Chapter 17

Text mining in life sciences

J. Fluck, H. Deneke & C. Gieger
Department of Bioinformatics,
Fraunhofer Institute for Algorithms and Scientific Computing,
Germany.

Abstract

Efficient information retrieval and extraction is a major challenge in molecular biology and genome-based clinical research. In addition, we realize an increasing demand to combine information from different resources and across different disciplines in life sciences. A large fraction of this information is only available in scientific articles. Now the immense volume of literature makes it almost impossible for biologists and clinical researchers to retrieve all relevant information on a specific topic and to keep up with current research. Text mining provides the methods to automatically retrieve and extract information contained in free text. Furthermore a structured representation formalized by ontologies is required to combine the extracted knowledge with other sources. Examples are descriptions of protein interactions or clinical phenotypes that can be used to analyze large scale experiments in drug development.

1 Introduction

The amount of information available in molecular biology is enormous due to the recent advent of high-throughput techniques and is expected to further increase at a nearly exponential rate. Moreover, if we are able to identify associations between the genome and proteome on the one hand and phenotypes and diseases on the other hand, other disciplines such as clinical research and drug development can greatly profit from this information. However, this requires the ability to extract relevant relations from a large number of potential candidates. This task typically exceeds the capabilities of manual work by humans. Efficient methods for computer-based knowledge management have to be applied to reduce and structure the information.

WIT Transactions on State of the Art in Science and Engineering, Vol 17, © 2005 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/978-1-85312-995-7/17

Currently, results of biological research are usually reported in journals as free text articles. In the last few years, text mining strategies have been developed to get access to the large amounts of information stored in free text to supplement biological databases. However, natural language processing poses a number of problems for computer-based knowledge management systems. In biology and medicine, as well as in any other domain, synonyms are used, and word meanings can be ambiguous. Hence, the identification of references by keyword searches will miss positive hits as well as provide false positives. In addition, the variety of grammatical structures that can be used to express a finding makes an automated extraction of information a difficult task.

Nevertheless, text mining can help to overcome problems induced by the limited coverage of information in current databases. It can provide strategies to quickly find all text articles or database annotations referring to a specific entity, and can extract relations between different entities. But there are also several drawbacks. Namely, text mining has only limited precision and recall. Until now, it is not broadly used to convert textual information in structured database entries. We discuss which methods are used today and which steps need to be taken to make text mining usable for a broader range of applications in life sciences.

2 Text mining – current state

2.1 Methodical development

Due to the obvious applications for text mining, many researchers have developed text analysis methods tailored to the specific needs in biology and medicine. Commonly used techniques in information retrieval are advanced search algorithms and text clustering. These approaches are able to retrieve information in an efficient way and/or to provide an overview over large collections of documents (*e.g.* Illiopoulos *et al* [1]). Approaches based on simple statistics can benefit significantly from incorporating structures obtained from domain dictionaries (*e.g.* Gieger *et al* [2]), or more generally ontologies (*e.g.* Glenisson *et al* [3]).

Currently, information extraction in biology is mainly focused on the identification of interactions between biological entities. This addresses the need of biologists to recognize, *e.g.* protein-protein interactions in text and get rid of the need to read every single document. A first fundamental requirement is the unique identification of biological entities or processes. Handling of unknown words, multiple synonyms for one entity, and the identification of terms composed of more than one word are some of the difficulties encountered in name recognition. Furthermore, the same names are used to identify different proteins, genes or other biological entities. Some names are also common English words, like the gene names ‘WAS’ and ‘KILLER’. Fukuda *et al* [4] found a solution to recognize protein names that uses special properties such as

the occurrence of uppercase letters, numerals, and special endings. Others extended the identification process to recognize chemical names (Wilbur *et al* [5]), and other biological entities (Narayanaswamy *et al* [6]). But all these methods still face the problem caused by the use of synonyms. This makes it hard to rely on the findings as building blocks for more advanced analysis. Hanisch *et al* [7] used a large, semi-automatically generated dictionary together with a token-based search algorithm as a conceptual backbone of information extraction. They showed a high specificity and sensitivity for identifying protein names and their corresponding objects in protein and gene databases. Techniques to extract interactions mostly rely on simple statistical measures, *e.g.* word co-occurrence (Stapley and Benoit [8]), and basic pattern matching (Ng and Wong [9]). More advanced linguistic approaches are used to handle relations in complex sentences (*e.g.* Park *et al* [10], Yakushiji *et al* [11]). Recently, natural language processing approaches have been developed that can process and extract relations across multiple sentences (*e.g.* Leroy and Chen [12], Putejovsky and Castano [13]). We refer to Hirschman *et al* [14] as well as Mack and Hehenberger [15] for two excellent reviews of work in this area.

The potential of such text mining techniques is obvious. However, to establish how well these techniques will work on specific problems, systematic evaluations of their efficiency are needed. Until now, each group has established its own validation sets which are not comparable to each other. The text mining community has decided to tackle this problem by defining benchmark datasets and arranging international competitions. The Knowledge Discovery and Data Mining Challenge Cup 2002 (http://www.biostat.wisc.edu/~craven/kddcup/) had the task to build a system that automatically decides if a document contains experimental results about gene expression. Interestingly, the winning system used an information extraction technique that worked only on figure titles, document titles and abstracts ignoring the rest of the document. This system outperformed approaches from 17 other teams including all approaches using text categorization. TREC Genomics is a new event added to the Text REtrieval Conference (TREC, http://trec.nist.gov/), which focuses on information retrieval in biology and genomics.

2.2 Applications in life sciences

The question is how scientists can benefit from the limited capabilities of existing text mining systems today. We outline several applications where a direct use of text mining seems promising. Firstly, text processing can be used for classification. Stapley *et al* [16] predicted cellular locations of proteins from descriptions in abstracts with a performance between 50% and 80%. In addition, they identified some false annotations in databases which show the potential of text mining for an improved database curation. Glenisson *et al* [3] combined database information with textual information to expand functional annotations of proteins. In particular curators could benefit from annotation tools to automate

part of their work. But better integration of these methods into the workflow of researchers is still necessary.

Extraction of gene or protein interactions is a second field where many applications have been developed (*e.g.* Putejovsky and Castano [13], Blaschke *et al* [17]). The evaluation of these methods is difficult because of the lack of common benchmarks mentioned previously. To make these applications usable for end users, *i.e.* biologists and clinical researchers, effective and easy to use interfaces for visualizing the results are needed. PASTAWeb (Demetriou and Gaizauskas [18]) is an example of a good presentation of extracted results. Again to get the maximal benefit of information extraction methods, integrated systems are necessary that combine text mining with facilities to store, query, visualize and curate the information. GeneWays (Krauthammer *et al* [19]) can serve as good example of such a system.

Linking gene functions to the multitude of clinical phenotypes by means of information extraction is still in its infancy. The variety of concepts and terminologies used in a clinical environment greatly exceeds those used in genome research. Prior to information extraction, possible relations of the different entities have to be formalized, defining the concepts required to describe the phenotype characteristics of interest. For instance, physical examinations of patients having, *e.g.* cancer or depression have to deal with completely different concepts. This leads to very different terminologies and interrelations of terms.

As can be seen from previous examples, the representation of extracted information is an important aspect and can be formalized by ontologies. An overview of the current status is given in the next section.

3 Ontology development

Ontologies provide a means to express knowledge in a formal and compact representation accessible to both computers and humans, which avoid the shortcomings of natural language stated above. An ontology is given by a controlled vocabulary of terms including definitions of their meaning, and a specification of their interrelation. Due to this formal nature, it is easily possible to incorporate logical constraints and hierarchical structures into ontologies. As a consequence, facts expressed by the means of an ontology can be easily matched to data structures and thus stored in databases or used as input to applications. Usage of ontologies to define the fields in databases allows for highly accurate information retrieval and easy combination of information from different databases due to the underlying structure. A prerequisite for the generality of this combinatorial approach is a high acceptance of the used ontology.

The *Gene Ontology* (GO, Gene Ontology Consortium [20]) is currently the most widely used and known ontology in biology. Its aim is the description of knowledge about the roles of genes and proteins in cells. It is split into three independent categories describing biological processes, molecular functions and

cellular components. The usage of GO annotations in almost all sequence- and organism-based databases enables an easy comparison of proteins from different organisms or proteins with a similar function. However, the scope of the Gene Ontology is limited and ontologies covering other domains need to be developed and adopted in wider use.

In the medical world, the Unified Medical Language System (UMLS) Database (McCray and Miller [21]) is currently the most promising collection of medical terms and includes a classification of their meaning. It contains information about biomedical concepts and terms from many controlled vocabularies and classifications used in patient records, administrative health data, bibliographic and full-text databases, and expert systems. But a variety of different terminologies and classifications exist in parallel within UMLS. This is a problem for a mapping of terms from different sources.

While the current state and coverage of ontologies is still incomplete, several efforts exist to address these shortcomings. Hence, they are likely to be resolved within the next few years.

4 Conclusion

In future text mining will have a great impact in life sciences because free text is the most widely used medium for storing and communicating information both in research and clinical environments. It provides the tools to quickly extract relevant knowledge from large corpora of free text. However, transformation of the extracted information in a structured representation of information is required. For that purpose, ontologies are currently established which allow an easy access to relevant facts and enable the linkage to other information sources. Furthermore tools, *e.g.* for visualization, curation and annotation, must be developed to reach a broader acceptability and maximal employment of text mining in the biomedical community.

References

[1] Iliopoulos, I., Enright, A. J. & Ouzounis, C. A., TEXTQUEST: Document clustering of MEDLINE abstracts for concept discovery in molecular biology. *PSB 2001*, pp. 384–395, 2001.

[2] Gieger, C., Gaudan, S. & Kschischo, M., Managing biological knowledge using text clustering and feature extraction. *CompStat 2002: Proceedings in Computational Statistics, Short Communications and Posters*, eds. Klinke, S., Ahrend, P. & Richter, L., 2002.

[3] Glenisson, P., Antal, P., Mathys, J., Moreau, Y. & De Moor, B., Evaluation of the Vector Space Representation in Text-Based Gene Clustering. *PSB 2003*, pp. 391–402, 2003.

[4] Fukuda, K., Tsunoda, T., Tamura, A. & Takagi, T., Toward information extraction: Identifying protein names from biological papers. *PSB 199*8, pp. 707–718, 1998.

[5] Wilbur, W.J., Hazard, G.F., Divita, G., Mork, J.G., Aronson, A.R. & Browne, A.C., Analysis of biomedical text for chemical names: a comparison of three methods. *Proc. AMIA Symp 1999*, Washington, 1999.
[6] Narayanaswamy, M., Ravikumar, K. E. & Vijay-Shanker, K., A biological named entity recognizer. *PSB 2003*, pp. 427–438, 2003.
[7] Hanisch, D., Fluck, J., Mevissen, H.T. & Zimmer, R., Playing biology's name game: identifying protein names in scientific text. *PSB 2003*, pp. 403–414, 2003.
[8] Stapley, B. & Benoit, G., Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *PSB 2000*, pp. 529–540, 2000.
[9] Ng, S.-K. & Wong, M., Toward routine automatic pathway discovery from on-line scientific text abstracts. *GIW*, **10**, pp. 104–112 , 1999.
[10] Park, J. C., Kim, H. S. & Kim, J. J., Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *PSB 2001*, pp. 396–407, 2001.
[11] Yakushiji, A., Tateisi, Y., Miyao, Y. & Tsujii, J., Event extraction from biomedical papers using a full parser. *PSB 2001*, pp. 408–419, 2001.
[12] Leroy, G. & Chen, H., Automated extraction of medical knowledge using underlying logic from medical abstracts. *PSB 2002*, pp. 350–361, 2002.
[13] Putejovsky, J. & Castano, J., Robust relational parsing over biomedical literature: Extracting inhibit relations. *PSB 2002*, pp. 362–373, 2002.
[14] Hirschman, L., Park, J.C., Tsujii, J., Wong, L., & Wu, C.H., Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, pp. 1553–1561, 2002.
[15] Mack, R. & Hehenberger, M., Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discovery Today*, **7** (Suppl.), pp. S89 – S98, 2002.
[16] Stapley, B.J., Kelley, L.A. & Sternberg, M.J.E., Predicting the Sub-Cellular Location of Proteins from Text Using Support Vector Machines. *PSB 2002*, pp. 374–385, 2002.
[17] Blaschke, C., Andrade, M.A., Ouzounis, C. & Valencia, A., Automatic extraction of biological information from scientific text: Protein-protein interactions. *ISMB*, **7**, pp. 60–67, 1999.
[18] Demetriou, G. & Gaizauskas, R., Utilizing text mining results: The Pasta Web System. *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, pp. 77–84, 2002.
[19] Krauthammer, M., Kra, P., Iossifov, I., Gomez, S. M., Hripcsak, G., Hatzivassiloglou, V., Friedman, C. & Rzhetsky A., Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, **18** (Suppl. 1), pp. 249S–257S, 2002.
[20] Gene Ontology Consortium, Creating the gene ontology resource: Design and implementation. *Genome Res*., **11**, pp. 1425–1433, 2001.
[21] McCray, A.T. & Miller, R.A., Making the conceptual connections: the Unified Medical Language System (UMLS) after a decade of research and development. *J Am Med Inform Assoc*. **5**, 129–30, 1998.