and disparate pieces of information". Other expert pointed out that since the tool shows relational information, it is easy to search for relevant terms within the results. One expert pointed out that this tool is more informative than other search engines.

Asked what they did like least, participants suggested not to use abbreviated names for proteins that we had employed to avoid clutter. Only one participant reported an inability to locate a protein name required to answer question 4 from the task list.

The participants made a number of recommendations on how to improve the SRS Browser. One requested an alternative search box to find a specific gene, protein or paper in a complex network. The matching gene/protein/paper could simply be highlighted to ease identification. Another participant suggested to also show neighbors of neighbors to provide further depth in relational information. A third recommendation relates to the identification of information that is preserved across different networks. For example, if a gene sub-network is selected, then upon switching to protein and paper network only the relevant proteins and papers could be displayed directly instead of the entire network. We intend to incorporate these recommendations in future versions of the SRS Browser.

7. DISCUSSION AND FUTURE WORK

This paper presented the SRS Browser, an alternative visual interface that extends SRS by association discovery and interactive visualization.

By utilizing SRS as a search engine, relevant data entities are already tagged, thereby eliminating the need to rely on NLP techniques to identify entities of interest. Given that SRS comprises of all the data from a particular database (e.g., Uniprot) there is no need to maintain a separate database like GeneNet.

The visualizations succeed in helping users to understand and examine the global association networks of diverse data entities, e.g., genes, proteins, or papers, and also provide easy access to details such as node attributes or neighboring nodes. Filtering out neighboring nodes as a novel means to reduce the complexity and clutter of network visualizations was identified to be of high value for expert users.

The results of a formal usability study show that the interface is easy to learn and use in the context of biologically relevant tasks.

Future work will comprise the extension of the SRS browser to other databases, usage of the new GDS package (see section 4.1) to include rank-ordering and the representation of edge weights – indicating how often entities co-occurred – via edge thickness will be explored. We are also in the process of implementing additional network analysis methods, e.g., the identification of highly interconnected nodes or weak links and visual representations of results to further augment the identification of major network features. The coupling of network analysis and visualization might also prove valuable for increasing the scalability of the visual browsing interfaces - nodes of low relevance can simply be omitted in the visualization, yet they can be examined if the user decides to zoom into a particular area of interest or is s/he selects a certain node and queries for its neighbors.

ACKNOWLEDGEMENTS

REFERENCES

1. Etzold, T., A. Ulyanov, and P. Argos, *SRS: information retrieval system for molecular biology data banks.* Meth. Enzymol., 1996. **266**: p. 114-128.
2. Börner, K. and C. Chen, *Visual Interfaces to Digital Libraries*, ed. C.E. Chen. 2002: Springer Verlag, LNCS 2539.
3. Eriksson, G., K. Franzen, F. Olsson, L. Asker, and P. Liden, *Exploiting syntax when detecting protein names in text.* In Proceedings of Workshop on Natural Language Processing in Biomedical Applications, 2002.
4. Fukada, K., et. al., *Towards information extraction: Identifying protein names from biological papers.* PSB, 1998: p. 707-718.
5. Tapanainen, P. and T. Jarvinen, *A non-projective dependency parser.* In Proceedings of 5th Conference on Applied Natural Language Processing, 1997: p. 64 -71.
6. Collier, N., et. al., *Extracting the names of genes and gene products with a hidden markov model.* International Conference of Computational Linguistics, 2000. **18**: p. 201-207.
7. Wilber, W., et. al., *Analysis of biomedical text for biochemical names: A comparison of three methods.* AMIA Symp, 1999: p. 176-180.
8. Yeuyu, F., J. Mostafa, and K. Seki, *Protein association discovery in biomedical literature.* International Conference of Digital Libraries, Proceedings of the third ACM/IEEE-CS joint conference on Digital libraries, 2002: p. 113-115.
9. Berry, M., S. Dumais, and T.A. Letsche, *Computational methods for intelligent information access.* in Proceedings of SuperComputing '95, 1995.
10. Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, *Indexing by latent semantic analysis.* Journal of the American Society for Information Science, 1990. **41**(6): p. 391-407.
11. Tou, J. and R.C. Gonzalez, *Pattern Recognition*. 1974: Addision-Wesley.
12. Wilkinson, D.M. and B.A. Huberman, *A method for finding communities of related genes.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(suppl 1.): p. 5241-5248.
13. Adamic, L.A., D. Wilkinson, B.A. Huberman, and E. Adar, *A Literature Based Method for Identifying Gene-Disease Connections.* CSB '02, 2002: p. 109-118.
14. Smalheiser, N.R. and D.R. Swanson, *Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypothesis.* Computer Methods and Programs in Biomedicine, 1998.
15. Lin, S.M., P. McConnell, K.F. Johnson, and J. Shoemaker, *MedlineR an open source library in R for Medline literature data mining.* Bioinformatics, 2004.
16. Batagelj, V. and A. Mrvar, *PAJEK-A Program for Large Network Analysis.* Connections, 1998. **21**: p. 47-57.
17. Goldovsky L, I. Cases, A.J. Enright, and C.A. Ouzounis, *BioLayout: An automatic graph layout algorithm for similarity and network visualization.* Bioinformatics, 2004. **17**(9): p. 853-4.
18. Fruchterman, T.M.J. and E.M. Reingold, *Graph Drawing by Force-Directed Placement.* Software-Practice & Experience, 1991. **21**(11): p. 1129-1164.
19. Kolchanov, N.A., et al., *Gene Networks.* Molecular Biology (Msk.), 2000. **34.**(4): p. 449-460.
20. Kolpakov, F.A., E.A. Ananko, G.B. Kolesov, and N.A. Kolchanov, *GeneNet: a database for gene networks and its automated visualization.* Bioinformatics, 1998. **14**(6): p. 529-537.
21. Börner, K., C. Chen, and K.W. Boyack, *Visualizing knowledge domains.* Annual Review of Information Science & Technology, 2003. **37**: p. 179-255.
22. Jeffrey, H., S. Card, K., and J.A. Landay, *prefuse: a toolkit for interactive information visualization.* Submitted paper draft, April 2004.
23. Rebhan, M., V. Chalifa-Caspi, J. Prilusky, D. Lancet, and, *GeneCards: encyclopedia for genes, proteins and diseases.* Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel), World Wide Web URL: http://bioinformatics.weizmann.ac.il/cards, 1997.