Overview of the Ninth Annual Meeting of the BioLINK SIG at ISMB: Linking Literature, Information and Knowledge for Biology

Christian Blaschke[1], Lynette Hirschman[2], Hagit Shatkay[3],
and Alfonso Valencia[4]

[1] Bioalma, Spain
blaschke@almabioinfo.com
[2] MITRE Corporation
lynette@mitre.org
[3] Computational Biology and Machine Learning Lab,
School of Computing, Queen's University, Kingston, Ontario, Canada
shatkay@cs.queensu.ca
[4] Structural Biology and Biocomputing Programme,
Spanish National Cancer Research Centre (CNIO)
valencia@cnio.es

1 About BioLINK

With the increasing availability of textual information related to biological research, such information has become an important component of many bioinformatics applications. Much recent work aims to develop practical tools to facilitate the use of the literature for annotating the vast amounts of molecular data, including gene sequences, transcription profiles and biological pathways. The broad area of biomedical text mining is concerned with using methods from natural language processing, information extraction, information retrieval and summarization to automate knowledge discovery from biomedical text. In the biomedical domain, research has focused on several complex text-based applications, including the identification of relevant literature (information retrieval) for specific information needs, the extraction of experimental findings for assistance in building biological knowledge bases, and summarization – aiming to present key biological facts in a succinct form.

Automated natural language processing (NLP) began in 1947 with the introduction of the idea of machine translation by Warren Weaver, and work on automated (still mechanical) dictionary lookup for translation by Andrew Booth [2,7]. This work was continued throughout the 1950s in research on automatic translation by Bar Hillel, Garvin and others. In the 1950s, work on transformational grammars by Zellig Harris [3], formed the basis for computational linguistics, which was continued by Noam Chomsky, relating natural languages to formal grammars. The field made rapid progress starting in the late 1980s, thanks to a series of conferences focused on evaluation of text mining and information extraction systems: the Message Understanding Conferences (MUCs).

C. Blaschke and H. Shatkay (Eds.): ISBM/ECCB 2009, LNBI 6004, pp. 1–7, 2010.

There is also a long history of research on applications of text mining and natural language processing in medicine going back to the late 1960's with Sager's early work on parsing of scientific 'sublanguages' [11,12]. Within biology, text-based methods were introduced in the late 90's. The rapid accumulation of data emerging from advances in sequencing and other high-throughput methods has made the literature a critical source of information for new biological findings. There has been an increasing need for tools to help researchers manage and digest this growing volume of information. Early work in text mining for biology included a 1997 MSc thesis by Timothy Leek [9], and the first publication of an article at the ISMB conference [1].

As is often the case in interdisciplinary fields, communication between the developers of the tools (here – text mining and information extraction tools) and the actual users (in this case – the biologists) is necessary for the development of truly beneficial tools. The BioLINK group was created to address the needs of communication within the field of text mining and information extraction as it is applied to biology and biomedicine, aiming to bring together developers and users. Regular open meetings have been held in association with the ISMB conferences since 2001, facilitating interactions among researchers to exchange ideas with the wider community interested in the latest developments. These meetings focus on the development and the application of resources and tools for biomedical text mining.

BioLINK is interdisciplinary in nature and involves researchers from multiple communities: the users of text mining tools, including curators of biological databases, bench scientists and bioinformaticians; and the researchers who develop methods in natural language processing, ontologies, text mining, image analysis, information extraction and retrieval in order to apply them to problems in the biomedical domain.

In the last decade, BioNLP methods have been maturing, as demonstrated by the results from the BioCreative assessments (Critical Assessment for Information Extraction in Biology). The first BioCreative was held in 2004 [6] and the second in 2007 [8]. During this period, the research community grew (44 team participated in BioCreative II) and the results improved; for instance, in BioCreative II, the best systems achieved almost an f-measure of almost 0.9, in identifying mentions of genes and proteins in text.

Moreover, the availability of the full content of scientific publications has increased to the point where new challenges can be posed and addressed. Recent text mining experiments have begun to use full text articles, in particular to extract information about experimental evidence. Moreover, much of the published material includes images which are of utmost importance for both scientists and database curators. Images typically provide critical supporting evidence for assertions occurring within articles. Image analysis is an important tool in understanding biomedical processes in multiple granularity levels, and also has great potential to enhance document retrieval and categorization. While there has been ongoing interest in developing systems for automatically extracting biological information from the literature, relatively little has been done so far to utilize information from images or on combining text and image data. Recently the challenge of automatically and effectively processing images and figures from the scientific literature is generating much interest in image analysis as a source of biomedical data.[4,10,13,15]

To take into account these new priorities, the scope of BioLINK was extended this year to include the analysis of images and figures within scientific publications. Furthermore, a session about the present and future of scientific publishing was included, in which representatives from scientific journals and community members discussed the impact of information extraction methods (whether from text or images) on producers and consumers of scientific information. Two overview papers, one pertaining to image analysis and the other to the future of scientific publishing, are included in this volume.

Since its inception, it has been part of the mission of the BioLINK SIG to formulate common goals and define standard data sets and uniform evaluation criteria for biomedical text mining systems. In one of the early meetings (in 2002) we proposed to organize an assessment inspired by the well known CASP evaluations for protein structure predictions. This initiative led to the BioCreative challenges that started at the end of 2003. BioCreative is now a well accepted forum for system assessment in the field and has helped to define shared tasks and standardized evaluation criteria, stimulating interaction and exchange of ideas between developers and users of text mining technologies within the biological domain.

2 History of BioLINK Meetings

2001, ISMB Copenhagen, Denmark: Lynette Hirschman and Alfonso Valencia organize the first workshop related to text mining and literature analysis at the ISMB in Copenhagen. This workshop became the predecessor of the BioLINK Special Interest Group that met since then at the annual ISMB conferences.

2002, ISMB Edmonton, Canada: This was the first year where people were invited to contribute publications and present them at the workshop. A wide range of themes were covered including term and entity recognition, augmenting the gene ontology with text mining, functional analysis of genes based on text and literature based discovery.

In addition, interesting discussions took place about the results of the KDD challenge cup [14] and the TREC genomics track [5]. Furthermore, a proposal was presented to organize a new evaluation of text mining systems closer to the needs of biologists. This discussion led to the creation of the BioCreative (Critical Assessment of Information Extraction in Biology) evaluations.

2003, ISMB Brisbane, Australia: The meeting in 2003 focused especially on developing shared infrastructure (tools, corpora, ontologies). The contributed publications discussed corpus resources, extracting protein-protein interactions from text, protein named entity recognition and automatically linking MEDLINE abstracts to the Gene Ontology. At that meeting the first BioCreative was discussed and the initial training data were released.

2004, ISMB Glasgow, Scotland: That year the BioLINK SIG meeting focused on resources and tools for text mining, with special emphasis on the evaluation of these tools. Contributions were in the area of named entity recognition in biomedical texts and infrastructures for term management. The discussions focused on the TREC genomics track and the results of BioCreative.

2005, ISMB Detroit, Michigan: In 2005 the BioLINK meeting was held jointly with the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. The contributions covered extraction of protein-protein interactions from text, named entity recognition and shallow parsing in biomedical texts, corpora for text mining, functional analysis of genes based on text, and user-oriented biomedical text mining.

2006, ISMB Fortaleza, Brasil: It was decided to organize a joint meeting of the Bio-ontologies and the BioLINK workshops covering both themes. This was done to build on the close relationships between bio-ontologies and biomedical text mining. For example, ontologists apply text-mining techniques to test, check and build ontologies, while the knowledge in ontologies is being used to augment and improve text-mining techniques. The meeting consisted of sessions that focused on the intersection of bio-ontologies and text mining, as well as individual sessions on the use of ontologies in the life sciences and on biomedical text mining.

Authors contributed presentations covering the analysis of cellular processes using text mining, the Protein Description Corpus, corpus annotation guidelines, formats and standards to enhance interoperability of TM systems, (deep) parsing of biomedical text, andfunctional analysis of genes based on text, and the use of images within text.

2007, ISMB Vienna, Austria: Following BioCreative II, the meeting focused on assessments, on standards for annotation both in biological databases and in biomedical text corpora, and on new tools for biomedical text mining.

Papers presented at the meeting discussed both the common themes of named entity recognition and parsing within biomedical texts, along with annotation tools, corpora, extraction of biomedical relationships, and more focus on user-centered text mining systems.

2008, ISMB Toronto, Canada: This BioLINK meeting focused on the theme of automated linkage of the literature to biological resources in support of applications such as: automated indexing of the biomedical literature, generation of structured digital abstracts and the use of text-mined data in biology and bioinformatics pipelines. To stimulate discussion, a forum of "end users" was invited to present their applications and text mining needs, with a specific goal of encouraging partnerships among the end users and the developers of text mining tools.

Papers presented at the workshop covered some of the (by now) traditional topics, such as named entity recognition and parsing of biomedical text, along with more recent topics including corpora and annotation tools, mining information from full text , linking text-based information to biological databases entries.

2009, ISMB Stockholm, Sweden: The latest BioLINK meeting moved beyond text analysis to take into account the analysis of images and figures in scientific publications, in recognition of the key information they provide. Furthermore, the meeting included a session about the future of scientific publishing, and the impact of information extraction methods on producers and consumers of scientific information. This is also the first year in which selected papers from the workshop are being published as conference proceedings within Lecture Notes in Bioinformatics.

Papers presented in this year's meeting discussed named protein-protein interactions, analysis of experimental data and hypothesis generation, linking text to databases entries, text mining systems in support of specific users and applications, augmentation of the gene ontology using text mining, corpus annotation tools, and image analysis in scientific publications.

The following papers, listed here by topic, have been included in this volume:

Effective training of document classifiers in support of database curation

Learning from Positives and Unlabeled Document Retrieval for Curating Bacterial Protein-Protein Interactions by Hongfang Liu, Guixian Xu, Manabu Torii, Zhangzhi Hu, and Johannes Goll.

The authors present a method for training classifiers to detect documents that contain information about protein-protein interactions based on publicly available data. Manually curating training data containing positive and negative examples is time consuming and in many situations not feasible. Often positive examples can be deduced from existing databases, but negative examples are not explicitly given. The authors explore different ways to create reliable negative training data and show that good classifiers can be trained from such automatically created training data.

Toward Computer-Assisted Text Curation: Classification is Easy (Choosing Training Data can be Hard...) by Robert Denroche, Ramana Madupu, Shibu Yooseph, Granger Sutton and Hagit Shatkay.

In this work the authors developed a system to identify abstracts that are likely to describe experimental characterization of proteins, as opposed to abstracts unlikely to contain such information, supporting the curation of characterized proteins. To train and test the classifiers, small hand-curated datasets, as well as a large set based on previously curated abstracts from Swiss-Prot and GO were constructed. The authors show that classifiers trained on relatively small hand-curated datasets perform at a high level very close to the level required by database curators. Another interesting finding of this study was that in more than 80% of the manually examined articles, the abstract and title alone contained sufficient information for determining the relevance of the article for database curation.

Text-mining in support for experimental data analysis

Combining Semantic Relations and DNA Microarray Data for Novel Hypotheses Generation by Dimitar Hristovski, Andrej Kastrin, Borut Peterlin, and Thomas C. Rindflesch.

One important application of text mining systems is to support scientists in interpreting experimental results. Hristovski et al. present a methodology that integrates the results of microarray experiments with a large database of semantic predications extracted by a text mining system from the scientific literature. Examples from microarray data on Parkinson disease are presented to illustrate the way semantic relations shed light on the relationship between current knowledge and information gleaned from the experiment, and help generate novel hypotheses.

Tools

Mining Protein-Protein Interactions from GeneRIFs with OpenDMAP by Andrew D. Fox, William A. Baumgartner Jr., Helen L. Johnson, Lawrence E. Hunter, and Donna K. Slonim.

Standard basic components that can be used as building blocks in biomedical text mining systems are only starting to emerge; during the last few years there are an increasing number of systems have been made available for public use. This work is an example of how publicly available tools for tokenizing, protein named entity recognition and information extraction can be put together to build a system to extract protein interactions from text. The work specifically makes use of a fairly limited and well-structured text, namely, GeneRIFs, and uses the UIMA framework to integrate the different components, thus increasing the possibilities for reuse. Fox et al. describe how the performance of individual processing steps influences the overall results; modules for detecting protein complexes and enhancements to their information extraction patterns are discussed.

Extracting and Normalizing Gene/Protein Mentions with the Flexible and Trainable Moara's Java Library by Mariana L. Neves, José María Carazo and Alberto Pascual-Montano.

The Moara system is an addition to the growing ecosystem of text mining components that are made available to the public and that can help developers focus on specific problems without having to re-implement existing methods. Moara is a Java library that can be easily integrated with other systems or used as a standalone application. It uses machine learning algorithms that are trained to detect and to normalize gene and protein names in the literature.

Integrating images and text

Structured Literature Image Finder: Extracting Information from Text and Images in Biomedical Literature by Luis Pedro Coelho, Amr Ahmed, Andrew Arnold, Joshua Kangas, Abdul-Saboor Sheikh, Eric P. Xing, William W. Cohen, and Robert F. Murphy.

The Structured Literature Image Finder (SLIF) is an example of an advanced, publically available system that can be used by researchers who are not familiar with the underlying technology. It uses a combination of text-mining and image processing to extract information from figures in the biomedical literature. To access the information a web-accessible searchable database is provided to the users. One can query the database for text appearing in figure captions or the images themselves and browse through the publications and their images.

References

1. Andrade, M.A., Valencia, A.: Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. In: Proceedings of the 5th Annual International Conference on Intelligent Systems for Molecular Biology, ISMB 1997 (1997)

2. Chan, S.W.: A Dictionary of Translation Technology. Chinese University Press (2004)
3. Harris, Z.S.: Transfer Grammar. International Journal of American Linguistics 20(4) (October 1954)
4. Hearst, M.A., Divoli, A., et al.: BioText Search Engine: beyond abstract search. Bioinformatics (June 2007)
5. Hersh, W., Bhupatiraju, R.: TREC genomics track overview. In: Proc. of the Twelfth Text Retrieval Conference, TREC 2003 (2003)
6. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6(Suppl. 1), S1 (2005)
7. Hutchins, J.: Warren Weaver Memorandum: 50th Anniversary of Machine Translation. In: MT News International, July 22, pp. 5–6 (1999)
8. Leek, T.R.: Information Extraction Using Hidden Markov Models. Master's thesis, Department of Computer Science, University of California, San Diego (1997)
9. Krallinger, M., et al.: Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biology 9(Suppl. 2), S1 (2008)
10. Murphy, R.F., Kou, Z., Hua, J., Joffe, M., Cohen, W.W.: Extracting and structuring subcellular location information from on-line journal articles: the Subcellular Location Image Finder. In: Proceedings of IASTED International Conference on Knowledge Sharing and Collaborative Engineering, KSCE 2004 (2004)
11. Sager, N.: Information Reduction of Texts by Syntactic Analysis. Seminar on Computational Linguistics. In: Pratt, A.W., Roberts, A.H., Lewis, K. (eds.) Division of Computer Science and Technology, National Institutes of Health, Bethesda, MD, pp. 46–56 (1966) (PHS Publication No. 1716)
12. Sager, N.: Syntactic Analysis of Natural Language. In: Advances in Computers, vol. 8, pp. 153–188. Academic Press, NY (1967)
13. Shatkay, H., Chen, N., Blostein, D.: Integrating Image Data into Biomedical Text Categorization. Bioinformatics 22(11) (2006); Special issue: Proc. of the Int. Conf. on Intelligent Systems for Molecular Biology (ISMB 2006) (August 2006)
14. Yeh, A.S., Hirschman, L., Morgan, A.A.: Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. In: Proceedings of the 11th Annual International Conference on Intelligent Systems for Molecular Biology, ISMB 2003 (2003)
15. Yu, H., Lee, M.: Accessing Bioscience Images from Abstract Sentences. Bioinformatics 22(11) (2006); Special issue: Proceedings of the 14th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2006) (August 2006)