Text mining for systems biology

Juliane Fluck[1] and Martin Hofmann-Apitius[1,2]

[1] Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[2] Bonn-Aachen International Center for Information Technology (B-IT), Dahlmannstraβe 2, 53113 Bonn, Germany

Scientific communication in biomedicine is, by and large, still text based. Text mining technologies for the automated extraction of useful biomedical information from unstructured text that can be directly used for systems biology modelling have been substantially improved over the past few years. In this review, we underline the importance of named entity recognition and relationship extraction as fundamental approaches that are relevant to systems biology. Furthermore, we emphasize the role of publicly organized scientific benchmarking challenges that reflect the current status of text-mining technology and are important in moving the entire field forward. Given further interdisciplinary development of systems biology-orientated ontologies and training corpora, we expect a steadily increasing impact of text-mining technology on systems biology in the future.

A substantial proportion of information relevant to the modelling and simulation of physiological and pathophysiological processes is not available from databases but is instead present in unstructured scientific documents, such as journal articles, reviews and monographies. Scientific communication in biomedicine is, by and large, still text based, because we all feel the need to report scientific advancements in a way that enables us to make use of the high expressiveness of natural human language. Technologies to identify useful biomedical information in unstructured text and to extract it automatically have been developed over the past 15 years. Initially focusing on finding and extracting information from PubMed abstracts, text-mining technology has advanced with impressive speed and is focusing increasingly on the extraction of complex biological context from full-text documents. A recent review on text-mining technologies enabling integrative biology provides a good overview of some of the academic technology developments made in this context [1].

Text-mining services for systems biology have to support the process of identifying and extracting information that is relevant to system description, modelling and simulation. Modelling of complex biological processes, spanning pathways to entire diseases, can be done at various levels of granularity using a range of mathematical modelling approaches [2]. Continuous models and quantitative models based on differential equations have been applied with great success where mechanistic details and kinetic parameters are known [3]. However, in cases where quantitative data are scarce, qualitative models, such as Boolean network models, have proven useful [4]. Another modelling approach that can deal with limited knowledge of mechanisms underlying systems behavior, but instead focuses on relationships represented as probabilities, are Bayesian network models or Belief Nets. Although Bayesian networks have been widely used in disease modelling [5], they depend on the availability of prior knowledge that can be used for the design of the Bayesian network and the computing of the prior distribution.

The first generation of text-mining applications has helped build Boolean models through entity recognition and co-occurrence networks [6]. Systems biology has since developed modelling strategies that represent information on causes and correlations in more detail; for example, OpenBEL [7] and, for pathway-related knowledge, BioPAX [8]. Disease classification systems and disease ontologies have facilitated the extraction of information that is relevant to modelling in systems biology; examples range from using ICD codes [9] on electronic patient records to the application of a dedicated ontology representing knowledge of Alzheimer' disease (AD)

Corresponding author:. Hofmann-Apitius, M. (martin.hofmann-apitius@scai.fraunhofer.de)

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2013.09.012

for comorbidity analysis [10]. Dedicated terminologies support the identification of biomarker candidates from the scientific literature [11], and the systematic analysis of scientific text using automated methods has led to the identification of putative biomarker candidates for AD [12]. Systems biology modelling is not only supported by the extraction of factual knowledge: in a dedicated literature-mining approach aimed at identifying scientific speculation and hypotheses, Malhotra *et al.* [13] were able to generate integrative models of biological processes that have been speculated to be causally involved in the generation of AD. Thus, such models represent the 'gray zone' of knowledge and are useful for the generation of new, testable hypotheses based on integrative models.

We foresee a need for more advanced text-mining methodologies that support the automated extraction of quantitative data from unstructured information sources. In this article, we briefly review the current state of automated named entity and relation extraction from scientific text, and shed light on the types of information needed for the above-mentioned modelling approaches in systems biology.

Overview of current state-of-the-art of named entity and relation extraction approaches

Biomedical systems can be represented as networks of triples, where entities are connected by relations. For the automated extraction of any kind of such triple, all entities connected through the relation have to be recognized by named entity recognition (NER) methods. The entity classes that are relevant for molecular systems biology models include genes, proteins, miRNAs, and chemical compounds. Additional relevant named entities for systems biology models are states such as phosphorylation or mutations; quantitative values (e.g. K_d values or concentration) provide useful information for the parameterization of models. Other entity classes that are essential for describing a biomedical system include descriptors of the experimental system, such as anatomical context (cell types or tissues), or phenotype information (e.g. clinical readouts).

Named entity recognition systems

For most of the molecular entities and clinical readouts, NER systems have already been developed. The usability of these systems depends on several issues: the recognition performance, the availability of the tool either as an open source or a commercial version, and the user-friendliness and support. In the text-mining community, several crucial assessments have been established over the past few years, providing a common platform for the comparison and exchange of methods and the standardization of input and output formats. For the recognition of gene and protein names, the BioCreative (http://www.biocreative.org) assessments focused on gene mention recognition and on normalization. The term 'normalization' describes the mapping of entities in text to reference entries in external resources, such as EntrezGene for gene names or UniProt for proteins. Normalization is essential for the automated extraction of systems biology networks from text. State-of-the-art recognition of gene and protein names has reached F measures of approximately 0.87 for gene mention recognition [14] and 0.81 for the recognition and normalization of human genes and proteins [15]. Database curators already use automated protein recognition in text and, in particular, the normalization of gene names to their corresponding sequence database entries, to save time in the manual extraction process from literature [16].

As another important class of entities for systems biology, chemical names are even more challenging. They span from single elements or ions, such as Ca^{2+} or Mg^{2+}, via the plethora of metabolites to chemical compounds and drugs. Owing to the specific challenges associated with the recognition of chemical entities in text, the best performance published for the recognition of chemical names in text reached an F score of 0.68 [17]. This is still far from a 'very good performance' and the research community is aware of this: the NER task in BioCreative IV in 2013 (http://www.biocreative.org/tasks/biocreative-iv/chemdner/) focused on the recognition of chemical compounds. However, the introduction of such community-wide assessment does lead to an overall improvement in the published results. The provision of published training data and separate test data leads to a reliable, open and transparent comparison of different approaches and to an uptake and evolution of the most successful methods.

Quantitative modelling

For quantitative modelling purposes (e.g. of pharmacokinetics), numerical values such as KD values or IC_{50} values need to be identified and extracted. The database BRENDA (BRaunschweig ENzyme Database; http://www.brenda-enzymes.org), which is a main information system of functional biochemical and molecular enzyme data, provides such literature data in one of its supplement databases, KENDA (Kinetic ENzyme Data). KENDA is a fully automatically generated resource based on extracting via text mining kinetic values and expressions (KM, K_i, k_{cat}, V_{max}, etc.), combined with the organism and enzyme name from more than 2.2 million PubMed abstracts [18].

In a recent paper, Wu *et al.* [19] demonstrated the value of a combination of a pharmacokinetics ontology, a well-annotated reference corpus, and a text-mining system that recognizes most types of information relevant to pharmacokinetics modelling. We expect to see more publications similar to this in the future: with a knowledge representation covering the major concepts in a scientific area of interest, an expert-annotated text corpus, and a text-mining system that builds on the controlled vocabulary that comes with the ontology.

Relation extraction

The recognition of entities is a prerequisite for the application of text mining in support of systems biology. Numerous methods have been published over the past decade, but we focus here mainly on those where the comparability of methods is given through an evaluation of the methods in the course of a public benchmarking competition.

For protein–protein interactions, five different corpora have been published over the past few years with slightly different focus in the annotation. Wang *et al.* [20] unified those corpora and, by doing, provided a common basis for the training and testing of new methods. Based on the aforementioned corpora, several machine-learning methods have been published that enable a simple classification on whether there is an interaction between two proteins in a sentence. Depending on the corpus used for

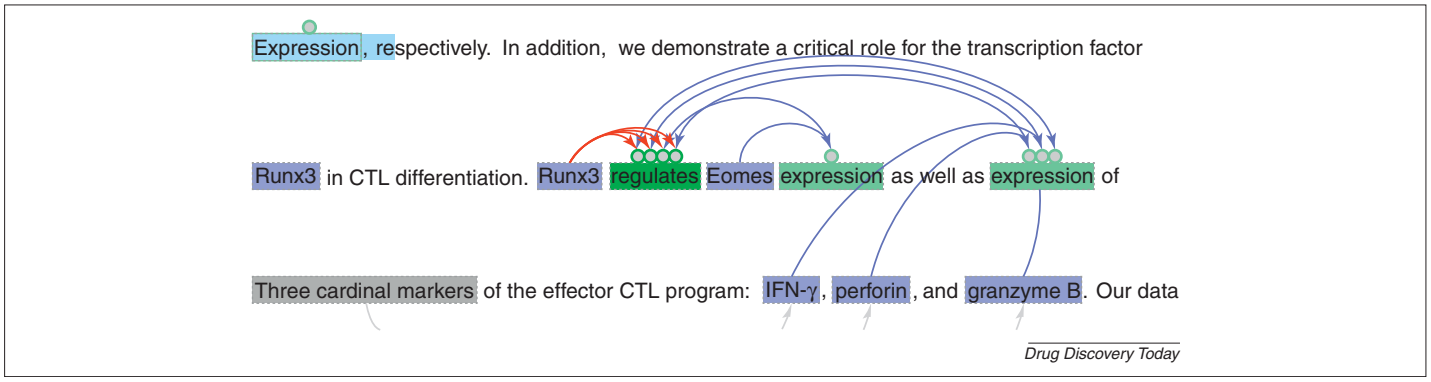

FIGURE 1

Example annotation for the GE task 2013. The BioLNP shared tasks provide training corpora for different relation extraction tasks. The depiction taken from http://2013.bionlp-st.org/tasks http://pubannotation.dbcls.jp/annsets/bionlp-st-ge-2013-training/pmcdocs/2626671/divs/0/annotations.json) shows the complexity of such annotations. All gene and gene products mentions are colored in blue, other unspecified entities in gray, trigger words for simple events are light green and the trigger word for the regulation is dark green. Relations are shown as arrows. Four simple events for expression (Eomes, IFN-gamma and granzyme B expression) and four regulation events from Runx3 to those expression events are annotated in this example.

training, higher precision of resulting relations or higher recall of all existing relations can be triggered.

These methods can be used for the automated construction of protein–protein networks or as an entry point for database curators. For integrative models representing causes and effects, controllers and effectors, detailed information on the type of relation between two entities, and the direction of the relation is required.

The BioNLP shared tasks

The relations described above can be extracted with approaches based on the BioNLP shared task data. The BioNLP shared task started in 2009 and takes place for the third time during 2013. It builds on the GENIA corpus, adding fine-grained information for relation extraction closer to the needs of systems biology. The BioNLP shared task provides well-annotated training and test data for event and relation extraction, and it attracted over 30 participating teams in 2011. Detailed information about the different tasks and the results for the 2011 shared task can be found at https://sites.google.com/site/bionlpst/ and the special issue of *BMC Bioinformatics* [21]. The BioNLP task addresses regulation relations as well as protein-modification events. An example annotation based on simple events such as 'Eomes expression' and regulation events such as 'Runx3 regulates Eomes expression' is shown in Fig. 1. In the 2013 shared task, a pathway curation challenge has been included that specifically addressed chemical compound relations (http://2013.bionlp-st.org/tasks/pathway-curation).

How far are the systems able to deliver?

Although an overall improvement in relation extraction systems can be observed over the past few years, the systems are not yet ready to deliver networks of causal relations in a fully automated way. The extraction accuracy for full text, at least for NER, but probably also for relationships, seems more challenging: the performance of protein mention detection has been reported to drop substantially in full text [22]. Another important issue is the availability and format of the full-text resources. Owing to copyright and terms of use imposed by the publishing industry, most full-text articles must not be parsed automatically; thus, automatic text analytics is substantially impaired by current copyright regulations. Furthermore, most of the full-text literature is only available as PDF and not as computer-readable, structured text or XML. This leads to more errors through the necessary preprocessing of documents, which sometimes requires OCR back-translation of documents. Document preprocessing and PDF representation of scientific knowledge are challenges that have

TABLE 1

Examples of experimental settings to prove certain relations[a].

Experimental variation	Example sentences
Different expression in wild-type versus mutant or know out, in or down organisms	In wasted mutant tissues, even though *eEF1A-2/S1* protein is absent
Engineered fusion vectors to enable easy measurement of expression (e.g. CAT, luciferase activity) in transfected cells	*HA+AFX* induced a pronounced increase in *CAT* activity
Engineered fusion vectors to enable constitutively expressed variants of proteins	*gagPKB* increased the cell surface density of *GLUT4myc*
Engineered fusion vectors to enable mutated inactive variants of proteins	*AAA-PKB* almost entirely blocked the insulin-dependent increase in surface *GLUT4myc*

[a]Experimental evidence is crucial for the reliability of extracted relations from literature but are not easy to extract and interpret using automatic systems. Problems for named entity recognition are the name variations or missing names in the sentences. For the relations, the correct relation partner is missing when authors state a changed expression in mutated tissue or of CAT activity. Correct interpretation of relations for engineered proteins is an additional problem for automatic systems.

not yet been tackled properly in current text-mining applications. Publishers, scientists and life-science companies should come together to think about mutually acceptable business models and better output formats, such as XML, for automated knowledge extraction.

An important aspect of text mining supporting systems biology concerns the experimental proof for the observed relation. Experimental evidence is usually crucially assessed by human curators; for example, in the IntAct data set covering protein interaction information, the interaction detection method is recorded routinely (http://www.ebi.ac.uk/intact/). As a support for IntAct curators, systems have been developed for the recognition of experimental protein interaction detection methods in addition to the protein interactions themselves (http://biocreative.sourceforge.net/bc2_ppi_ims.html).

The quality of published data and expert confidence of shown relations depends mainly on the kind of experiment performed. Examples of such experimental details are shown in Table 1. To enable the generation of high-quality data, this experimental evidence has to be provided to experts in addition to the relations found. Existing text-mining systems do not yet focus on the extraction of such evidence. Moreover, additional interpretation steps are necessary to translate the information given in text, such as 'the mutated inactive variant AAA-PKB almost entirely blocked increase in surface *GLUT4myc*,' into the formalized triple 'PKB increase surface expression of GLUT4'. So far, no text-mining system is able to do such an interpretation.

Further issues are the extraction of information from images and tables, which can be easily read by human experts. For image recognition and annotation, research for classification and annotation are underway (e.g. [23]). Similar table recognition and the resolution of table columns and rows are under development (e.g. [24]). However, at present, the automatic semantic interpretation of images and tables is not feasible.

End-user interfaces

Although text mining is not yet ready for the fully automated extraction of accurate content, current systems can substantially accelerate the building of knowledge bases relevant to systems biology modelling and model parameterization. Intuitive and well-designed user interfaces are a key component to integrate automatic systems into a knowledge acquisition workflow. In the BioCreative assessments, database curators directly take part in the design and the evaluation of the challenge. In the user-interactive task (http://www.biocreative.org/tasks/biocreative-iv/track-5-IAT/), database curators assess the quality of the extracted information as well as the usability aspects of the tools. Information retrieval and question-answering applications are another area to support end users with appropriate tools to retrieve relevant information. Over the past few years, the TREC (Text Retrieval Conference) assessments have provided several genetic (http://ir.ohsu.edu/genomics), medical (http://www-nlpir.nist.gov/projects/trecmed/2011/), and chemical (http://www.ir-facility.org/trec-chem) question-answering and retrieval tasks. In 2013, the Question Answering for Machine Reading Evaluation (QA4MRE) supported a new task 'Machine reading of biomedical texts about Alzheimer's Disease in CLEF (Conference and Labs of the Evaluation) 2013 (http://celct.fbk.eu/QA4MRE/index.php?pagePages/biomedicalTask.html). The questions to be answered focused partly on biochemical questions such as 'What experimental approach was successful to inhibit *in vivo* the production of amyloid β' or ?What is the product of the transformation of testosterone carried out by aromatase'. The resulting applications could lead to tools directly supporting information retrieval and question answering in the systems biology context.

BOX 1

Conclusions

- There are already methods and applications available and in use to support the task of extraction of published knowledge.
- There is the need to develop existing methods further and to provide relevant training data.
- Interdisciplinary work is required to develop methods that support systems biology directly.
- Equally essential are appropriate tools for the visualization, analysis and correction of extraction results for end users to work with the extracted data and to transform the information into the high-quality knowledge needed for systems biology.
- The availability and the quality of full-text corpora is currently problematic.

Outlook

For direct support to develop systems biology text-mining applications, the OpenBEL community [7] (http://www.openbel.org/) might be a platform for interdisciplinary work to develop relation-extraction approaches further. BEL documents normally contain extracted relations together with relevant evidence sentences and could serve as perfect training data. The proper preparation of BEL documents containing relevant relations and the associated evidence sentences from freely available full-text articles might serve as a good training corpus to provide optimal results for the automated generation of causal models encoded in OpenBEL. It would provide text-mining developers with good annotated data with which to train their systems and give system biologists direct feedback to the extent that text-mining systems are able to deliver. In addition, current text-mining applications could directly be used when conversion processes from text-mining output formats to system biology formats are available. First steps have been taken by providing conversion from the BioNLP ST format to the BEL format but additional layers of the semantic interpretation to translate linguistically based relations to relevant relations are still missing (Box 1) [25].

References

1 Rebholz-Schuhmann, D. *et al.* (2012) Text-mining solutions for biomedical research: enabling integrative biology. *Nat. Rev. Genet.* 13, 829–839

2 Tegnér, J.N. *et al.* (2009) Computational disease modeling: fact or fiction? *BMC Syst. Biol.* 4, 56

3 Bachmann, J. *et al.* (2012) Predictive mathematical models of cancer signalling pathways. *J. Intern. Med.* 271, 155–165

4 Hickman, G.J. and Hodgman, T.C. (2009) Inference of gene regulatory networks using Boolean-network inference methods. *J. Bioinform. Comput. Biol.* 7, 1013–1029

5 Schwartz, S.M. *et al.* (2012) A systematic approach to multifactorial cardiovascular disease: causal analysis. *Arterioscler. Thromb. Vasc. Biol.* 32, 2821–2835
6 Ananiadou, S. *et al.* (2010) Event extraction for systems biology by text mining the literature. *Trends Biotechnol.* 28, 381–390
7 Slater, T. *et al.* (2013) Speaking of systems biology: knowledge representation languages for pathway maps and computable biological networks. *Drug Discov. Today* http://dx.doi.org/DRUDIS-D-13-00062
8 Demir, E. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* 28, 935–942
9 Roque, F.S. *et al.* (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.* 7, e1002141
10 Malhotra, A. *et al.* (2013, July 3) ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimers Dementia* http://dx.doi.org/10.1016/j.jalz.2013.02.009
11 Younesi, E. *et al.* (2012) Mining biomarker information in biomedical literature. *BMC Med. Inform. Decision Making* 12, 148
12 Greco, I. *et al.* (2012) Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. *J. Transl. Med.* 10, 217
13 Malhotra, A. *et al.* (2013) 'HypothesisFinder': a strategy for the detection of speculative statements in scientific text. *PLoS Comp. Biol.* 9, e1003117
14 Smith, L. *et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome Biol.* 9 (Suppl. 2), S2
15 Morgan, A.A. *et al.* (2008) Overview of BioCreative II gene normalization. *Genome Biol.* 9 (Suppl. 2), S3
16 Hirschman, L. *et al.* (2012) Text mining for the biocuration workflow. Database http://dx.doi.org/10.1093/database/bas020
17 Rocktäschel, T. *et al.* (2012) ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* 28, 1633–1640
18 Schomburg, I. *et al.* (2013) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* 41, D764–D772
19 Wu, H.Y. *et al.* (2013) An integrated pharmacokinetics ontology and corpus for text mining. *BMC Bioinform.* 14, 35
20 Wang, Y. *et al.* (2010) Improving the inter-corpora compatibility for protein annotations. *J. Bioinform. Comp. Biol.* 8, 901–916
21 Kim, J.D. *et al.* (2012) Selected articles from the BioNLP Shared Task 2011. *BMC Bioinform.* 13, S11
22 Cohen, K.B. *et al.* (2010) The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinform.* 11, 492
23 Giordano, M. *et al.* (2013) iMole, a web based image retrieval system from biomedical literature. *Electrophoresis* 34, 1965–1968
24 Fang, J. *et al.* (2011) A table detection method for multipage PDF documents via visual separators and tabular structures. In *Proceedings of the 2011 International Conference on Document Analysis and Recognition*, IEEE Computer Society. pp. 779–783
25 Fluck, J. *et al.* (2013) BEL networks derived from qualitative translations of BioNLP shared task annotations. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, Association for Computational Linguistics. pp. 80–88