An Experiential Approach to Interacting with Biological Information

Naureen Moon, Bibek Dev Bhattarai, and Rahul Singh

Department of Computer Science, San Francisco State University
San Francisco, CA, USA
{numoon, bdb}@sfsu.edu, rsingh@cs.sfsu.edu

Abstract. Technological advancements in the life sciences have enabled biologists to generate high volumes of heterogeneous, multimedia data. The challenge today lies in correlating and integrating the information in ways that promote a holistic understanding of the underlying biological phenomena. This paper presents our research in designing an experiential information interaction environment for query-exploration of complex biological information. This entails development of a unified presentation-query-exploration environment that incorporates and relates multimodal data repositories and views. Specifically, our interface captures textual, sequence, and structural views of biological entities and presents semantic correlations between them using Gene Ontology annotations. Additionally, the system extracts and displays the spatial-temporal characteristics of the data to facilitate querying and discernment of relationships. The different views of the data are interactive, in order to facilitate information exploration and assimilation. Experiments and examples demonstrate the system's features, efficacy, and ability to facilitate both concept discovery and information querying.

1 Introduction

Recent technological advancements in the life sciences have enabled researchers to generate vast quantities of data about various bio-chemical entities and phenomena from genes to therapeutic drugs. These technologies have also induced a major paradigm shift in biological research. Gone are the days of detailed analysis focused on a few entities in isolation. Instead the research paradigm today seeks a more holistic view and is based on a multitude of experimental methodologies which probe the domain of interest from multiple perspectives. The various types of information generated as a consequence present interesting challenges towards development of systems meant to support querying, interaction, exploration, and assimilation of such data. This challenge has striking overlaps with issues in management of physical and logical heterogeneity which is currently a central focus of the multimedia and database research communities. Consider for example the problem of developing new therapeutics. The necessary research today requires among others: (a) exploration of the available literature in the area to ascertain, for instance, disease-gene relationships, (b) exploration/query/analysis of the associated genes, and (c) research

G. Bebis et al. (Eds.): ISCV 2006, LNCS 4292, pp. 1586 – 1595, 2006.

involving structural information associated with the disease (such as structure of proteins, enzymes, and small molecules associated with the disease). Even though the aforementioned steps cover but a portion of the associated research spectrum, facilitating these steps requires solving a significant set of critical problems which include:

- Access to multiple repositories each with their own storage structure and logic. Additionally such systems must support handling the large volumes of data present in these repositories.
- Support for interacting with heterogeneous data types, such as textual data, genomic data, and structural data.
- Support for users to assimilate this complex information by exploiting semantic correlations in the data, preserving data and user context, supporting efficient query-retrieval, and aiding in data exploration.

The past few decades have witnessed significant development in databases for life sciences. One comprehensive collection of repositories is the set of NCBI databases [1], which include PubMed, a literature index containing over 15 million citations. NCBI also maintains databases of non-textual information, among them, databases containing protein sequence and structure information, such as Protein and Structure. These databases may be searched, singly or collectively, through the Entrez web query interface [2], whereby results for each database are generated separately and accessible through hyperlinks. The retrieved documents on each results page are also accessible through sets of hyperlinks, and frequently contain links to related information in the other databases.

The development of the NCBI resources has been a significant milestone in supporting access to information related to the life sciences. Further, as part of this effort, paradigms and interfaces to interact with specialized data types (genomic and structural data) have also been developed. However, these solutions only partially address the challenges enumerated earlier. For instance, while the Entrez interface does support unified (keyword-based) querying across multiple repositories, it provides minimal functionality in terms of utilizing semantic correlations that may exist between hits originating from within a single repository as well as those originating from different repositories. This results in deluging a user with data and the user typically has little recourse beyond following each link. The hyperlink-based information traversal strategy often leads to context switching, thus increasing the cognitive load on users [3]. Additionally, there is no effective mechanism by which to search for multiple entities simultaneously to determine relationships or similarities between them. Furthermore, there is no provision to utilize meta-data such as the temporal or spatial characteristics of the information, which can not only be useful to support rapid access [4], but also to explore spatial-temporal and evolutionary trends in the data. Compounding the problem is the sheer volume of data researchers may encounter. For example, a biologist searching for a particular gene on PubMed will often obtain results in the hundreds or thousands. However, it is possible and beneficial to utilize the wealth of data from the NCBI databases as well as others to reveal substantial latent biological information, as we describe below.

2 Research Approach

The paradigm of experiential computing has recently been proposed [3], in multimedia research to develop systems that support assimilation of complex information. Such systems are characterized by: (1) Directness in terms of interaction with the data, (2) Unified query and presentation spaces, (3) Support for user and data contexts, (4) Presentation of information independent of the source, and (5) Facilitation of perceptual analysis and exploration. We note that the characteristics of experiential systems correspond well with many of the issues we have outlined earlier. For instance, relevant biological information can easily reside in multiple sources. Additionally, such information inherently includes interrelationships between entities (e.g., protein interactions and gene-disease connections) as well as multiple modalities (such as textual, genomic, and structural). This makes desirable presentation of data in manners that are source-independent and that simultaneously display and correlate heterogeneous data types. Moreover, the support for information assimilation espoused in experiential systems through direct interactions with the data, context support, and facilitation of perceptual analysis are essential to interacting with complex data such as that in the biological domain.

In this paper, we present the results of our ongoing investigations in developing an experiential system for managing and interacting with information from research in the life sciences. Towards this end, we propose a unified presentation-exploration-query interface that encompasses various modalities of biological data in a reflective manner. Specifically, we provide for a multiple perspective "view" of the information by combining textual data from the literature with Gene Ontology (GO) [5] code assignments and gene/protein sequence data, as well as a graphical representation of relationships and any corresponding structural information (Fig. 3). The various views are *reflective*, that is, that interactions in one view are instantaneously reflected in the other views. This serves not only to mimic real-life interactions (and thereby to allow users to maintain context) but also to provide a perspective on the interrelationships that exist in the information. Additionally, we introduce useful meta-information in the form of spatial-temporal characteristics of the information that display location and publication dates of the retrieved documents. This can be used to support intuitive access to the information as well as analyze the development of knowledge. Finally, our approach uses available manually-assigned GO codes of entities to reveal semantic correlations between them according to a fixed vocabulary of biochemically relevant concepts.

A number of techniques have been developed for automated processing of the literature in order to extract meaning from the vast quantity of available information. These efforts include document clustering [6] and extraction of gene and protein names and interactions [7,8,9,10]. A typical shortcoming of such text-analysis systems, however, is that they are standalone, namely that they do not extend beyond relationship extraction to encompass effective presentation of results, much less interaction. The TransMiner system [11] overcomes this limitation by combining extracted relationships with a graph visualization tool to suggest indirect relationships between entities. While it does include effective visual representation of relationships, interaction is minimal, being limited to user selection of association strength thresholds. The

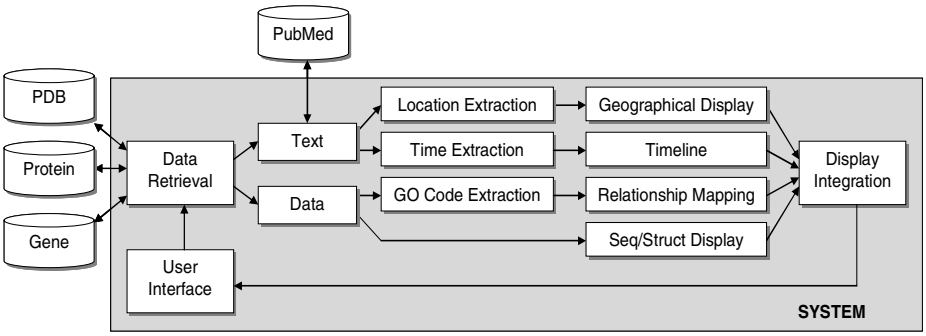

Fig. 1. The system architecture

ALFA architecture [12] enables user-guided information extraction from literature as well as non-textual data in an integrated representation. While it includes an impressive suite of tools for user interaction, its emphasis on user-driven selection of information precludes exploration and full reflectivity of the various components of the interface.

Our system differs most fundamentally from related research due to its combination of attributes, namely multiple perspective information search/exploration (by text (literature), sequence, and structure), support for text-based biological information analysis, inclusion of spatial and temporal characteristics of the information, andsupport for experiential user-information interactions. The proposed approach seeks to underline relationships in the data and facilitate its exploration through reflective interfaces that provide interlinked multiple-perspective views of the information. Such a system can therefore not only be used for query-retrieval, but also for information analysis, and ultimately for hypotheses generation and information exploration. Integrated support of this type is especially necessary given the modern paradigm of systems biology that seeks to span disciplines in an environment of increasing specialization.

3 System Description

The system consists of five major modules: Data Retrieval, Relationship Extraction/Display, GO Code Extraction/ Mapping, Location Extraction/Display, and Time Extraction/Display (Fig. 1). Most data is retrieved through dynamic access of the NCBI databases, PubMed, Gene, and Protein, in particular, using the efetch utilities [13]. Protein structures are obtained from PDB (Protein Data Bank) [14].

The system is initialized through entity identifier-based queries to the user interface. These may be one or more Gene IDs or Protein IDs. The Data Retrieval Module issues the query to the appropriate database using the efetch utility whereby the document for each entity under investigation is obtained. This document generally includes PubMed references for the gene or protein and sequence data as well as the GO codes assigned to gene or protein (when available). Both the relevant PubMed IDs and GO

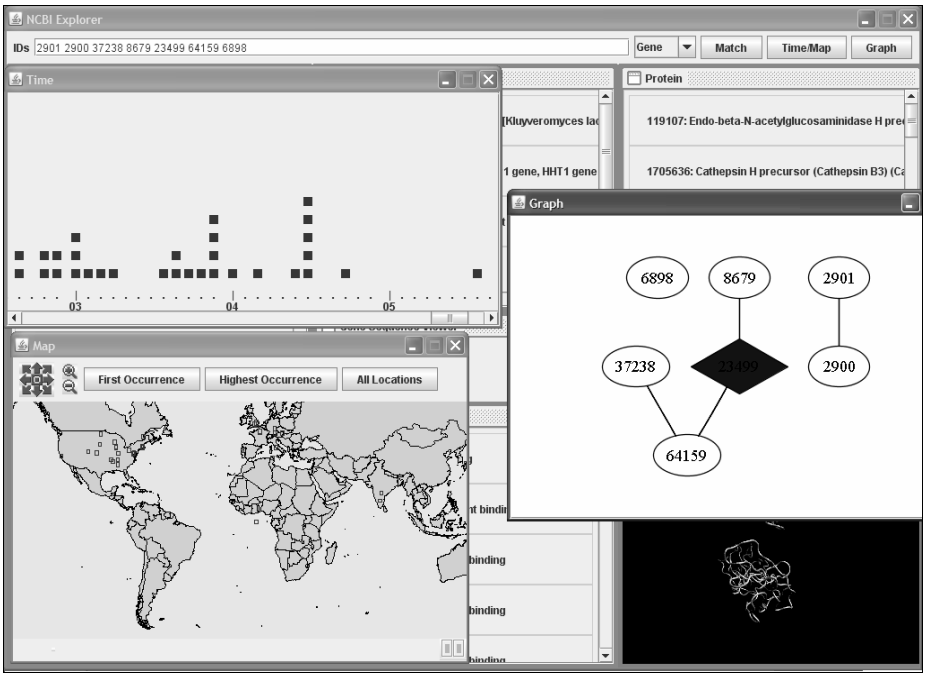

Fig. 2. The interface with spatial, temporal, and relationship visualization overlaid

codes are extracted and stored for display on the user interface as well as for correlating the information about the different genes or proteins. For queries of proteins, a diagram of the protein structure (when available) is obtained from PDB as well.

In the next step, the literature for each entity is accessed using the aforementioned PubMed IDs in conjunction with the efetch utility. The document thus obtained is parsed for information relevant to the views of the data in the UI, namely the title of the literature reference and its spatial-temporal data. Additionally, when a single entity is queried, the MetaMap Transfer Program (MMTx) [15] of the UMLS Metathesaurus is used to map the text to a controlled vocabulary of biological concepts, specifically, to extract names of related genes or proteins for display as a network of relationships (using [16]). When multiple entities are queried, the graph displays each as a node with edges representing GO codes shared by entities (Fig. 2). (It is noteworthy, however, that the visualization may be extended to encompass relationships between different entity types, such as genes and diseases.) Extraction of the spatial-characteristics of the information requires parsing of the data (specifically, the "affiliation" field) and cross-referencing the terms therein with an index of locations. Once found, the city and country names are used to determine the latitude and longitude values of the location by which they are displayed on the map.

Temporal information of the documents is used to indicate each on the timeline and is, in contrast, much easier to obtain, as it is given in the "date published" field. However, since spatial-temporal information is not fundamental to biological under-

standing, the timeline and map views are displayed as pop-up windows rather than having dedicated space on the interface (Fig. 2).

The user interface is populated with various panes displaying a plethora of information about the entities of interest. The leftmost pane lists the entities along with their literature references, as represented by the title of the publication. Elsewhere, the sequence data of the gene/protein is displayed, possibly with the structure as well for proteins (using Java Molecular Viewer [17]). The upper right-hand pane shows, in graphical form, gene-gene or protein-protein relationships suggested by the literature. Lastly, the middle panels contain the GO codes associated with the entities of interest. The lower of these shows the GO codes extracted from the entities, while the higher one lists the entities themselves.

Whether a single or multiple genes or proteins are searched, the graph visualization and displayed GO codes may be used to infer relationships between or co-functionality of entities. As stated above, support for data context entails changes or interactions in one view being reflected in all views. While there are multiple ways that this may be implemented, our system does so mainly by highlighting the information in the other views that corresponds to the selection in the active view. With respect to the GO annotations, for example, selecting an entity in the literature view would show its GO codes only in the lower pane, as well as all of the other entities queried sharing one or more of the codes in the upper pane. Selecting one of the seemingly related entities would serve to highlight those of the GO code(s) it is described by to shed light upon the nature of the similarity or relationship between the entities (Fig. 3) and may be used for functional clustering of entities.

The other views of the data are also reflective. For example, the proteins encoded by the selected gene are highlighted, and its nucleotide sequence displayed in the sequence viewer. In addition, the locations and times corresponding to the selected gene's documents would be highlighted on the map and timeline. Furthermore, selecting an area of the map or timeline would in turn cause the document(s) corresponding to that location or date to be highlighted. This capability allows users to observe spatial-temporal trends in the research.

In accordance with the experiential paradigm, the various features of the system afford users a view of the data from multiple sources and perspectives as well as allowing users to directly interact with the information. As such, users may go beyond simple querying to explore the data and the various trends and relationships present therein. In addition, the unification of the views and query space into a single interface minimizes context switching, thereby easing the cognitive load on the user.

4 Experimental Results

While the system is designed primarily to facilitate exploration of information, some of its components (in particular, the location and time modules) can in fact speed up information extraction. As such, the proposed system was evaluated from two perspectives, that of information discovery and information extraction. The former,

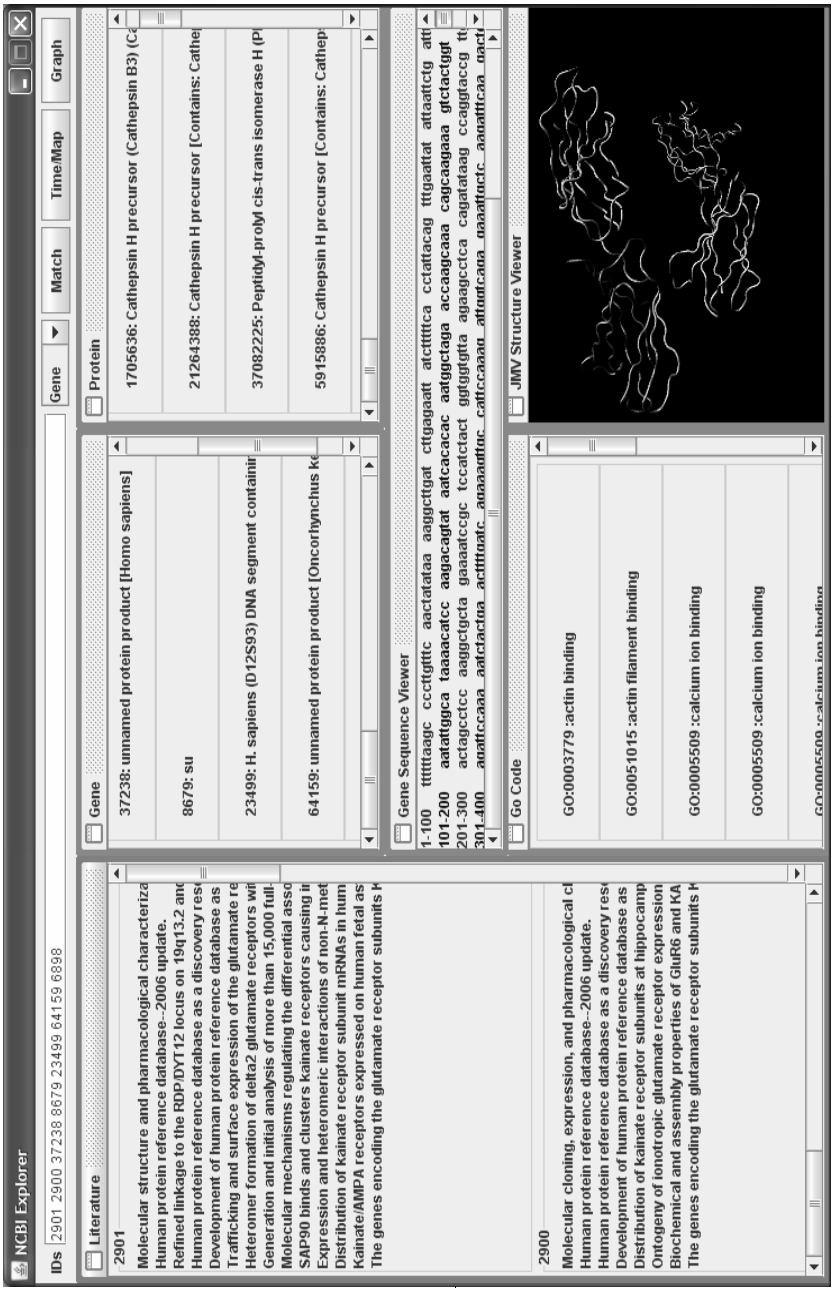

Fig. 3. The unified presentation-query-exploration interface

Table 1. The test set of genes, comprising 4 clusters (from [20])

GROUP	GENES	FUNCTION
1	nmda-r1 glur6 ka2 glur1 glur2 glur4 glur3 ka1	Glutamate receptor channels
2	dopamine beta-hydroxylase, tyrosine hydroxylase, phenethanolamine N-methyltransferase, catechol- O-methyltransferase, dopa decarboxylase, monoamine oxidase A, monoamine oxidase B	Catecholamine synthetic enzymes
3	alpha-tubulin, beta-tubulin, dynein, actin, alpha-spectrin	Cytoskeletal proteins
4	tyrosine transaminase, chorismate mutase, prephenate dehydratase, prephenate dehydrogenase	Tyrosine and phenylalanine synthesis

though essentially qualitative in nature, was assessed by demonstrating the system's ability to facilitate information exploration and discovery by presenting an example of display of entity relationships. In contrast, the latter analysis quantifies the contribution of the system's features towards improving performance of information extraction.

In order to meaningfully assess the value of the system vis-à-vis information discovery, it is necessary to compare entity relationships suggested by GO code correlations with known information. As such, we ran a test set of 24 human genes consisting of 4 well-defined functional clusters (Table 1) to gauge the degree to which genes within the same cluster appear more related than those between clusters.

The results (Table 2) seem erratic at first glance. Genes in the first cluster have, on average, over 10 GO codes in common with other genes in the same cluster, while having nearly zero GO codes in common with genes outside of the cluster. Conversely, the remaining three clusters show no such difference, with average values ranging between zero and one codes in common both within and outside the cluster. These values, however, point not to a lack of coherence within clusters, but rather to a paucity of data. It is noteworthy that no definitive negative correlation is observed. When available, the data may indicate strong relationships between entities, however, the approach suffers from the lack thereof.

In the second evaluation, the performance of the system in the realm of information extraction was evaluated by comparison with searches of the NCBI databases through Entrez. In particular, four queries were formulated with information goals constructed in such a way as to gauge the efficacy of the map and timeline modules, with performance determined by two methods: counting the number of mouse clicks and measuring the accession time until the information goal was reached. The results are shown below (Table 3).

Table 2. GO codes as indicators of functional coherence

	Average Number of GO Codes in Common	
Cluster	Within Cluster	Between Clusters
1	10.8	0.4
2	0.8	0.8
3	0.6	0.7
4	0.0	0.2

Table 3. Comparison of information extraction using NCBI and our system

Query	NCBI		Proposed System	
	No. of clicks	Access Time (s)	No. of clicks	Access Time (s)
1	8	67	2	27
2	12	85	2	32
3	2	33	2	29
4	2	46	2	31

The first information goal was, for a given protein, to find a reference that was published in Japan. The second query was similar, but had as its objective the country that produced the most publications about another protein. It is evident that our system is at a clear advantage for queries involving geographical information, with information access accelerated by a factor of 3 to 4 using both time and number of clicks.

The third information goal was to find the number of reference articles published after 2000 about a certain protein, while the fourth was the year in which the most articles were published about the same protein. In this case, there is no clear advantage in speed of information extraction as a result of the timeline module. This is likely due to the fact that the year an article is published is often given along with its title. As such, the temporal information is readily visible and accessible, which is not the case for relevant spatial information. However, despite the ready availability of the date, the timeline provides more intuitive access to the data, and a better sense of its distribution.

5 Conclusion

In this paper, we have presented a novel metaphor to facilitate search of biomedical information. Development of approaches such as these is critical at this juncture in the field due to the proliferation of data. Tools for analysis and synthesis of data have not kept pace with those for its generation, which serves to drastically limit the information that may be derived therefrom. Our approach seeks to aid in the exploration and assimilation of knowledge using an experiential paradigm. As such, users are presented with an interface that combines query and presentation capabilities with interactive views of the data. Moreover, the data itself is derived from various repositories and multimodal in its nature, to further allow the user to "experience" the information from multiple perspectives. Additional features of the system include display of semantic correlations between entities via common GO codes as well as display of the spatial-temporal characteristics of the data. Examples of queries submitted to the interface demonstrate the ability of the system to promote retrieval, exploration, and discovery of information, and thereby illustrate the promise of the approach.

References

1. NCBI, http://www.ncbi.nlm.nih.gov.
2. NCBI Entrez, http://www.ncbi.nlm.nih.gov/Entrez.

3. R. Jain, "Experiential Computing," Communications of the ACM, Vol. 46, No. 7, July 2003.
4. R. Singh, R. L. Knickmeyer, P. Gupta, and R. Jain, "Designing Experiential Environments for Management of Personal Multimedia," ACM Multimedia, 2004.
5. Gene Ontology, http://www.geneontology.org.
6. I. Iliopoulos, A. J. Enright, and C. A. Ouzounis, "TextQuest: Document Clustering of MEDLINE Abstracts for Concept Discovery in Molecular Biology," Pacific Symposium on Biocomputing, 2001.
7. M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman, "Using BLAST for Identifying Gene and Protein Names in Journal Articles," Gene, Vol. 259, 2000, 245-252.
8. D. Proux, F. Rechenmann, L. Julliard, V. Pillet, and B. Jacq, "Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction," Genomic Informatics Workshop, Vol. 9, 1998, 72-80.
9. K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi et al., "Toward Information Extraction: Identifying protein names from biological papers," Pacific Symposium on Biocomputing, 1998.
10. T. Rindflesch, L. Tanabe, J. Weinstein, L. Hunter et al., "EDGAR: Extraction of Drugs, Genes, and Relations from the Biomedical Literature," Pacific Symposium on Biocomputing, 2000.
11. V. Narayanasamy, S. Mukhopadhyay, M. Palakal, D. A. Potter, "TransMiner: Mining Transitive Associations among Biological Objects from Text," Journal of Biomedical Science, Vol. 11, 2004, 864-873.
12. A. Vailaya, P. Bluvas, R. Kincaid, A. Kuchinsky, M. Creech, and A. Adler, "An Architecture for Biological Information Extraction and Representation," Bioinformatics, Vol. 21, No. 4, 2005, 430-438.
13. NCBI E-Utilities, http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.htm.
14. RCSB Protein Data Bank, http://www.rcsb.org/pdb.
15. A. R. Aronson, "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program," Proc. American Medical Informatics Association Symposium, 2001.
16. Grappa: A Java Graph Package (AT&T Labs - Research) http://www.research.att.com/~john/Grappa.
17. Java Molecular Viewer (JMV is developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign with NIH support), http://www.ks.uiuc.edu/Research/jmv.