Drug Discovery Today: Technologies
Vol. 2, No. 3 2005

Editors-in-Chief
Kelvin Lam – Pfizer, Inc., USA
Henk Timmerman – Vrije Universiteit, The Netherlands

Knowledge management

Information extraction technologies for the life science industry

Juliane Fluck[1,*], Marc Zimmermann[1], Günther Kurapkat[2], Martin Hofmann[1]

[1]Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[2]TEMIS Deutschland GmbH, Kurfürstenanlage 3, 69115 Heidelberg, Germany

Access to relevant information and knowledge is essential for all steps of the drug discovery process. However, keeping track of relevant information in publications and patents becomes a real challenge for scientists and managers in industrial research. Computer-aided information extraction (IE) systems have been developed to support the work of scientists by extracting relevant information from scientific publications and presenting it in an aggregated, condensed form. In this review, we will give an overview on current information extraction strategies in the life sciences with a special focus on biological entity recognition and more recent developments towards the identification and extraction of chemical compound names and structures.

Section Editor:
Manuel Peitsch – Novartis, Basel, Switzerland

Introduction

Despite the fact that we hear a lot about rational approaches in genomics, biochemistry and pharmacology; life sciences are still by and large descriptive sciences. Most of the scientific results relevant for drug discovery, for example, the link between interacting proteins, molecular pathways and diseases are described in unstructured scientific text. To make the vast amount of published information in the life sciences accessible for intelligent, computer-based services such as information retrieval, information discovery and clustering, several groups have started to work on strategies to automatically identify and extract relevant information from scientific publications. In the following, we will review current approaches in information extraction (IE) in molecular biology with a particular focus on their application in genomics research. Furthermore, we will allude to recent attempts at extending these approaches beyond the current scope on biological term recognition towards an application of IE technologies in chemistry and pharmacology.

*Corresponding author: J. Fluck (juliane.fluck@scai.fraunhofer.de)

Information extraction: state of the art in biomedicine

The majority of IE approaches in the life sciences have focused on molecular biology and genomics information so far. An increasing number of reviews on text mining in the biomedical domain is indicative of the strong interest of the scientific community in this sort of approach (cf. [1,2]). In essence, a typical IE system comprises at least some of the modules listed in Table 1.

Applications of IE technology in the life sciences typically concentrate on genes and proteins and the relationships between these entities. The strategies vary from using simple co-occurrence of names or dictionary entities in abstracts or in sentences to deployment of natural language processing (NLP)-based methods of tagging and parsing for the recognition of protein–protein or protein–gene relations. In the following, we give some examples of typical applications.

Curation of databases

Database curation is among the applications where IE helps to reduce the workload for human experts. Recent reports on

1740-6749/$ © 2005 Elsevier Ltd. All rights reserved. DOI: 10.1016/j.ddtec.2005.08.013

Glossary

F-score: 2 × precision × recall/(precision + recall).
Precision: true positive matches/(true positive matches + false positive matches).
Recall: true positive matches/(true positive matches + false negative matches).

the use of *Textpresso* for the curation of entries in Flybase [3] or the use of IE for the extraction of gene–disease relationships for BRENDA [4] underline the steadily growing role that IE plays for high-quality information sources in biology and genome research. It is more than probable that the work of most database curators will be supported by IE technology in the near future.

Interpretation of gene expression data

Protein or gene interaction networks generated by IE systems have been widely used to explore the context of significantly regulated genes found in large-scale expression experiments. Pure statistical approaches for the interpretation of gene expression experiments are often highly reliable with respect to the statistical analysis of correlation. However, the resulting models are difficult to interpret from a biological point of view. Therefore, several groups have integrated IE approaches with statistical analysis for a contextual interpretation of gene exp-ression data. Jenssen *et al.* [5] used a simple co-occurrence approach of gene names similarity of Gene Ontology (GO) annotations for the interpretation of expression data. Tifin *et al.* [6] applied a combination of text- and data-mining technology for the interpretation of gene expression data and their possible association with disease relevant terms. Pan *et al.* [7] deployed text-mining technology to identify associations of transcription factors with biological processes as defined by Gene Ontology. Chiang *et al.* [8] introduced a system that uses multiple dictionaries for diseases and gene names to assign biological and medical context to individual genes. Finally, Albert *et al.* [9] used text mining to establish a protein-interaction database focusing on the target protein family of nuclear hormone receptors.

Our own experience using co-occurrence for the generation of protein and gene interaction networks showed a high rate of false positive correlations in these networks. Recently, we integrated grammar-based interaction networks to identify subnetworks highly relevant according to expression measurements [10]. *P*-values quantifying differential expression were mapped directly onto the extracted interaction network. Regions of interests within the large knowledge-based network could easily be identified using this approach (Fig. 1, for example, network).

Despite the broad spectrum of applications for IE systems, the use of IE technology is not yet routined in academic and industrial life science research. Among the reasons for the slow adoption of IE technology are concerns about the performance and a lack of standards for the comparison of the different methods. All the IE systems published so far have in common that they have been evaluated on very small benchmark sets, and it remains to be demonstrated how these systems perform on larger corpora. None of these IE systems seem to fit all the requirements in the life science domain. As mentioned above, co-occurrence networks often suffer from low PRECISION rates. Rule-based systems (cf. [11,12]) as well as machine learning-based systems (cf. [13,14]) reach higher precision rates (above 80%) at the expense of comparably low RECALL rates. As most of the IE tools in the biomedical domain use algorithms originally developed for the analysis of newspapers (cf. [15,16]), significant parts of these IE systems (e.g. part of speech tagging [17] and noun phrase chunking [18]) need to be adapted to the special requirements of the biomedical domain. Adaptations such as the integration of biological lexica and the tuning of tokenization algorithms are essential to cope with the performance loss owing to the shift of the application domain. In addition to the adaptations of existing core technology, new, specific challenges arise in the life science domain through the use of large domain-specific terminology, which is often highly ambiguous. Moreover, the identification and extraction of representations of chemical compounds (e.g. as IUPAC names) provides a completely new challenge to IE systems. In the following, we will emphasize on the state of the art of different techniques for biomedical named entity recognition and we will highlight recent developments in the field of IE for chemistry.

Table 1. Elements of an information extraction (IE) system

- **Tokenization** is a process of breaking the text up into its constituent units which are words (e.g. token), sentences or also parts of text
- **Part of speech tagging** labels the words with different categories based on the role the respective word plays in the sentence (e.g. verb, noun, preposition, adjective)
- **Named entity recognition** identifies complex noun phrases as entities and classifies them in high-level categories (e.g. protein or cell-type) or maps them directly to fine-grained instances (e.g. to a certain gene entity defined through a gene database identifier)
- **Co-occurrence analysis** identifies relations through the joint occurrence of entities in certain text parts (e.g. abstract or sentence) sometimes accompanied with further statistical frequency analysis
- **Syntactic analysis** tries to identify the structure of sentences and to associate syntactic roles to the different parts of each sentence (e.g. subject, predicate, object). Whereas a complete analysis (=*parsing*) of whole sentence structures is too slow for unrestricted text, *shallow parsing* methods have been established. They focus on problem-orientated and pragmatic solutions (e.g. find protein–protein relations)

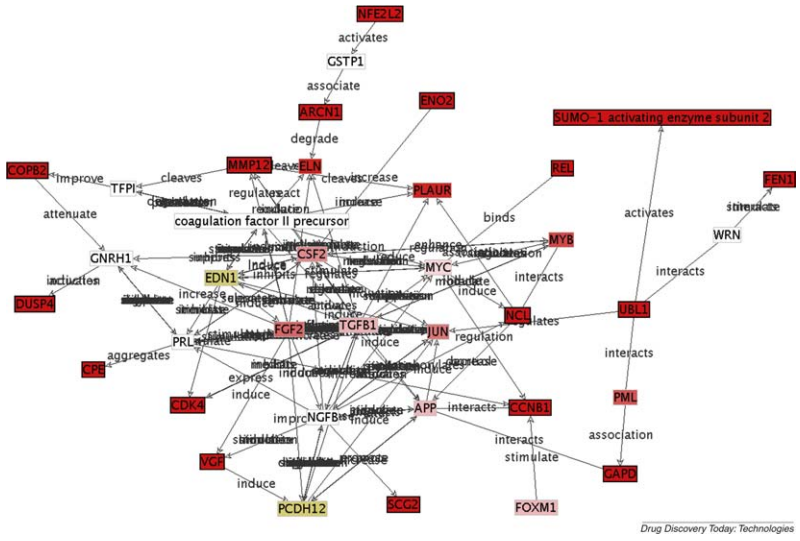


Figure 1. Visualization of expression data with text-mining networks. Proteins in the network are visualized as nodes and interaction between the proteins as edges. The *P*-values of expression data from microarray experiments were mapped to the corresponding proteins in the extracted network. The network was extracted using a shortest path search with maximum step-length of 2 between genes with a defined range of *P*-values. The degree of red coloration correlates to the significance of the calculated *P*-value (beige: no *P*-value determined).

Named entity recognition of protein and gene names

The unambiguous identification of biological entities or processes is a fundamental requirement for information extraction in the biomedical domain. Some of the common challenges associated with biological name recognition concern the handling of yet unknown words (e.g. new gene names), multiple names for the same entity (*synonymy*) and the identification of entities composed of more than one word (*multiword terms*). In addition, identical names are used to identify different proteins, genes or other biological entities (*polysemy*). Also common word names are often used as gene names. Some examples are shown in Box 1.

Box 1. Examples of gene names and their synonyms:

- MAPK14 – synonyms: mitogen-activated protein kinase 14; EXIP; Mxi2; **CSBP**; Csaids binding protein; MAX-interacting protein 2; stress-activated protein kinase 2A; p38 mitogen activated protein kinase; mitogen-activated protein kinase p 38; MAP KINASE 14, among others.
- HNRPK – synonyms: heterogeneous nuclear ribonucleoprotein K; **CSBP**; TUNP; **ROK**; hnRNP K.

Examples of acronyms and their long forms (extracted from Medline abstracts):

- **CSBP**: clinostatic systolic blood pressure; casual systolic blood pressure; carotid-subclavian bypass.
- **ROK**: Republic of Korea; rokitamycin.

Examples for common word synonyms:

- WAS, HAT, ICE, AGAR, BRAIN, DREAM, FORM, FOX, among others.

During the past years, several different methods for named entity recognition in the biomedical domain have been developed which can be categorized in two main classes. The first class handles recognition and classification of names in text (e.g. protein, DNA-molecule, cell type). Entities will be recognized through special text features (e.g. uppercase, contains numbers, ending -ase) which can be used in rule-based (e.g. [19–21]) as well as machine learning-based techniques (e.g. [22,23]). Known names as well as new names can be recognized without the use of dictionaries. However, entities identified in text could not be mapped directly to biological entities in databases. The ability to link information extraction results to experimental data through mapping of, for example, gene or protein names to the respective database entities is essential for contextual data interpretation. This problem was tackled by the second class of named entity recognition approaches. Here, dictionaries are used to identify the corresponding names in the text (cf. [24,25]). The performance of these systems of course depends on the comprehensiveness of the dictionaries and the ability to recognize spelling variants of dictionary names.

In the past 2 years independent assessments for the evaluation of these methods have been set up. In the Bio-Entity Task at JNLPA (for an overview cf. [26]) and the BioCreAtIvE (http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html) assessment task 1a (for an overview cf. [27]), recognition and classification of names in text (e.g. protein, DNA-molecule and cell type) were compared by different methods. The performance rates in these assessments are far below the rates, which could be achieved in the general newspaper domain where balanced precision and recall between 93

and 95% can be reached for the identification of person, organization and location names. The reasons for this difference are most probably the higher degree of ambiguity of domain names in biomedical text corpora and/or annotator disagreements in the provided corpora.

The ability to recognize gene and protein names and map them directly to the respective database entries was assessed in BioCreAtIvE Task 1B where, depending on the organism, good annotation performance could be achieved (cf. [28] for an overview). For yeast, most participants in the competition reached more than 90% recall and precision owing to the very stringent nomenclature that is in use for this organism. For *Drosophila*, despite the high amount of ambiguous and common word gene names, precision and recall values of about 80% were reached. For the mouse dictionary, only slightly better results than in fly could be achieved. The reasons for this are most probably ambiguous names of acronyms, which are used in other context and the fact that many names are also shared with other organisms (e.g. human).

Specific challenges in chemical entity recognition

In the field of compound names and chemical nomenclature, we have different issues to consider. In publications dealing with small molecules and their pharmacological and toxicological effects, we are confronted with a large variety of syntactical and semantically different compound descriptions. Chemicals can be described in publications by trivial names (e.g. brand or trade names), by registry numbers (e.g. database or project identifiers), by systematic naming schemes (i.e. nomenclatures such as IUPAC [29], formal descriptions like SMILES [30] or sum formula) and even by depictions of the chemical structure.

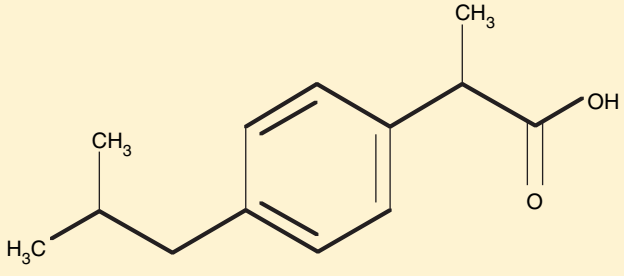
The complexity of the name space that exists for well-known small molecules is illustrated in the following example of 2-(*p*-isobutylphenyl)propoinic acid, a compound widely known as Ibuprofen. We count more than 500 synonyms in the literature and in compound databases for Ibuprofen (PubChem database, http://pubchem.ncbi.nlm.nih.gov) [31]; and even following the rules of IUPAC nomenclature we can assign three different names to this compound. The CAS registry assigns a single code to Ibuprofen; however, different salts of the molecule are represented by different codes (Box 2).

Only a very limited number of named entity recognition approaches are described in the literature for the recognition of chemical entities. Narayanaswamy *et al.* [16] presented a rule-based method; however, this approach was tested only on a very small benchmark set. Other systems used dictionaries and simple string matching without any evaluation of recall and precision (e.g. [32]). One potential problem is that the possible chemical space of small, drug-like molecules is incredibly huge and, therefore, the dictionaries will become large and cumbersome. Therefore, we are looking at alternative ways of name recognition, for example, through generalized chemical 'scaffold' patterns for IUPAC nomenclature and SMILES representations. Preliminary tests in our laboratory indicate that new tokenization strategies (in comparison to the newspaper and biological domain) have to be integrated to make this approach feasible.

By contrast, we deal with trivial names where only the use of dictionaries allows mapping of the chemical structure. For the recognition of trivial names we tested the ProMiner system, which was assessed on benchmark sets for protein and gene entity recognition in different organisms (see named entity recognition of protein and gene names) and performed very well in the BioCreAtIvE [33] competition. The ProMiner system [34] possesses a semi-automatic dictionary generation module. Compound names have been taken from the chemical part of MeSH (http://www.nlm.nih.gov/mesh/meshhome.html) and ChEBI (http://www.ebi.ac.uk/chebi/) to construct a simple chemical dictionary. Trivial names were merged with the help of the CAS registry number and incor-

Box 2.

Depiction:



Synonyms: Abbifen, Aches-N-Pain, ACT 3, Actifen . . . Zofen
IUPAC: 2-(*p*-isobutylphenyl)propoinic acid, *p*-isobutylhydratropic acid, α-methyl1-4-(2-methylpropyl)benzeneacetic acid
Molecular formula: $C_{13}H_{18}O_2$
SMILES: CC(C)CC1=CC=CC=C1C(C)C(=O)O (others can be created)
CASRN: 15687-27-1 (different salts: 31121-93-4, among others)

porated in a base thesaurus. Based on such dictionary, a string match procedure is used for name detection in the text. In a first approach, this procedure led to about 80% precision for randomized integration of the trivial names in a biomedical corpus [35].

For effective recognition of chemical names, trivial names as well as IUPAC and SMILES nomenclature, dictionary-based approaches and pattern-based recognition have to be combined in further developments.

Information extraction from images

Images representing chemical structure information can be found in nearly all electronic sources of chemical information (e.g. journals, patents and web interfaces of chemical databases). Nowadays these images are generated with special drawing programs [36], either automatically from computer-readable file formats (i.e. collections of molecules) or by the chemist through a graphical user interface. Although these programs can produce and store the information again in a computer-readable format, the information is published as a bitmap image. As a consequence, the structure information can no longer be used as input to chemical analysis software packages.

To make chemical information contained in drawings of chemical structures accessible for computer programs, a few projects have been started to convert bitmap images of molecules back into machine-readable form. In general, there are two different strategies for the recognition of patterns in images. *Statistical pattern recognition* [37] uses the so-called features from the image representation (e.g. color frequencies, pixel patterns). Supervised learning algorithms can be trained on the feature vectors to recognize patterns in test sets of images. The other is the *structural pattern recognition* [38]. This strategy uses abstract data types such as strings, trees and graphs instead of numerical types. These *concepts* can describe relationships between objects (e.g. geometric, spatial) and allow the hierarchical composition of substructures. To be recognized, objects are compared with model objects using matching algorithms (e.g. string alignment, graph matching). Both strategies can be combined with each other. Three projects on chemical structure reconstruction have been described in the literature: the Kekulé and the CLiDE project and our own approach called compound structure reconstruction (CSR):

- Kekulé [39]: the workflow of the Kekulé project consists of vectorization of TIFF images. Optical character recognition (OCR) techniques and neural networks are used to identify special symbols such as chiral bonds and text representations.
- CLiDE [40]: the CLiDE project uses segmentation algorithms for monochrome bitmaps to identify connected components. These components are grouped into graph primitives. There are special primitives for the chemical context (e.g. superatoms and bond types).
- CSR [41]: we use image readers and vectorization software in combination with classic OCR, machine learning, graph matching and a chemical structure editor. Our prototype executes a workflow comprising image preprocessing, graph matching and molecule reconstruction. We use a template database of common scaffolds in drug-like molecules and compare them with the extracted graph from the image. In the last step, the information on the atom types from the OCR process is merged with the molecular skeleton.

Similar to the situation with textual IE in chemical literature, we are missing a test corpus consisting of images and the associated structural information. We, therefore, assembled our own test corpus consisting of a diverse set of molecules and their depictions (T. Fey, Master's thesis, Fachhochschule Bonn-Rhein-Sieg, 2004). The resulting test corpus comprises structures of the top 100 'blockbusters' (most sold drugs) from the year 2002 (RxList LLC, http://www.rxlist.com).

In a first benchmark test with CLiDE and CSR, we found that the same problems pose major challenges for both systems: recognition of chiral bonds (there are many different ways to draw them), overlapping entities (e.g. bridged ring systems) and OCR (misleading text as bonds and vice versa). However, the ability of CSR to learn from human interference will help in overcoming these problems.

Commercial implication of information extraction solutions in drug discovery

The current information landscape is characterized by two extremes, unstructured data in textual documents and structured data stored in database records. Access to information is mediated through search engines and database systems, respectively. Both approaches, although are helpful and essential and therefore widely deployed, do have inherent limitations. The major limitations of the information retrieval (IR) systems are: the need to know what to look for, the lack of semantics, that is, missing concept searches and finally the fact that the search results in a list of documents containing facts but not a presentation of aggregated facts themselves. Database systems are limited by the low flexibility and fixed granularity of the information provided; the lack of context and the difficulties in integrating information from other sources such as proprietary documents or databases from other vendors.

IE systems are ideal tools to complement the current approaches. Such systems transform unstructured textual data into structured information while keeping the context available. This mainly results in two types of applications: a dynamic text analysis system and a topic-driven extraction system (Fig. 2).

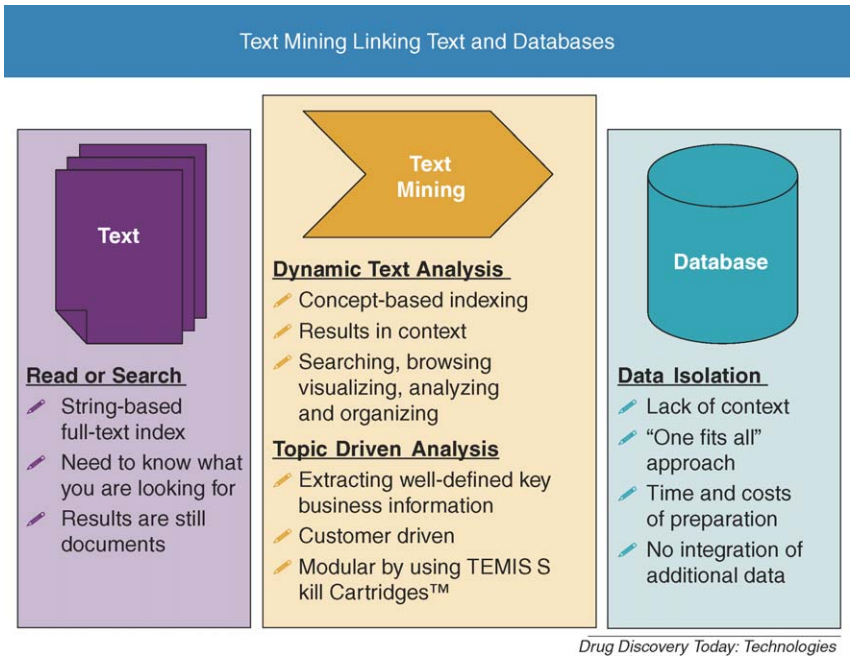


Figure 2. Role of commercial text-mining applications: closing the gap between documents and database information. Generalized role of text mining in commercial application. Indexing of text documents allows for rapid searching and retrieval of relevant documents; however, the result of such search is still a list of documents that need to be read to find relevant information. By contrast, databases store relevant information in a fixed format with the lack of context and the difficulty to integrate data. Text mining, however, renders the user flexible and allows to define and to extract relevant information defined by a scientific or management task.

Text analysis recently raised a lot of interest, particularly in the field of patent analysis. This is related to a new analysis paradigm of textual data, which allows not only a much easier way to search for information but also to explore the information hidden in patent documents. Such an analysis includes as a first layer, the extraction and normalization of biomedical entities which enables a larger group of end-users to identify relevant patents using their day-to-day tools such as chemical substructure searches. Furthermore, as the extracted information is stored, the content of patent documents can directly be explored, that is hidden information is directly accessible (e.g. show all targets for a given indication area) and patents can, therefore, be used to generate new research ideas. By adding semantic relationship analysis as a second component, highly relevant information such as bioactivity data or disease–compound interactions are directly accessible from the literature. Finally, clustering and categorization within different dimensions (e.g. disease versus targets) enables an in depth analysis of the patent scope of interest. In summary, the application of IE technologies introduces a new and powerful analysis methodology for textual data, eases search and exploration and thus, makes crucial business information accessible to a larger group of end-users.

The topic-driven analysis focuses on the extraction of well-defined key business information. With its high data volume and frequent updates, competitive and scientific intelligence are formidable examples areas. IE systems systematically extract and categorize in the background the information from the data stream according to user needs. Examples are the extraction of competitive (financial, restructuring, law cases, licensing, partnership, co-development) [42] or scientific information (protein–protein interactions, protein–ligand interactions, epidemiological information and side effects) (Atlanta Pharma, TEMIS group and iAS interactive systems, press release, unpublished). Thus, the approach supports innovation, knowledge sharing and enables the user to keep pace with the growing information.

In addition to the approaches outlined above, there is a strong impact of text mining on knowledge integration and navigation. Entity recognition is a substantial step towards semantic integration of data coming from different sources. This approach has actually been followed by some larger corporation such as Novartis [43] (Atlanta Pharma, TEMIS

Links

- http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html
- http://www.temis-group.com/
- http://www.scai.fhg.de
- http://www.biowisdom.com/
- http://www.simbiosys.ca/clide/index.html
- http://www.insightful.com/
- http://www.clearforest.com/index.asp

Related articles

Cohen, A.M. and Hersh, W.R. (2005) A survey of current work in biomedical text mining. *Brief Bioinform.* 6, 57–71
Shatkay, H. and Feldman, R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.* 10, 821–855
Krallinger, M. *et al.* (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today* 10, 439–445
Rebholz-Schuhmann, D. *et al.* (2005) Facts from text – is text mining ready to deliver? *PLoS Biol.* 3, e65
Hale, R. (2005) Text mining: getting more value from literature resources. *Drug Discov. Today* 10, 377–379

group and iAS interactive systems, press release, unpublished), similar efforts are going on at GSK based on ontologies developed at BioWisdom, Cambridge, UK. The extracted concepts have been used to greatly extend the functionality of corporate search engines, enabling searches of the type 'show me all proteins that have been described in the context of Alzheimer's Disease'.

Conclusions

IE approaches are about to become an evident and integrated part of knowledge management in the life sciences industry. The initial main focus of academic research on gene and protein name recognition has been complemented by approaches addressing the requirements of the pharmaceutical and biotechnology industry on chemical entity recognition. Current IE prototypes are capable of extracting chemical and pharmacological information from text and chemical structure information can be reconstructed from depictions with acceptable precision. Therefore, we expect IE technologies to develop into a corner stone of IT solutions supporting drug discovery in the near future.

Outstanding issues

- Improvement of precision and recall for chemical entity recognition.
- Efficient mapping of chemical entities from text to structure representations.
- Generation of large biological, pharmacological and toxicological literature test corpora for training and benchmarking purposes.
- Integration of chemical entity recognition from text with chemical structure reconstruction from images.

References

1 Shatkay, H. and Feldman, R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Cell Biol.* 10, 821–855
2 Gieger, C. *et al.* (2003) The future of text mining in genome-based clinical research. *Biosilico* 1, 99–103
3 Muller, H.M. *et al.* (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* 2, e309
4 Hofmann, O. and Schomburg, D. (2005) Concept-based annotation of enzyme classes. *Bioinformatics* 21, 2059–2066
5 Jenssen, T-K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28
6 Tiffin, N. *et al.* (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucl. Acids Res.* 33, 1544–1552
7 Pan, H. *et al.* (2004) Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining. *Nucl. Acids Res.* 32, W230–W234
8 Chiang, J.H. *et al.* (2004) GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics* 20, 120–121
9 Albert, S. *et al.* (2003) Computer-assisted generation of a protein-interaction database for nuclear receptors. *Mol. Endocrinol.* 17, 1555–1567
10 Gieger, C. *et al.* (2003) Using text mining networks for the context specific interpretation of gene expression data. *Biometric. J.* 46 (Supplement Abstracts of the Joint Meeting of the IBS-DR and the DAE, 56)
11 Saric, J. *et al.* (2004) Extracting regulatory gene expressions expression networks from PubMed. *ACL'04/EACL'04 – Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics & 10th Conference of the European Chapter of the Association for Comptational Linguistics*
12 Friedman, C. *et al.* (2001) Genies: A natural-language processing system for the extraction of molecular pathways from journal articles. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB)* pp. 74–82
13 Xiao, J. *et al.* (2005) Protein–protein interaction extraction: a supervised learning approach. *First International Symposium on Semantic Mining in Biomedicine (SMBM)*, 10th–13th April, European Bioinformatics Institute, CEUR Workshop Proceedings (vol. 148)
14 Yakushiji, A. *et al.* (2005) Biomedical information extraction with predicate–argument structure patterns. *First International Symposium on Semantic Mining in Biomedicine (SMBM)* 10th–13th April, European Bioinformatics Institute, CEUR Workshop Proceedings (vol. 148)
15 Pustejovsky, J. *et al.* (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. In *PSB 2002 – Proceedings of the Pacific Symposium on Biocomputing 2002* (Altman, R.B. et al. eds), pp. 362–373, World Scientific Publishing
16 Narayanaswamy, M. *et al.* (2003) A biological named entity recognizer. In *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003* (Altman, R.B. et al. eds), pp. 427–438, World Scientific Publishing
17 Wermter, J., and Hahn, U. (2004) Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. *Proceedings of the 11th World Congress on Medical Informatics*. In *Studies in Health Technology and Informatics* (Vol. 1, number 107) (Marius, F. *et al.*, eds), pp. 560–564, IOS Press
18 Wermter, J. *et al.* (2005) Recognizing noun phrases in biomedical text: an evaluation of lab prototypes and commercial chunkers. *First International Symposium on Semantic Mining in Biomedicine (SMBM) – 10th–13th April* European Bioinformatics Institute, CEUR Workshop Proceedings (vol. 148)
19 Proux, D. *et al.* (1998) Detecting gene symbols and names biological texts: a first step toward pertinent information extraction. *Genome Informatics Workshop* pp. 72–80
20 Fukada, K. *et al.* (1998) Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.* 701–713
21 Tamames, J. (2005) Text detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinform.* 6 (Suppl. 1), S10
22 McDonald, R. and Pereira, F. (2005) Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinform.* 6 (Suppl. 1), S6
23 Zhou, G.D. *et al.* (2005) Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinform.* 6 (Suppl. 1), S7
24 Fundel, K. *et al.* (2005) A simple approach for protein name identification: prospects and limits. *BMC Bioinform.* 6 (Suppl. 1), S15
25 Crim, J. *et al.* (2005) Automatically annotating documents with normalized gene lists. *BMC Bioinform.* 6 (Suppl. 1), S13
26 Kim, J-D. *et al.* (2004) Introduction to the bio-entity task at JNLPBA. *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 28–29 August, Geneva, Switzerland
27 Yeh, A. *et al.* (2005) BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinform.* 6 (Suppl. 1), S2

28 Hirschman, L. *et al.* (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinform.* 6 (Suppl. 1), S11

29 International Union of Pure and Applied Chemistry, Organic Chemistry Division, Commission on Nomenclature of Organic Chemistry (III.1) (1993) *A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993)*, Blackwell Scientific Publications

30 Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36

31 Negwer, M. and Scharnow, H-G. (2001) *Organic-Chemical Drugs and their Synonyms*. Wiley-VCH

32 Singh, S.B. *et al.* (2003) Text Influenced Molecular Indexing, TIMI: a literature database mining approach that handles text and chemistry. *JCICS* 43, 743–752

33 Hanisch, D. *et al.* (2005) ProMiner: rule-based protein and gene entity recognition. *BMC Bioinform.* 6 (Suppl. 1), 14

34 Hanisch, D. *et al.* (2003) Playing biology's name game: identifying protein names in scientific text. *Pac. Symp. Biocomput.* 403–414

35 Zimmermann, M. *et al.* (2005) Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology. *Curr. Top. Med. Chem.* 5, 785–796

36 Li, Z. *et al.* (2004) Personal experience with four kinds of chemical structure drawing software: review on ChemDraw, ChemWindow, ISIS/Draw, and ChemSketch. *J. Chem. Inf. Comput. Sci. (Software Rev.)* 44, 1886–1890

37 Jain, A.K. *et al.* (2000) Statistical pattern recognition: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 4–37

38 Bunke, H. (1995) *Handbook of Pattern Recognition and Computer Vision*, pp. 163–209, World Scientific Publishing (chapter 1.5)

39 Brown, J.R. and Balmuth, J.R. (1992) Kekule: OCR – optical chemical (structure) recognition. *J. Chem. Inf. Comput. Sci.* 32, 373–378

40 Ibison, P. *et al.* (1993) Chemical literature data extraction : the CLiDE project. *J. Chem. Inf. Comput. Sci.* 33, 338–344

41 Zimmermann, M. *et al.* (2005) Combating illiteracy in chemistry: towards computer-based chemical structure reconstruction. *ERCIM News* 60, 40–41

42 Barsalou, T. (2004) Text mining for competitive intelligence. *Research & Development SCIP's European Conference 2004*, 27–29 October, Milan, Italy

43 Vachon, T. (2003) Text mining in life sciences informatics at Novartis. *Symposium on Text Mining in the Life Sciences*, 1 October, Sankt Augustin, Germany