Visualization and Language Processing for Supporting Analysis Across the Biomedical Literature

Carsten Görg[1], Hannah Tipney[2], Karin Verspoor[2], William A. Baumgartner Jr.[2], K. Bretonnel Cohen[2], John Stasko[1], and Lawrence E. Hunter[2]

[1] School of Interactive Computing & GVU Center,
Georgia Institute of Technology, Atlanta, GA 30332
[2] Center for Computational Pharmacology,
University of Colorado Denver School of Medicine, Aurora, CO 80045
goerg@cc.gatech.edu,{Hannah.Tipney,Karin.Verspoor}@ucdenver.edu,
William.Baumgartner@ucdenver.edu,kevin.cohen@gmail.com,
stasko@cc.gatech.edu,Larry.Hunter@ucdenver.edu

Abstract. Finding relevant publications in the large and rapidly growing body of biomedical literature is challenging. Search queries on PubMed often return thousands of publications and it can be a tedious task to filter out irrelevant publications and choose a manageable set to read. We have developed a visual analytics system, named Bio-Jigsaw, which acts like a visual index on a document collection and supports biologists in investigating and understanding connections between biological entities. We apply natural language processing techniques to identify biological entities such as genes and pathways and visualize connections among them via multiple representations. Connections are based on co-occurrence in abstracts and also are drawn from ontologies or annotations in digital libraries. We demonstrate how Bio-Jigsaw can be used to analyze a PubMed search query on a gene related to breast cancer resulting in over 1500 primary papers.

Key words: Visual analytics, investigative analysis, entity identification, language processing, biomedical literature.

1 Introduction

The emergence of biomedical technologies, such as microarrays, genome-wide association studies, and methods exploiting low-cost sequencing, has made the simultaneous observation of all gene products in a genome both easily accessible and routine. The ability to assay biological systems at a genomic scale has enabled the collection and documentation of biomedically relevant information at a level of unprecedented scale and detail. The resulting explosion of knowledge contains information critical to the advancement of biomedical research and the understanding of human health and disorder.

While many aspects of this knowledge is captured in structured form within freely available gene- and protein-centric databases (some 1,170 peer-reviewed databases were cataloged in 2009 [7]), the PubMed bibliographic database housed at the National Center for Biotechnology Information (NCBI) at the National Library of Medicine

(NLM), remains the largest, most comprehensive source of biomedically important knowledge [20]. Peer-reviewed biomedical literature is not only the richest and most reliable of data sources, it is also the most overwhelming. Currently PubMed contains references to more than 19 million biomedical articles, and in 2008 expanded at a rate of approximately 2,200 new entries per day (calculated from PubMed 2008 indexed entries) [10].

This flood of knowledge has been accompanied by a breakdown of disciplinary boundaries, which traditionally made it possible to at least keep up with advances in a single field. Genome-scale research is intentionally broad, assaying potentially every gene in a genome, meaning pertinent prior knowledge could come from almost any biomedical discipline. Such blurring of boundaries means it is becoming increasingly common for new and important functions to be discovered for previously characterized genes, for example Relaxin 1 (RLN, NCBI Entrez Gene GeneID: 6013[1]), originally characterized as a cervical ripening hormone in the 1950s [8], has recently been implicated in processes as diverse as osteoarthritis and heart failure [16, 19]. Therefore, not only is the volume of available biomedical knowledge captured in the literature growing at an unprecedented rate, but more and more of it is relevant to a larger number of biomedical scientists than ever before.

The challenge facing biomedical researchers is one of how to effectively and efficiently extract and interpret valuable and relevant knowledge trapped within this wealth of biomedical literature. Failure to do so can be extremely costly in terms of the wasted time, effort and money chasing weak leads, inadvertently duplicating already published results and missing important discoveries.

The current strategy most biomedical researchers use to identify articles of interest is to query PubMed using keywords and Boolean search terms via its online interface.[2] Keywords can include gene or protein names (*i.e.*, "ELN" or "Elastin"), diseases or disorders (*i.e.*, "cancer" or "Down Syndrome"), concepts or symptoms (*i.e.*, "pain" or "high glucose"), methodologies or technologies (*i.e.* "gel electrophoresis" or "microarray"), in addition to author names, publication dates and publication types. For each article a title and abstract is returned and the full text article can typically be accessed via a hyperlink. Currently, the only way for a researcher to parse the information returned to them is to manually read each abstract (and maybe even some of the full text articles) in an attempt to identify themes of interest.

Although popular, this strategy is challenging for a variety of reasons. Firstly, the sheer size of many of the document sets returned from these searches is frequently unmanageable; not only is it incredibly difficult and time consuming to manually read hundreds of abstracts, it is also virtually impossible to remember more than a handful of details and themes from such large document sets without taking notes or using other memory aids. Failure to identify and retain key themes and entities can result in knowledge being discarded as unimportant or missed completely. Secondly, within biomedicine the use of synonyms for gene and protein names is rife, which can cause confusion when trying to disambiguate between such entities within a large document set. Thirdly, each time the scientist encounters a new concept or entity (such as a dis-

[1] `http://www.ncbi.nlm.nih.gov/gene`
[2] `http://www.ncbi.nlm.nih.gov/pubmed`

ease or gene) within the document set they must often invest additional time exploring supplementary datasources (*i.e.*, gene- or disease-centric databases such as EntrezGene or OMIM) to determine if these new terms are in fact of interest to them. Not only are these concepts and entities thought of as individual features, the investigating scientist will also be attempting to determine if, and what, relationships may exist between them.

Again, a failure to fully explore such relationships can result in the inappropriate disregard of critically important knowledge. Finally, when manually reading these document sets, the reader is inadvertently biased towards themes and concepts they already deem important, making it tricky to be aware of and receptive to knowledge from far outside their sphere of specialism. When a reader is unaware of just how much they do not know about a biomedical concept or process it is difficult to gauge when something is actually of interest.

In this paper, we describe a visual analytics system, Bio-Jigsaw, which supports biologists in investigating connections between biological entities grounded in the biomedical literature. The system identifies mentions of biological entities in text, specifically genes, using natural language processing strategies, visualizes relationships among those entities based on document co-occurrence, and allows a biological analyst to explore the documents from which those relationships are derived. We present an analysis scenario in which we demonstrate how the system supports biologists in exploring the biomedical literature.

2 Navigation of the Biomedical Literature

There are a number of tools available to the biomedical researcher to aid in navigation of the literature. GoPubMed[3] [5] supports the organization of the abstracts returned from a PubMed query according to Gene Ontology[4] and MeSH (Medical Subject Headings) concepts through recognition of those concepts in the text. Textpresso[5] [17] enables identification of terms from 33 categories, including genes, cells, phenotypes, cellular components, etc. through regular expressions that capture a significant amount of the variation in the surface form of those terms. It further supports document retrieval through queries that can combine categories with specific words, e.g. documents that mention two specific genes and a term belonging to the "regulation" category. The iHOP system[6] [9] provides an interface that links genes based on co-occurrence in sentences in PubMed abstracts, and provides hyperlinks among those sentences for navigation and exploration of literature relationships among genes. Reflect[7] is a Firefox plug-in which recognizes and highlights mentions of proteins and small molecules on any webpage, providing direct access to structured information about the highlighted entities through a pop-up window. Each of these tools plays some role in facilitating interpretation and navigation of the biomedical literature, by providing a more conceptual analysis of the content of the literature.

[3] http://www.gopubmed.org
[4] http://www.geneontology.org
[5] http://www.textpresso.org
[6] http://www.ihop-net.org/UniPub/iHOP
[7] http://www.reflect.ws

These tools are oriented towards supporting the user who is manually exploring the literature, and they do provide important assistance in identifying and organizing relevant literature. They do not, however, emphasize visual exploration of the document space. Out of these systems, only iHop takes any advantage of connections among documents, and none provide analysis over concepts in the document set as a whole. To support such analysis, specific biological facts must be extracted from the literature and represented in computable format. Such information extraction has been the subject of recent research. Protein-protein interactions have been the most common target for biological event extraction from the earliest studies [2, 4] to the latest competitions like BioCreative II [15] and II.5 [14]. Research has also extended to other event types including those addressed in the recent BioNLP'09 challenge [13]: gene expression, transcription, protein catabolism, protein localization, binding, phosphorylation, and regulation. Other studies have addressed the extraction of gene-disease relations [3], protein residue annotation [18], among others.

In this work, we treat co-occurrence of two genes within one publication as evidence of a relationship between them, without employing more detailed event or interaction extraction. We perform information extraction, specifically the gene mention detection and gene normalization algorithms of [1], to recognize occurrences of specific genes in PubMed abstracts and to associate them with the appropriate Entrez Gene identifier. It is our intention that in future work we will take advantage of our OpenDMAP concept recognition system [11] to identify more specific relationships among genes. However, with the proper weighting, literature co-occurrence is by itself a valuable indicator that we can take advantage of [6].

3 Visualizing Connections Across the Biomedical Literature

We have developed a visual analytics system, Bio-Jigsaw, to support biologists in investigating connections between biological entities or concepts grounded in the biomedical literature. Bio-Jigsaw is a customized version of the Jigsaw system [21, 12], tailored to the bio-informatics domain.

Connections in Bio-Jigsaw are based on co-occurrence in abstracts or drawn from ontologies and annotations in digital libraries. Bio-Jigsaw is a multiview system, including a number of different visualizations of the documents in the collection and the entities or concepts (genes, MeSH terms, KEGG pathways, GO biological processes, etc.) within those documents.

The two List Views in Figure 1 show connections between PubMed abstracts and MeSH terms and between MeSH terms and genes, respectively. Lists can be sorted alphabetically, by the frequency of occurrence in the whole document set (the larger an item's bar the more frequently it occurs), or by connection strength (the darker the shade of orange the stronger the connection). Figure 2 shows the Word Tree [22] for Tamoxifen. A Word Tree shows all occurrences of a word or phrase from the documents in the context of the words that follow it. The analyst can navigate through the tree by clicking on its branches. The Graph View in Figure 3 shows connections between documents (white rectangles) and entities (colored circles) using a node link diagram. The Document View in Figure 4 displays documents and highlights identified entities

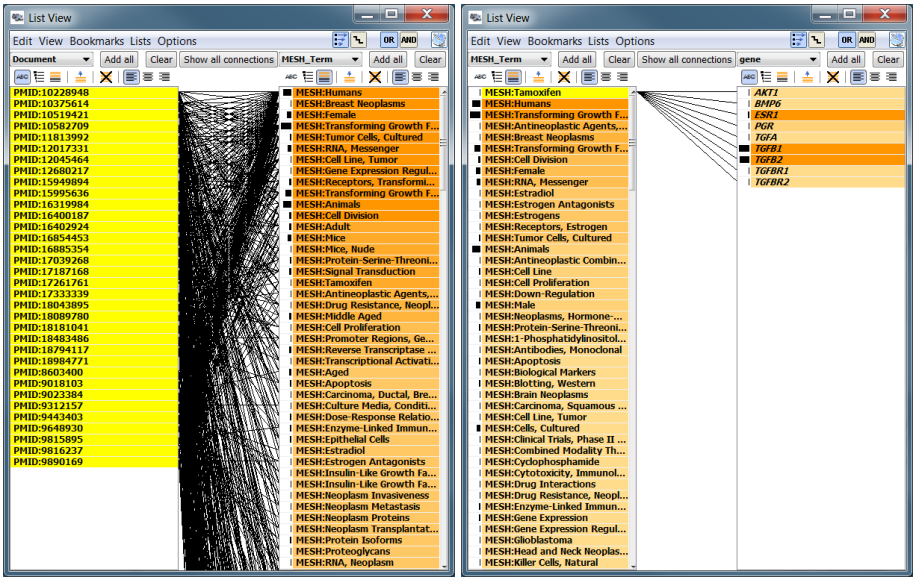

Fig. 1. The List View showing connections between PubMed abstracts, MeSH terms, and genes. Selected items are shown in yellow; connected items are in orange.

within them. The word cloud at the top shows the most frequent keywords from the set of currently loaded abstracts.

A textual search query interface allows users to find particular entities and the documents in which they occur. In addition, entities and documents can be explored directly by interacting with those objects in the views. For instance, new entities can be displayed and explored by user interface operations in the views that expand the context of entities and documents. In practice these two approaches are often combined: search queries serve to jump-start an exploration and view interaction then yields richer representations and exploration.

4 Scenario

In this section we walk through an analysis scenario to demonstrate how Bio-Jigsaw supports biologists in exploring the biomedical literature. We also provide a video that demonstrates the scenario actions in more detail.[8]

After looking at clinical breast cancer data our analyst has become interested in the tgfb2 gene. After searching PubMed for 'tgfb2', over 1500 primary papers are returned to her. To start exploring these papers in Bio-Jigsaw, she opens a List View and searches for documents containing the phrase 'breast cancer'; 34 documents are found and displayed in the left column of the List View (see Figure 1(a)). The analyst then views the

[8] The video is available at `http://www.gvu.gatech.edu/ii/jigsaw/BioJigsaw.avi`.

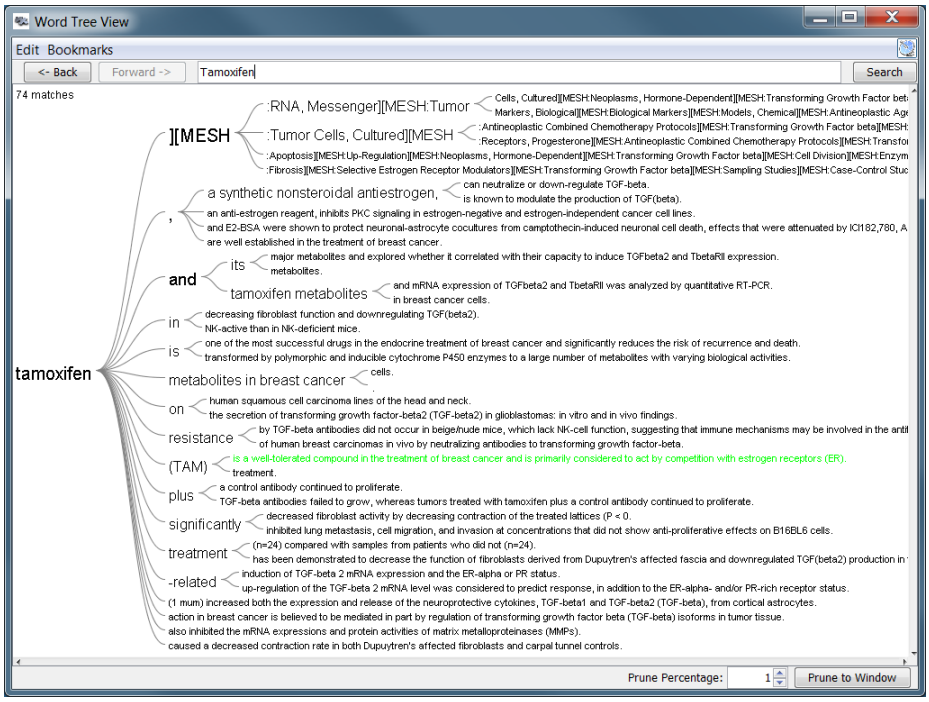

Fig. 2. Word Tree for Tamoxifen across a collection of journal articles. All trailing words are shown, sized by their frequency of occurrence.

MeSH terms (a controlled vocabulary used to describe the content of biomedical articles within PubMed[9]) associated with the 34 documents in a second list and sorts them by connection strength. She identifies the term "Tamoxifen" as being of interest because although it is not frequently observed across the document collection as a whole (as illustrated by the short bar), it is well connected to the 34 selected documents (dark color).

The term Tamoxifen is unknown to the analyst. By launching the Word Tree View she investigates this term further (see Figure 2). The Word Tree displays all sentences from the document set containing the word Tamoxifen, grouped by common suffixes. The analyst can quickly see that Tamoxifen is a compound used during the treatment of breast cancer.

By adding an additional list to the List View, the analyst next explores other entities that are associated with the subset of documents identified by the phrase breast cancer. The analyst wants to know if other genes are also affected by Tamoxifen so she displays the gene entity in the new list (see Figure 1(b)). She can see from their highlighting that three genes (tgfb1, tgfb2, and esr1) are particularly well connected to Tamoxifen within this document set.

[9] `http://www.nlm.nih.gov/mesh`

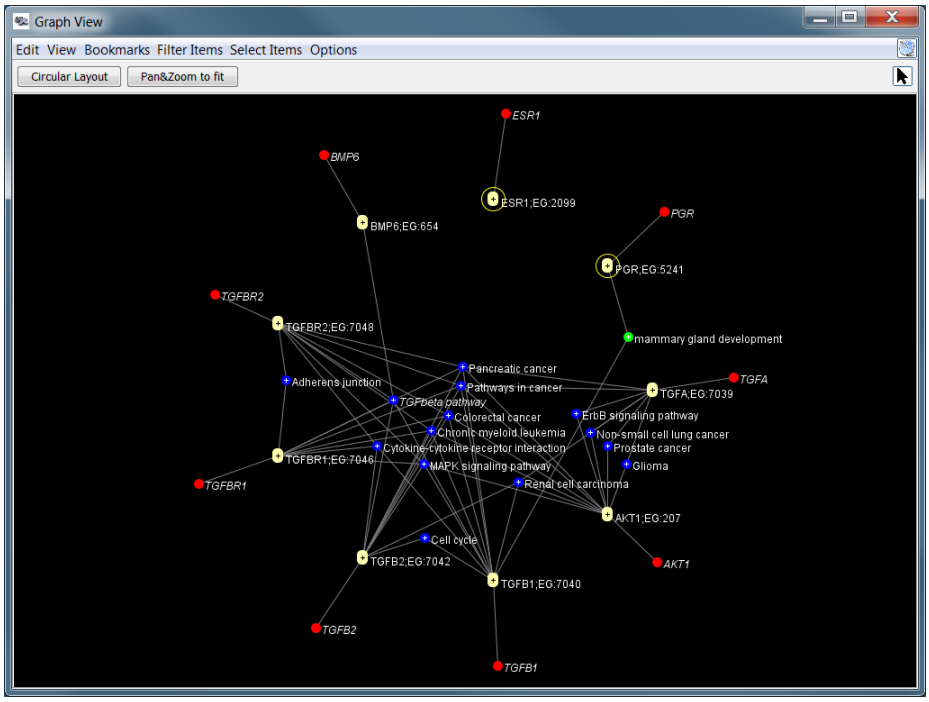

Fig. 3. The Circular Layout in the Graph View showing connections between ontology documents (white rectangles) and KEGG pathways. All entities (colored circles) connecting to more than one document are drawn in the middle making it easier to focus on them.

Now the analyst turns to the Graph View (see Figure 3) and displays all the genes (red nodes) connected to Tamoxifen along with their ontology documents (yellow nodes), as well as connected KEGG pathways (blue nodes) and GO biological processes (green nodes). By applying the circular layout (documents are displayed on a circle), entities connected to only one document are positioned outside the circle, while entities connected to multiple documents are shown inside the circle; the more connections entities have, the closer to the center they are positioned. The analyst then filters out irrelevant entities and notices that although a number of cancer associated KEGG pathways are shared by the tgfb family and their receptors, the progesterone and estrogen receptors (highlighted with yellow circles) do not share these annotations. However, the GO biological process term “mammary gland development” is common to both tgfb and the progesterone receptor and the analyst wonders if the progesterone receptor has a role in breast cancer which is affected by Tamoxifen and if it is mediated by members of the tgfb family.

Now the analyst turns back to the List View to find the documents most connected to the tgfb genes, their receptors, and the progesterone receptor. She displays the eight most connected documents in the Document View (see Figure 4) to read their abstracts. She learns not only that Tamoxifen increases the expression of tgfb2 resulting in inhi-

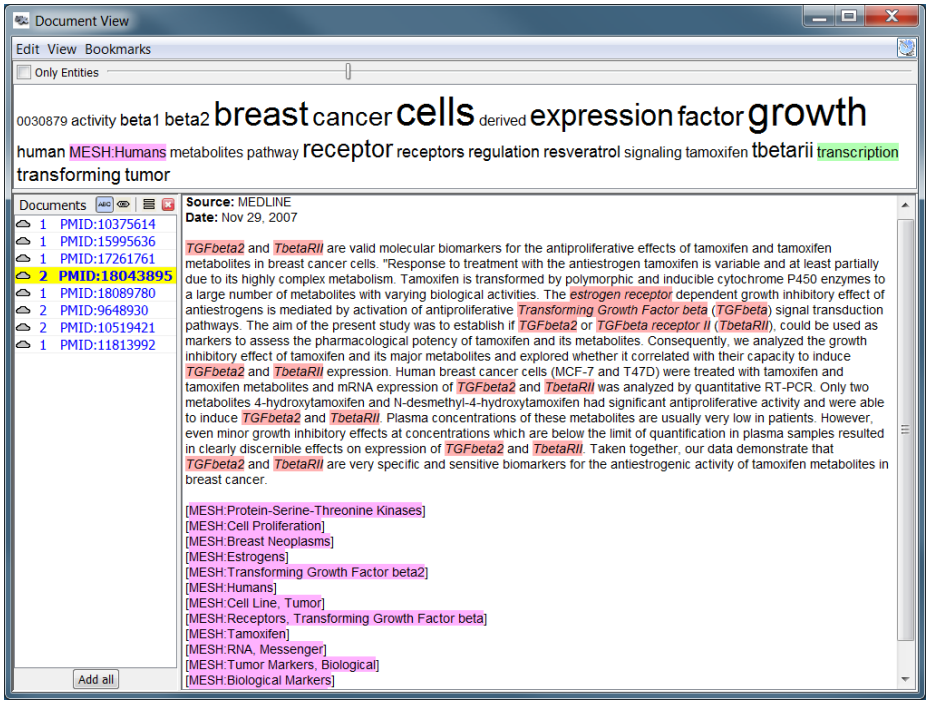

Fig. 4. The Document View showing a set of abstracts relevant to tgfb2 (list on the left side). The word cloud at the top shows the most frequent keywords from the set of abstracts. Document PMID:18043895 is selected and its abstract and related terms are visible. Entities are highlighted: genes in red and MeSH terms in purple.

bition of breast cancer growth, but also that the steroid receptor status of tumors (*i.e.*, if the tumor expresses progesterone receptors or not) can be predictive of the tumors response to Tamoxifen.

5 Conclusion

In this paper, we introduced the Bio-Jigsaw system which integrates natural language processing for entity identification and normalization in the biomedical literature with sophisticated visualization strategies. We provided an example of how the visualization of relationships among entities occurring in a document set can be navigated and explored by a biological analyst to gain new insights into a gene of interest. Through the visualization, the analyst is able to quickly identify new concepts that are relevant to the gene under investigation, and to hone in on concepts that are strongly indicated by the document set. The different views that the system makes available each play an important role in exploring the document set, and the analyst can move smoothly among them as well as manipulate which entities and concepts the visual analysis focuses on. Through the use of this tool, the biologist has the ability to navigate and explore the

biomedical literature in such a way that they can much more effectively extract, manipulate, and interpret the knowledge that exists there.

Acknowledgments

This research is based upon work supported in part by the National Science Foundation via Award IIS-0915788 and by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. We also acknowledge the support of NIH National Library of Medicine grants 2R01LM009254 and 2R01LM008111 to Lawrence Hunter, and NIH grant 1R01LM010120-01 to Karin Verspoor.

References

1. William A. Baumgartner Jr., Zhiyong Lu, Helen L. Johnson, J. Gregory Caporaso, Jesse Paquette, Anna Lindemann, Elizabeth K. White, Olga Medvedeva, K. Bretonnel Cohen, and Lawrence Hunter. Concept recognition for extracting protein interaction relations from biomedical text. *Genome Biology*, 9, In press.
2. Christian Blaschke, Miguel A. Andrade, Christos Ouzounis, and Alfonso Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Intelligent Systems for Molecular Biology*, pages 60–67, 1999.
3. H.W. Chun, Y. Tsuruoka, J.D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Pacific Symposium on Biocomputing*, pages 4–5, 2006.
4. Mark Craven and Johan Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Intelligent Systems for Molecular Biology*, pages 77–86, 1999.
5. A. Doms and M. Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33:783–786, 2005. GoPubMed.
6. Aaron Gabow, Sonia M. Leach, William A. Baumgartner Jr., Lawrence E. Hunter, and Debra S. Goldberg. Improving protein function prediction methods with integrated literature data. *BMC Bioinformatics*, 9(198), 2008.
7. M.Y. Galperin and G.R. Cochrane. Nucleic acids research annual database issue and the nar online molecular biology database collection in 2009. *Nucleic Acids Research*, 37:D1–4, 2009.
8. E.F. Graham and A.E. Dracy. The effect of relaxin – mechanical di-latation of the bovine cervix. *Journal of Dairy Science*, 36:772–777, 1953.
9. R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nat Genet*, 36(7), July 2004.
10. Lawrence Hunter and K Bretonnel Cohen. Biomedical language processing: what's beyond PubMed? *Mol Cell*, 21(5):589–94, 2006.
11. Lawrence Hunter, Zhiyong Lu, James Firby, William A. Baumgartner Jr., Helen L. Johnson, Philip V. Ogren, and K. Bretonnel Cohen. OpenDMAP: An open-source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-specific gene expression. *BMC Bioinformatics*, 9(78), 2008.
12. Younah Kang, Carsten Görg, and John Stasko. The evaluation of visual analytics systems for investigative analysis: Deriving design principles from a case study. *IEEE VAST*, pages 139–146, Oct. 2009.

13. Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of BioNLP'09 shared task on event extraction. In *BioNLP 2009 Companion Volume: Shared Task on Entity Extraction*, pages 1–9, 2009.
14. Martin Krallinger, Florian Leitner, and Alfonso Valencia. The BioCreative II.5 challenge overview. In *Proceedings of the BioCreative II.5 Workshop 2009 on Digital Annotations*, 2009.
15. Martin Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, and Alfonso Valencia. Evaluation of text-mining systems for biology: overview of the second biocreative community challenge. *Genome Biology*, 9(Suppl 2):S1, 2008.
16. M. Kupari, T.S. Mikkola, H. Turto, and J. Lommi. Is the pregnancy hormone relaxin an important player in human heart failure? *Eur J Heart Fail*, 7:195–198, 2005.
17. H. M. Muller, E. E. Kenny, and P. W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11), 2004.
18. Kevin Nagel, Antonio Jimeno-Yepes, and Dietrich Rebholz-Schuhmann. Annotation of protein residues based on a literature analysis: cross-validation against uniprotkb. *BMC Bioinformatics*, 10(Suppl 8):S4, 2009.
19. K. Santora, C. Rasa, D. Visco, B.G. Steinetz, and C.A. Bagnell. Antiarthritic effects of relaxin, in combination with estrogen, in rat adjuvant-induced arthritis. *J Pharmacol Exp Ther*, 322:887–893, 2007.
20. Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Eugene Yaschenko, and Jian Ye. Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.*, 37(Suppl 1):D5–15, 2009.
21. John Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.
22. Martin Wattenberg and Fernanda B. Viégas. The word tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, 2008.