Information Extraction in the Life Sciences: Perspectives for Medicinal Chemistry, Pharmacology and Toxicology

Marc Zimmermann, Juliane Fluck, Le Thuy Bui Thi, Corinna Kolárik, Kai Kumpf, and Martin Hofmann*

Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53754 St. Augustin, Germany

Abstract: Information extraction approaches have been successfully applied to mine the scientific literature in biology and medicine. So far, the main focus of research and development in this domain was on the recognition and extraction of gene and protein names in the context of molecular biology and genome research and on disease names and other medical terms in the context of clinical research. Similar to biology and medical sciences, medicinal chemistry, pharmacology and toxicology are descriptive sciences. However, information extraction approaches in these disciplines encounter a number of problems that are specific to the fact that these scientific areas are essentially centred at chemical compounds and their structures. In this review, we will give a short overview on general information extraction strategies in the life sciences and we will introduce new approaches to apply information extraction to the domain of pharmacology, medicinal chemistry and toxicology. Finally, we will emphasize on how information extraction approaches will support public and commercial research in medicinal chemistry, pharmacology and toxicology by linking information on chemical structures to biological information.

Keywords: Information extraction, named entity recognition, chemical structure reconstruction, biological effects, semantic mediation.

INTRODUCTION

Life sciences have been predominantly descriptive sciences for centuries. Only recently, in the second half of the 20th century, the advancement of analytical biochemistry and molecular biology made it possible to unravel some of the molecular mechanisms underlying biological processes. In general, these biological processes tend to be rather complex and their formal representation is not trivial. Due to the lack of a unified, comprehensive model for molecular biology and molecular pharmacology, the description of scientific results in unstructured text is still the method of choice to communicate information in these knowledge domains. In order to make the vast amount of published information in molecular biology and clinical research available in a computer-readable format, several groups have started to work on strategies for information extraction and information reconstruction from scientific publications over the last few years. Two basic modes of information extraction can be distinguished:

- information extraction from textual sources
- information extraction from images

In the following we will review recent approaches in information extraction in the life sciences and we will sketch, how this approach can be extended beyond the current scope on biological and medical term recognition towards an application of this technology in medicinal chemistry, pharmacology and toxicology.

INFORMATION EXTRACTION: STATE OF THE ART IN BIOMEDICINE

Scientific publications comprise all sorts of written scientific communication, therefore this term covers publications not limited to scientific journals but also other types of publications, e.g. patents or reports. Information extraction approaches in the life sciences, however, have focused largely on molecular biology and genomics information so far. Gene and protein name recognition and information extraction approaches have widely used abstracts from MEDLINE, the largest public bibliographic data collection comprising more than 15 million abstracts of biomedical journal publications [1] as a source. Recent reviews on text mining in the biomedical domain give insight into different strategies for information extraction in these fields and underline the importance of this technology for biomedical data analysis and knowledge management (e.g. [2-3]).

In contrast to information retrieval, which deals with the finding of documents according to a query and organizing them by relevance or topic, information extraction (IE) is centred at the identification of explicit entities and extraction of information related to these entities. By combining different natural language processing (NLP) tools, lexical resources and semantic constrains, information extraction provides effective modules for mining the literature. Here, we confine ourselves to giving a short overview about the state of the art of these techniques. We will also emphasize

*Address correspondence to this author at the Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53754 St. Augustin, Germany; Tel +49-2241-14-2802; Fax +49-2241-14-2656; E-mail: martin.hofmann@scai.fhg.de

1568-0266/05 $50.00+.00

on problems and solutions that accompany the adaptation of information extraction techniques to the biomedical domain.

ARCHITECTURE OF INFORMATION EXTRACTION SYSTEMS

A comprehensive information extraction system possesses the following major components [2]: tokenization, part of speech tagging, noun phrase chunking, named entity recognition and syntactic analysis (cf. Table **1**).

Most of the information extraction tools in the biomedical domain make direct use of algorithms for tokenization, part of speech tagging and chunking that were developed for the general purpose newspaper domain (e.g. [4-6]). Therefore, results on part of speech tagging [7] and noun phrase chunking [8] show a significant loss of performance when these methods are applied to the biomedical domain. Special adaptations like the integration of biological lexica are needed to cope with this loss. Also, different tokenization of texts, in particular in complex noun phrases containing hyphens, parenthesis and other special tokens can result in a performance loss of methods based on tokenization.

NAMED ENTITY RECOGNITION OF PROTEIN AND GENE NAMES

The unambiguous identification of biological entities or processes is a fundamental requirement for information retrieval and extraction in the biomedical domain. Some of the difficulties encountered in name recognition include the handling of yet unknown words (e.g. new gene names), multiple names for one entity (*synonymy*), and the identification of terms composed of more than one word (*multi-word terms*). Furthermore, identical names are used to identify different proteins, genes or other biological entities (*polysemy*). During the last years a number of different methods for named entity recognition in the biomedical domain have been developed and independent assessments for the evaluation of these methods have been set up. Current systems for named entity recognition 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); the second class deals with the recognition and mapping of names to unique database entities (e.g. mapping of mouse gene and protein names to the MGI database entities). In the first class machine learning based approaches are predominant; the performance of the best systems show precision and recall rates of about 70 percent in the assessment of the GENIA corpus (cf. [9] for an overview, the best system has 76% recall and 69% precision) and around 83 percent in the BioCrEAtIvE assessment (cf. [10] for an overview). These rates are far below the rates, which could be achieved in the general newspaper domain where balanced precision and recall between 93–95% can be reached for the identification of person, organization and location names. The reasons for this difference are most likely the higher degree of ambiguity of domain names in biomedical text corpora and annotator disagreements in the provided corpora. Moreover, different tokenization of text may be a significant source for error.

The BioCrEAtIvE assessment also provided training and test corpora together with name dictionaries for three different organisms (mouse, fly and yeast) for the second class of named entity recognition. The gene and protein name should be recognized and mapped to the according database entity. Here, most of the participating systems are rule based but two systems use a combination of a rule based matching and a machine learning based filter for false positive hits, which lead to a good overall performance (cf. [11] for an overview). For yeast, most approaches reach more than 90 percent recall and precision due to the very stringent nomenclature that is in use for this organism. The ProMiner system, which is rule based, reached the best results of all participating systems in the fly (*Drosophila melanogaster*) and mouse corpus [12]. For drosophila, despite the high amount of ambiguous and common word gene names, precision and recall values of about 80 percent were reached. For the mouse dictionary, only slightly better results than in fly could be achieved. Reasons for this are 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). Furthermore, annotator disagreement was higher for this organism, which leads also to erroneous annotations for the training and test corpus.

In summary, for both classes of named entity recognition similar performance rates where reached in the case of the BioCrEAtIvE assessment. For the class 1 recognition, clear boundary definitions in the corpus annotation could improve performance, but due to the ambiguity of naming in the

Table 1: Major Components of Information Extraction Systems

Tokenization	a process of breaking the text up into its constituent units which are words (= token) sentences or also parts of texts.
Part of speech tagging	labels the word with different categories based on the role the word play in the sentence (e.g. verb, noun , preposition, adjective).
Noun phrase chunking	focuses on the identification of basic structural relations between groups of words.
Named entity recognition	identifies complex noun phrases as entities in the simplest case, classifies them in different categories (e.g. protein or cell-type) or maps them directly to very granular instances (e.g. to a certain gene entity defined through a gene database identifier).
Syntactic analysis	establishes the connection between different parts of each sentence. This is done in the simplest case through co-occurrence and statistical analysis or with different syntactic parsing methods.

biomedical terminology it will be difficult to reach the same performance as in other fields, e.g. the general newspaper domain. Straightforward solutions such as using class 1 results (recognition of gene and protein names) as an input for the class 2 task (mapping of these names to the database entries) would probably result in a performance loss. The mapping, however, to the corresponding entities is an indispensable step in extracting relevant information coming from multiple synonyms in the text and linking extracted information to experimental data.

SOURCES OF TERMINOLOGY FOR THE CONSTRUCTION OF DICTIONARIES

An important aspect for term recognition and in particular for the mapping of terms to certain information classes such as database entries or ontology categories) is the availability of terminological resources for the generation of dictionaries. The major sources for gene and protein related terminologies are genome and protein databases like the sequence database ENTREZ GENE [13] and the protein database SWISSPROT [14]. The gene ontology (GO) [15] provides the terminology for the classification of genes and proteins in functional and process categories. In the text mining system *Texpresso* [16] gene ontology was used to search corresponding terms in the *Caenorhabditis elegans* literature. Most gene and protein databases contain GO annotations and therefore mapping of the extracted gene and protein names to the gene name dictionary also provides a mapping to GO categories. The similarity of GO annotation of genes and their co-occurrence in MEDLINE abstracts was used by Jennssen *et al.* for the interpretation of gene expression data [17]. The Medical Subject Headings (MeSH) [18] and the Unified Medical Language System (UMLS) metathesaurus [19] provide large terminological sources for medical terminology. But these terminologies are very complex, often not directly useful for name recognition and no mapping is directly available between the different UMLS resources. Because of these reasons the applicability of this terminology resource is limited.

APPLICATIONS FOR INFORMATION EXTRACTION SYSTEMS

For information extraction the applications typically concentrate on finding information about genes and proteins and relationships between these entities. The strategies vary from using simple co-occurrence of entities in abstracts or in sentences to dictionary approaches combined with statistical approaches. This leads to high recall rates but also to the extraction of a high amount of false positive relations. Further work concentrates on training classifiers for the recognition of sentences containing the desired information. Craven and Kumlien [20] used classifiers for the extraction of sentences discussing the location of proteins within the cell and Ray and Craven [21] used hidden markow models to identify sentences describing the association of genes and diseases. The latter approach reaches a performance of 77% precision at 30% recall or 92% precision combined with a recall rate of 21%. In comparison to this classifier, simple co-occurrence led to a higher recall rate (70%) with a considerable lower precision (40%).

Numerous information extraction (IE) systems for the recognition of protein-protein or protein-gene relations with more NLP based methods of tagging and parsing have been developed. Rule based systems (e.g. [5, 22]) as well as machine learning based systems (e.g. [23-24]) were developed and reach high precision rates (above 80%) accompanied with comparably low recall rates (significantly below 50%). Text mining assessments for IE tackle problems of extracting information about expression experiments for Drosophila genes (KDD cup 2002 [25]) or experimental evidence for the GO annotation of mouse genes (BioCreAtIvE, 2003 [26]).

In summary, all information extraction systems have been evaluated so far on very small benchmark sets and it remains to be demonstrated, how these systems perform on larger corpora. The user can choose between co-occurrence systems, allowing high recall but low precision rates, or NLP based applications leading to higher precision but substantial loss in recall. The hope here is that important interactions are typically described more than once and the recall is higher in respect of finding existing relationships. Our own experience using co-occurrence as well as grammar based systems for the generation of protein and gene interaction networks, which in turn were used for the interpretation of gene expression data, clearly underlines the need for highly precise systems [27].

We are, however, aware that depending on specific applications a user defined tuning of recall and precision levels would be desirable.

SPECIFIC CHALLENGES IN CHEMICAL ENTITY RECOGNITION

Named entity recognition is a slightly different problem in the field of compound names and chemical nomenclature. In publications dealing with small molecules and their pharmacological and toxicological effects we distinguish two major classes of compound names:

- trivial names (including brand and trade names or database identifiers and registry numbers)

- regular expressions (based on nomenclatures, e.g. IUPAC [28], or linear notations, e.g. molecular formula, SMILES [29])

Disambiguation of multi-word terms and semantically correct mapping of synonyms to reference terms pose significant challenges to both, the field of biological as well as the world of chemical entity recognition. Despite the fact that chemistry follows physical rules, whereas rule-based nomenclatures for the description of small molecules have been already developed as early as 80 years ago [30], the representation of complex molecules by IUPAC nomenclature can result in complicated regular expressions that are not unambiguous. Moreover, in many publications trivial names and abbreviations are used, in particular when the scientific report focuses on the biological effects of drugs.

The preferred representation of chemical entities is, however, the chemical structure of the molecule. Molecules are most often represented as atoms and the way, how these atoms are connected (i.e. bonds). This is some kind of universal language of chemistry, which can be understood by the human expert and by the computer. Some of the representations are more easily interpretable by humans like depictions or drawings and some are better suited for the machine (e.g. connection tables [31]). Based on the chemical structure several of the above mentioned problems for word terms can be solved. The more technical chemical structure representations (e.g. SMILES; SD files) can be used as input formats for most computational chemistry programs such as tools for similarity searches and exact matches (i.e. allowing disambiguation). Hence, the task of transforming knowledge from scientific publications into a computer-accessible form can be separated into two sub-tasks:

1. Linking an entity name contained in free text to its chemical structure,
2. Extracting chemical structure information from images

Both steps will be described in more detail in the following section.

FROM ENTITIES TO STRUCTURE

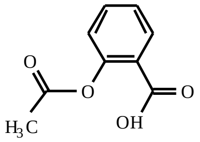
Chemical entities can appear in scientific text as trivial and brand names, assigned catalogue names or structure description formats (cf. Table **2**).

Table 2. Textual Representations of Chemical Compounds

Representation	Examples
Trivial and brand names	Aspirin (there are 163 synonyms listed in PubChem [32] and even more can be found in Negwer [33])
Assigned catalogue names	EC numbers [34], CAS registry numbers [35], Beilstein numbers [36], patent identifier, NCI identifier [37], and vendor identifiers [38]
Structure description formats	Molecular formula, SMILES, IUPAC, IUPAC International Chemical Identifier (InChI [39])

The complexity of the name space, that exists for well known small molecules, is illustrated in the following example of 2-(Acetyloxy)benzoic acid, a compound better known as Aspirin. We count more than 163 synonyms in the literature and in compound databases for Aspirin; and even following the rules of IUPAC nomenclature we can assign three different names to the compound with the molecular formula C9H8O4. CAS registry assigns a single code to Aspirin, however, different salts of the molecule are represented by different codes (cf. Fig. **1**).

Depiction:



Synonyms: A55, A.A, AAS, Aasito, AC 5230, Acard ... Zorprin

IUPAC: 2-(Acetyloxy)benzoic acid, O-Acetylsalicylic acid, 2-Acetoxyben-zoic acid

Molecular Formula: C9H8O4

SMILES: CC(=O)OCl=CC=CC=C1C(=O)O (others can be created)

InChl: 1.0RC/C9H804/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H, 1H3,(H, 11, 12)

CASRN: 50-78-2 (different salts: 60364-30-9 ...)

Fig. (1). These are different examples of representations of the drug known as Aspirin.

In order to extract relevant information for chemical entities, all different kinds of representations have to be marked, extracted and linked to one unique object (i.e. the chemical structure). In the chemical domain we are still lacking a common compound index, where every structure can be uniquely mapped to its different representations. This problem has been – at least in part – addressed in biology for genes [40]. The number of small molecules which can be synthesized at least in theory is so vast [41] that we cannot expect to create a comprehensive index. On the other hand, a lot of information on interesting molecules like marketed drugs has been independently collected. For most trivial names, the link to the corresponding chemical structure is actually available in scientific text sources. There are public efforts to create collections of brand names of drugs and to establish a common unambiguous name (examples are INN [42], USAN [43]). Catalogue names have been introduced as a way to formalize the nomenclature in the chemical domain. However, they still lack the information required to link to the chemical structure. In order to access the information associated with a catalogue name or identifier, we have to know the data source (naming organization) and then we have to be able to query this catalogue for the structure.

As a consequence, to establish this sort of link, databases containing all trivial names and catalogue names are required. The quality of information extraction is thus dependent on the completeness of such a database. Typical problems encountered are multiple synonyms being used for a single entity, and the completeness due to new entities being discovered at a rapid rate. Due to the large effort in keeping these catalogues up-to-date most of them are commercial and very expensive for the academic community. Some published catalogues are: Negwer, Merck Index, CAS Registry, and the Beilstein Crossfire database.

In spite of creating a single large compound index by copying all the information from public, commercial and proprietary sources a virtual index could be created which only stores the links between them. The main advantage would be that changes and extensions of the original source are automatically included in the virtual index. We have created a prototype, named TUAM, for the purpose of mediating between different data sources which can be unstructured text, ontologies, tabular data or even relational databases (cf. Fig. **2**). TUAM stands for “Tool for Universal Annotation and Mediation”. It allows the linking of single entities (words, phrases, molecules, ontology classes, database columns or rows) by relations. Relation types can express general dependencies like “similarity” or more specific ontological relations like inheritance (“is-a”) or containment (“part-of”). Two entities are mapped onto each other in RDF [44] style (triples in the form: subject – predicate – object, where subject and object are the individual entities and the predicate is a relation type). The

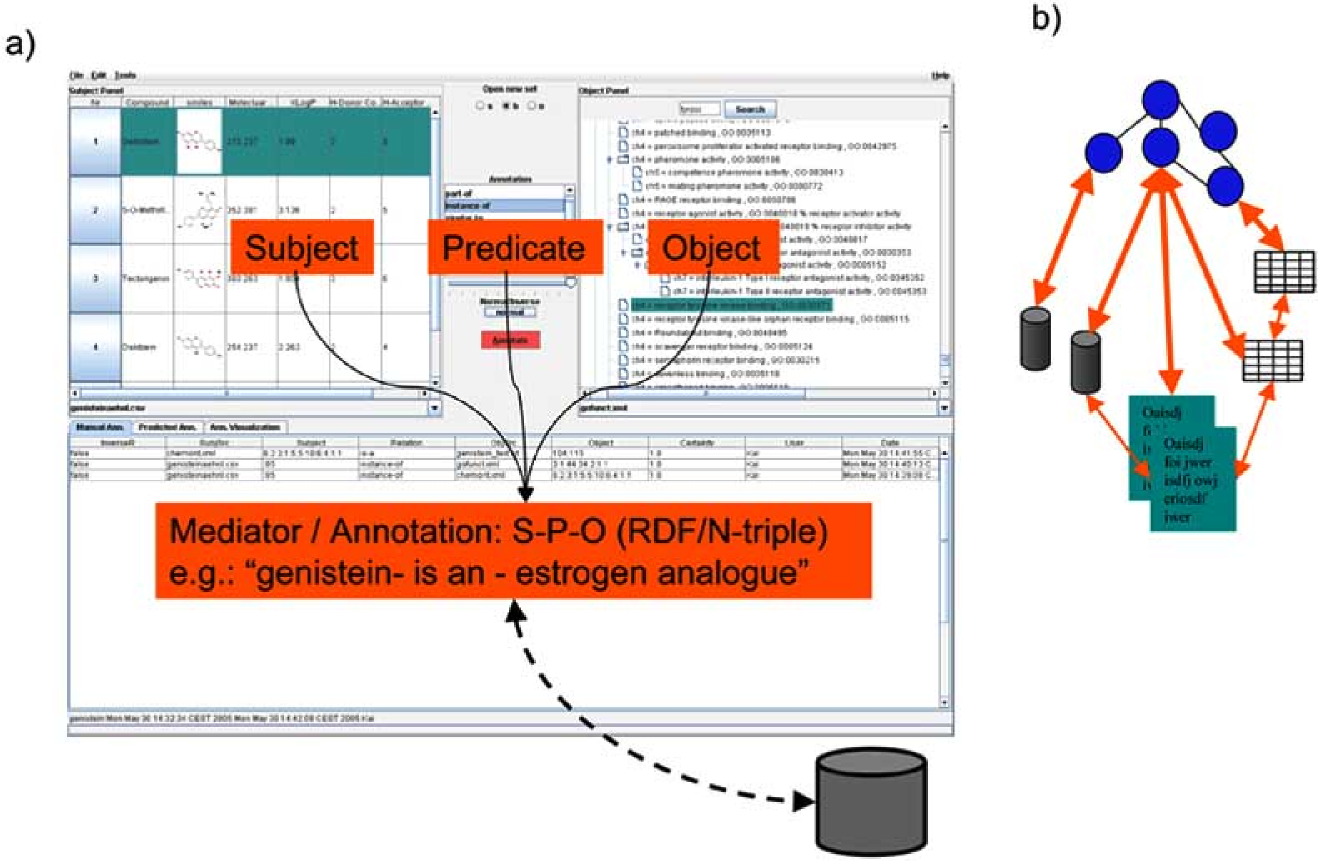

a)
b)
Subject
Predicate
Object
Mediator / Annotation: S-P-O (RDF/N-triple)
e.g.: "genistein- is an - estrogen analogue"
Oaisdj
Oaisdj
foi jwer
isdfj owj
eriosdf
jwer

Fig. (2). a) TUAM provides semantic mediation for workgroups dealing with multiple data sources, e.g. chemical databases (left application panel) and biological ontologies (right panel). Annotations are stored as subject-predicate-object-triples and made persistent in a relational database.

b) Semantic mediators can be n:m (many-to-many) between arbitrary data sources. Usually, an ontology would be chosen as the semantic "hub" (graph structure at top, thick arrows). Other mediators can be assembled to refine the semantic network (thin arrows).

data sources are left untouched by the annotations. The annotations can be exploited as semantic annotations, i.e. the addition of new attributes to existing data. Cardinality of annotations is usually n:m (many to many) so that dense semantic nets can be produced. Ontologies are usually chosen as the semantic hub for other kinds of data in order to have a semantic foundation, which is both, well controlled and itself semantically rich. The semantically enriched target data, e.g. chemical compound data sets, can be subsequently exported or directly accessed from the annotation database to make it amenable to data mining.

In an attempt to capture the extremely large number of potential variants in the domain of small molecules without having to invent new names, descriptive formats such as IUPAC names and SMILES strings have been developed. These are normally string representations derived from the chemical structure. Because these systems are rule-based, an identifier can be automatically created for every possible molecule. This has several advantages. For assigning the molecules an identifier, they do not have to be enumerated and ordered. It is possible to reconstruct the structure from the name by reversal of the naming process. Several commercial and open source tools exist to convert from IUPAC or SMILES names to structure and vice versa. These are often associated with structure drawing tools. A few examples are given in [45-49].

RECOGNITION OF CHEMICAL ENTITIES THROUGH NAMED ENTITY RECOGNITION

Only a very limited number of name entity recognition approaches tackle the recognition of chemical entities. Narayanaswamy *et al*. [5] presented a rule-based approach, which exploits surface clues, simple linguistics, and domain knowledge in identifying the relevant terms. The authors define token features, which support the classification of protein or chemical name classes and present data with impressive high precision (above 90 percent) and also relatively high recall rates (up to 73 percent) using their method. However, it could be that they over-fitted their system to the very small training and benchmark set of 55 MEDLINE abstracts. Other systems use dictionaries and simple string matching without any evaluation of recall and precision (e.g. [50]).

For a more detailed discussion of the problems and possible solutions for the recognition of compound names and the mapping to their structures we would like to deal with the two classes of trivial names and regular names separately. The mapping of IUPAC and other regular names to their respective structure can be done by already available software as discussed above. Therefore, the recognition of this type of names in text is sufficient to link textual sources to chemical structures.

For the recognition of trivial names we used the ProMiner system which, was tested on benchmark sets for protein and gene entity recognition in different organisms (human, mouse, fly, and yeast) and performed very well in the BioCreAtIvE assessment. The ProMiner system [51] consists of two parts: The first part is engaged with the dictionary generation. The gene and protein dictionary is assembled from entries of different biological databases. As the name and synonym fields in these databases often contain physical descriptions (cDNA clone, RNA, 5'-end), family names (membrane protein) or other annotation remarks, the dictionary is cleaned in an automated process. This process can also be adapted to the purpose of generating chemical dictionaries. Based on a dictionary, a string match procedure is used for name detection in the text. The system was initially designed for human protein and gene names, which often consist of multiple words. In that setting, name variants, e.g. permutations, insertions or deletions of words were observed. For example, the name ”Interleukin type 1 beta” is a spelling variant of ”Interleukin-1 beta”. However, ”Interleukin 1” is a different protein entity than ”Interleukin-1 receptor”. The ProMiner match procedure assigns weights to different classes of words to reflect their importance for name detection. These are termed token classes. Defining different token classes allows for fast adaptation of ProMiner to the chemical name recognition problem.

For the generation of chemical dictionaries we can obviously not make use of any thesaurus of chemical entities available to the public. In principle, it should be possible, to create dictionaries from given chemical structures (even new ones) automatically. A potential disadvantage of this approach is, that all the chemical naming conventions are not as regular as they seem. Depending on the conversion tool used, we will generate several different names for the same compound. Moreover, the possible chemical space of small, drug-like molecules is incredibly huge and therefore the dictionaries will become cumbersome large. That is why we think the name recognition through generalized chemical ‘scaffold’ patterns should be possible for IUPAC nomenclature and SMILES representations and we generate dictionaries for these names only for evaluation purposes.

As a good starting point for dictionary generation names were extracted out of the chemical part of MeSH [18] and ChEBI [52]. Trivial names were merged with the help of the CAS registry number and incorporated in a base thesaurus, which was used as a dictionary for the ProMiner approach. In such a way we have collected a list of drug names, which contains a total of 4673 drug identifiers with 11366 synonyms. 9125 of these synonyms are different, 1294 belong to more than one drug definition. Additionally, some of the synonyms are ordinary English words.

A dictionary of IUPAC names, generated from the same resources, serves as an evaluation base for general recognition of these terms. Curation of this dictionary, to resolve name conflicts and ambiguities, was done using the curation routine in ProMiner. Additionally, a curation step is indispensable for the generated IUPAC name dictionary. To correct the most obvious mistakes, different checks (braces have to match, no textual description, no trailing dash or an ending ‘ic’) have been done. The resulting dictionary contains 128,736 individual strings. Nevertheless, a number of mistakes can still be identified in the training corpus (e.g. spelling mistakes such as “m ehtyl”), so that perfect accuracy is impossible to achieve even with a perfect algorithm.

The next problem is a lack of test collections (“test corpora”) of text sources containing chemical entities. The GENIA corpus provides an annotation for chemical compounds but the provided corpus contains mostly ion names (e.g. Ca^{+} or Fe^{++}) and only few IUPAC and trade names. For this reason we assembled an artificial test corpus using the GENIA corpus. Chemical entities tagged in this corpus are removed and replaced by compounds randomly picked from the generated chemical compound dictionaries. The recall and precision can be computed after applying the system to the modified corpus.

Using the trivial name dictionary and the artificial GENIA corpus the ProMiner system returned 11321 hits, 9026 correct terms have been identified and 99 were missing. This leads to about 80 percent precision and a recall rate of almost 99 percent. The recall rate is definitely due to the artificial test corpus which contain only the exact dictionary names but the precision rates are very similar to the rates achieved with the gene and protein name dictionaries. Here, future work in adapting ProMiner parameters and filter functions to the chemical domain is likely to generate higher precision rates.

For the identification of IUPAC names our results are of very preliminary nature. For tackling the problems of name identification we only used the dictionary for evaluating the recall of our artificial system. During the evaluation, it has become clear, that the handling of braces is somewhat critical for an accurate system. IUPAC names contain dashes, numbers and words within each token. To identify these, each token has been matched against a JavaCC grammar used for generating a parse tree of such names. The grammar can be found in (Table **3**).

Table 3. JavaCC Grammar for the Generation of IUPAC Parse Trees

Rules for parsing IUPAC tokens:
- Braces delimit subparts of the expression and the content of each brace has to match the generation of a normal IUPAC token or a word
- A IUPAC name is a sequence of words and special IUPAC fragments, which are linked by dashes
- A special IUPAC fragment can consist of smaller atomic parts delimited by comas (“ ,”) , or colon (“: ”)
- The atomic part can consist of a single letter or an integer, optionally followed by multiple hyphens, optionally followed by [abc], a Greek letter, an R or S, and a “*”. Such a chain is also allowed to start with a Greek letter.

Once a token has been assigned to the “IUPAC” class, this match has been expanded to neighboring tokens, if these tokens correspond to words commonly used as part of chemical names.

For testing purposes, the grammar used for parsing the IUPAC tokens has been applied to the training dataset. Out of these 128,736 names, 109789 or 85.2% were successfully parsed by the grammar. Shortcomings for the recall rate in the grammar are the handling of braces. The current system does not check matching of braces across multiple tokens. For the search on the chemical names in text the systems we have identified the following shortcomings which lead to unsatisfying precision:

Ambiguous usage of commas and other punctuation (e.g. used as enumeration or sentence separator, but also at the end of different parts of the same chemical entity name).

Definition of borders of chemical entities (what should be part of the chemical name, e.g. which leading adjectives should be included).

These problems are leading to different tokenization requirements inside and outside of IUPAC names. There is a drastic difference in the usage of brackets, dash, commas and other punctuation. Therefore integration of chemical name entity recognition in information extraction will need two different tokenization approaches.

In summary the established gene and protein recognition approaches, like the ProMiner system, can easily be adapted for the detection of trivial chemical names, but for the recognition of regular names further research and the generation of an appropriate training corpus is indispensable. Our work will be continued to combine dictionary based methods with the recognition of regular terms into a single system.

FROM STRUCTURE TO UNIQUE IDENTIFIERS

After the molecular structure of an extracted entity is retrieved, there are several ways to generate unique compound identifiers from it. The algorithms are based on the translation of the molecular graph in a simple string (canonical SMILES or InChI) or a number by using graph traversals [53]. The molecular graph is the formal description of the chemical structure. Simply speaking a graph is a collection of objects (called nodes) and a collection of relationships between these objects (called edges). All the information of a single atom is stored into an attributed node. The attributes are the atom properties like the elemental type. The bonds are assigned to the edges. A graph traversal is an ordering of the nodes and edges. Starting from one node the neighboring nodes (these are the nodes which are directly linked by an edge) are visited being the next starting point. There are several rule systems (algorithms) to do this in a unique way and therefore providing a unique identifier for each graph: being the ordered node-edge list. Most of these algorithms have been developed for database and compound registration systems where unique keys are needed in order to store information. All chemical software providers offer a solution [54-61].

INFORMATION EXTRACTION FROM IMAGES

One universal representation of molecules is the two-dimensional depiction of the molecular structure, which is based on the molecular graph. Depictions can be found as images in nearly all electronic sources of chemical information (e.g. journals, patents, and web interfaces of chemical databases). Nowadays theses images are generated with special drawing programs, 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 bitmap image (e.g. GIF for web interfaces or BMP for text documents). As a consequence the structure information can no longer be used as input to chemical analysis software packages. Normally the extracted images have to be manually converted by redrawing every structure. This is a time-consuming and error-prone process. There are initiatives to exchange more informative formats like Chemical Markup Language CML [62] (for webpages) or SDF (supporting material in journals) and to use plug-ins or special renderers for the viewing process.

To make chemical information contained in drawings of chemical structures accessible for computer programs, projects have been started to convert bitmap images of molecules back into machine-readable form. This process is called chemical structure reconstruction (CSR). In general there are two different strategies (not restricted to the chemical domain) for the recognition of patterns in images: statistical pattern recognition [63] and structural pattern recognition [64], (cf. Fig. **3**).

Both strategies can be combined with each other. For CSR, two projects have been described in the literature: the Kekulé and the CLiDE project:

Kekulé [65]: 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 like chiral bonds and text representations. For the atoms and bonds a connection table is computed. The last step comprises a post-processing phase which normalizes the reconstructed graph (e.g. bond lengths).

CLiDE [66]: The CLiDE projects uses segmentation algorithms for monochrome black-white bitmaps in order 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).

EVALUATION OF STRUCTURE RECOGNITION TOOLS

As is true for the textual information extraction in chemical sources, we are missing a test corpus consisting of images and the associated structural information. In order to evaluate CLiDE (no software is available from the Kekulé project) we have been forced to assemble our own test corpus. This test set should cover a diverse set of molecules and depictions. It should consist of relevant compounds and provide the information on the structure in a chemical file format. We therefore picked the top 100 “blockbusters” (most sold drugs) from the year 2002 [67] to assemble our test corpus. For each molecule an image was generated using ChemDraw [68] (cf. Fig. **4**). ChemDraw is a typical drawing tool providing templates for frequent scaffolds (e.g. typical

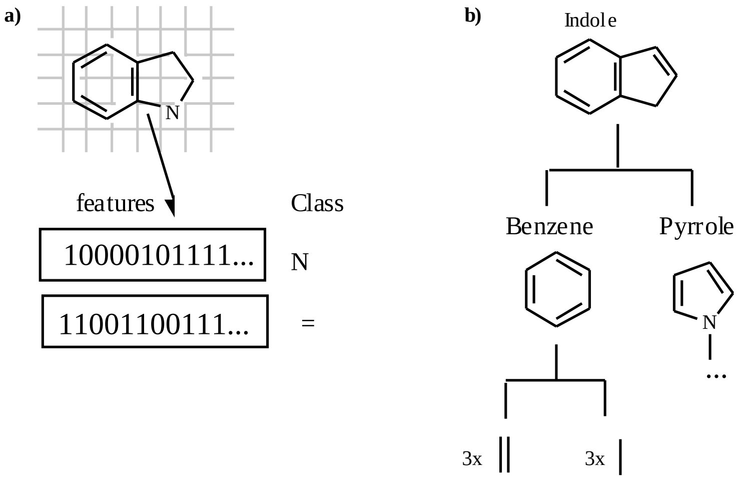


Fig. (3). a) *Statistical pattern recognition*: This method derives *n* so called features from the image representation (e.g. colour frequencies, pixel patterns). In the next step the image is decomposed into smaller parts. For each segment an *n*-dimensional feature vector is computed. Supervised learning algorithms can be trained on the feature vectors in order to recognize patterns in test sets of images.
b) *Structural pattern recognition*: This strategy is using abstract data types like strings, trees, and graphs instead of numerical types. These concepts can describe relationships between objects (e.g. geometric, spatial). It allows the hierarchical composition of substructures. In the first step the image is converted into a collection of objects (e.g. by vectorization of the bitmap). Then the relationships between the objects are computed. Unknown objects are compared to model objects using matching algorithms (e.g. string alignment, graph matching).

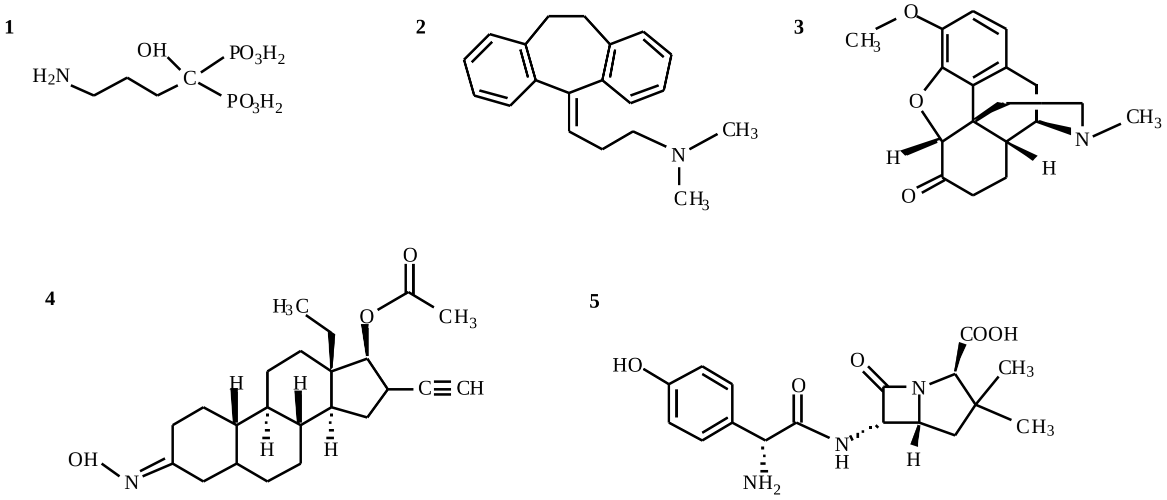


Fig. (4). Example molecules from 100 blockbuster set: 1 Alendronate, 2 Amitriptyline, 3 Hydrocodone, 4 Norgestimate, 5 Amoxizillin.

ringsystems like steroids). It allows exporting the structure as an image or in a chemical file format. The image can be manipulated in several ways: scaling, rotating, line thickness, colouring, different font types for atom symbols. Depictions of the same molecules have been extracted from the graphical user interface of the Merck Index. There are some limitations on the input for CLiDE. Namely the images have to be monochrome (no grey levels are allowed), the format should be BMP or PDF and the resolution should be 300 dpi. Although the software has been developed more than 10 years ago, the problem of structure recognition is still not solved. The results from the validation study are not encouraging. In about 45% of the images there was an error [69]. The most common errors have been made in the area of identification of chiral and double bonds and a wrong assignment of the letter ‘N’ from the Merck Index pictures or ‘Cl’ for the ChemDraw generated images. The letter ‘I’ is sometimes converted to a single bond (cf. Fig. **5**). In some cases CLiDE asks the user for manual intervention, but the system does not learn from expert intervention. This means: the same type of error will be reproduced by using even the same input. The system cannot be trained to perform better after correction through human experts.

REDESIGN AS A MODERN STRUCTURE RECOGNITION TOOL

Our idea was to assemble a prototype for structure reconstruction reusing standard technology which has been developed in fields outside of the chemoinformatics domain.

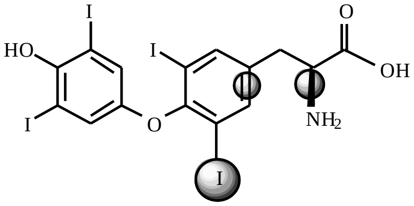


Fig. (5). The most common error sources for the structure recognition process are chiral and double bonds, misclassification of text, and bridged rings (grey circles). The picture of Thyroxine is taken from the Merck Index.

A workflow has been assembled which uses image readers and vectorization software in combination with classic OCR, machine learning, graph matching, and a chemical structure editor. The prototype consists of three phases: image pre-processing, graph matching, and molecule reconstruction (cf. Fig. **6**). The segmentation of the raster image is the first step. First of all pixels are assigned to the background, to the boundary of a graphical element or to the interior of an element. Then the image is broken down into connected components. Each component is classified by machine learning as text, a bond or a special graphical symbol (chiral bonds). Some of the components have to be grouped together to form one object. The text is separated from the image and passed on to an OCR program. The remaining parts of the raster image are converted into scalable vector graphics (SVG [70]). The resulting lines are searched for single, double and triple bonds. In the next step the connections of the bonds are analysed and minor errors like disconnections, lines resulting from noise or dented lines are fixed based on rules looking at the local neighbourhood and checking bond lengths. The bonds form the skeleton of the molecular graph. The molecular graph is invariant to scaling, translations, and rotations. We can therefore use a template database of common scaffolds in drug like molecules and compare them to the extracted graph from the image. The task is to find all subgraphisomorphisms between the template and the query graph. All templates have been assembled by extracting the most common fragments from the input molecules. The templates are efficiently stored in a precomputed decomposition network [71] and the sub-graphisomorphisms are constructed by a bottom-up algorithm. The complete molecule is reconstructed using the list of the found non-overlapping templates. New templates can be automatically identified and added to the library. In the last step the information on the atom types from the OCR process is merged with the molecular skeleton. A library of common superatoms like ‘COOH’ is used to construct the complete chemical structure. In the final phase chemical rules are used to check for bond order or bridged ringsystems. Afterwards the result is presented in a chemical editor and can be manually corrected or exported to a chemical file format. The resulting molecule is annotated with the number of corrections made by each filtering step and a score for the matched templates.

The results of a first evaluation using the CLiDE benchmark set is showing us that we are dealing with the same problems: recognition of chiral bonds and OCR. As the different algorithms are not developed for this special task, there are some errors and short comings which have to be taken into account. Especially the vectorization process introduces some unwanted effects (disconnecting lines, double bonds are no longer parallel …). At the moment we are working on replacing some of the standard algorithms with our own specialized version. The next step is to

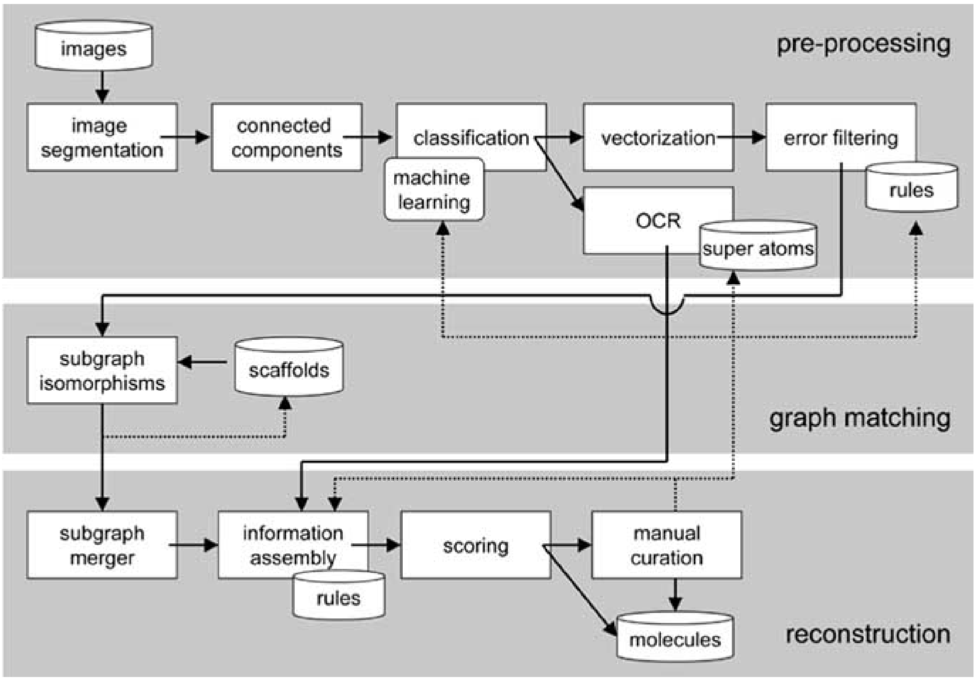


Fig. (6). The structure reconstruction process of CSR can be broken down into three main phases: image pre-processing, graph matching of predefined drug like scaffolds, and reconstruction of the molecular graph (i.e. combining matched fragments and assigning atom and bond types).

assemble a real large dataset in order to train our machine learning methods.

ROADS TO GO I: LINKING CHEMICAL COMPOUND STRUCTURE INFORMATION TO BIOLOGICAL EFFECTS

Medicinal chemistry, pharmacology and toxicology are scientific disciplines in which information on compounds and compound classes are tightly linked to information on biological objects, networks and processes. Meaningful relationships between compounds are not only encoded by their structural similarity, but could also exist at the phenotypic level (e.g. the activation of identical de-toxification pathways). In an attempt at making both types of relationships accessible in the same information system, Sing *et al*. [50] developed a system named “Text Influenced Molecular Indexing” (TIMI). The approach is based on the representation of textual indices as a matrix of terms and documents, whereas chemical indices are represented by a matrix comprising molecules and descriptors. This system uses singular value decomposition (SVD) for a mapping of compounds to text sources and contextual information. However, context is defined as co-occurrence of terms in this approach and true information extraction based on natural language processing is not part of the approach.

The report by Singh *et al*. is one of the first public communications on combining text based searches with chemical similarity searches. Although the primary focus on the paper is on document retrieval, it can easily be envisioned how this approach could be applied to true information extraction approaches. One major drawback of the approach is, that biological objects retrieved through chemical similarity search are often only weakly *associated* with chemical compounds. They are not necessarily *semantically linked* to them (e.g. organisations that conduct research on the compound of interest). This can be explained by a lack of a grammar that limits the biological search terms to concepts relevant for pharmacology and toxicology.

More specific semantic structures representing the intersection between chemistry and biology, e.g. an appropriate pharma-ontology, would allow for more dedicated approaches towards combined chemical and biological information mining in the literature. The ultimate aim would be to relate chemical structures and variations thereof to pharmacological and toxicological effects. Textual information sources, that link chemical structures to biological effects, are not confined to journal publications; actually we expect toxicological tables [72] (for more information see [73]), pharmacological bulletins (e.g. [74]), and free text fields in databases [75] to be valuable sources for this approach, too.

Information extraction approaches, described in our review, will allow us to efficiently link chemical structures to biological effects. There are numerous problems to be addressed on the road towards using information extraction approaches for the purpose to automatically assign information on targets, pathways, side effects, stability in serum, induction of cytochrome genes, and so forth to chemical compounds. The following list is certainly incomplete, but tries to highlight some of the major problems we will encounter on our way towards information extraction in pharmacology, medicinal chemistry and toxicology:

- efficient recognition of chemical entities
- efficient “decoding” of chemical names from regular expressions (e.g. IUPAC)
- automated construction of chemical dictionaries
- availability of ontologies representing concepts relevant for the description of toxicology, pharmacology, and medicinal chemistry
- extraction of information on chemical and biological conditions under which pharmacological or toxicological effects have been observed

Even though this list is filled with significant challenges, the success of information extraction approaches in the area of molecular biology and medicine encourages us to continue to work on dedicated systems for the extraction of pharmacological (and related) knowledge. However, the special form of communication of chemical information, like structures represented as molecular graphs, prompts us to expand the current work towards the reconstruction of chemical structure information from images:

ROADS TO GO II: MULTIMODAL INFORMATION EXTRACTION

We have shown in this paper that information extraction technologies have demonstrated their potential in the domain of molecular biology including functional genomics and that the time has come to adapt the information extraction approach to the fields of medicinal chemistry, pharmacology and toxicology. We have also shown, information extraction should not be restricted to textual sources, as chemistry-centred sciences, e.g. pharmacology and medicinal chemistry, widely use images of chemical structures to transport information on chemical entities.

One of the major challenges in information extraction approaches in medicinal chemistry, pharmacology and toxicology is the multimodal extraction of information from both, text and images. In patents and in other scientific publications we often find images of chemical core structures (so called *Markush structures* [76]) with R-groups associated with this core structure. The chemical composition of these R-groups is normally specified in the text. Frequently, we find a complex set of chemical variants of these R-groups described in text. This combinatorial way of description results in a tremendous variability of molecules. The multi-modal reconstruction of the chemical space represented by Markush structures and their R-groups in the literature marks another “grand challenge” in information extraction in the field of pharmacology.

The perspectives of the application of information extraction approaches to medicinal chemistry, pharmacology, and toxicology are auspicious, not only for commercial (industrial) research, but also for academic (public) research. At present, academic research in the field of *in silico* pharmacology and toxicology is hampered by the fact that a significant portion of the knowledge associated with compounds is available only in commercial databases. Although we have understanding for the fact, that the generation of high quality information is expensive. But we also have to realise, that limitations in access to public information on molecules and their biological effects are restricting the advancement of science in pharmacology and toxicology. The public genome projects may give an idea how publicly accessible information stimulates research and development. We therefore see an unparalleled opportunity of the current developments in the field of information extraction technology, to use this sort of approach to make high quality information on compounds, their synthesis, their variants, and their biological targets and effects available to the public. The PubChem project, initially started with a similar motivation, is currently facing strong resistance from commercial information suppliers in the US [77]. The legal situation as well as the intention of public scientific funding institutions in Europe might be different and the chances to start a European initiative on using information extraction technology to generate a public database for compounds, their synthesis, and their biological effects are very good. We are convinced, that providing this sort of information to the research community as a whole, will certainly boost academic and commercial research in medicinal chemistry, pharmacology, and toxicology.

REFERENCES

[1] PubMed. A service of the National Library of Medicine. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed.

[2] Shatkay, H.; Feldman, R. Mining the biomedical literature in the genomic era: an overview. *JCB,* **2003,** *10,* 821-855.

[3] Gieger, C.; Deneke, H.; Fluck, J. The future of text mining in genome-based clinical research. *Biosilico.,* **2003,** *1,* 99-103.

[4] Pustejovsky, J.; Castano, J.; Zhang, J.; Kotecki, M.; Cochran, B. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *PSB 2002 – Proceedings of the Pacific Symposium on Biocomputing 2002,* Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Kevin Lauderdale, and Teri E. Klein, editors. Singapore: World Scientific Publishing, **2002**; pp 362–373.

[5] Saric, J.; Jensen, L. J.; Ouzounova, R.; Rojas, I.; Bork, P. Extracting regulatory gene expressions expression networks from PubMed. In *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*. Association for Computational Linguistics, **2004**.

[6] Narayanaswamy, M.; Ravikumar, K. E.; Vijkay-Shanker, K. A biological named entity recognizer. In *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003,* Russ B. Altman, A. Keith Dunker, Lawrence Hunter, and Teri E. Klein, editors, Lihue, Singapore: World Scientific Publishing, **2003**; pp 427–438.

[7] Wermter, J.; Hahn, U.. Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. In *Proceedings of the 11thWorld Congress on Medical Informatics. Vol. 1*, number 107 in Studies in Health Technology and Informatics, Marius Fieschi, Enrico Coiera, and Yu-Chan Jack Li, editors, Amsterdam: IOS Press, **2004**; pp 560–564.

[8] Wermter, J.; Fluck, J.; Stroetgen, J.; Geißler, S.; Hahn, U. *Recognizing noun phrases in biomedical text: An evaluation of lab prototypes and commercial chunkers.* First International Symposium on Semantic Mining in Biomedicine (SMBM), European Bioinformatics Institute, Hinxton, Cambridgeshire, UK. 10th -13th April, **2005.**

[9] Kim, J.-D.; Ohta, T.; Tsuruoka, Y.; Tateisi, Y.; Collier, N. *Introduction to the Bio-Entity Task at JNLPBA* Yeh, A.; Morgan, A.; Colosimo, M.; Hirschman, L. BioCreAtIvE Task 1A: gene mention finding evaluation *BMC Bioinformatics,* **2005**, *6*(Suppl 1), 2.

[10] Hirschman, L.; Colosimo, M.; Morgan, A.; Yeh, A. Overview of BioCreAtIvE task 1B: normalized gene lists.*BMC Bioinformatics,* **2005**, *6*(Suppl 1), 11.

[11] Hanisch, D.; Fundel, K.; Mevissen, H.-T.; Zimmer, R.; Fluck, J. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics,* **2005**, *6*(Suppl 1), 14.

[12] Epstein J. A.; Kans J. A.; Schuler G. D.: WWW Entrez: A Hypertext Retrieval Tool for Molecular Biology. *Proceedings of the Second International WWW Conference* **94**. Chicago, I1, http://*www3.ncbi.nlm.nih.gov/WWW2/WWW2_ paper2.html*.

[13] ENTREZ, *http://www.ncbi.nlm.nih.gov/Database/index.htm*.

[14] SWISS-PROT, http://www.expasy.org/sprot/.

[15] The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nature Genet.,* **2000**, *25*, 25–29. *www.geneontology.org*.

[16] Muller H. M.; Kenny E. E.; Sternberg P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.,* **2004**, *2*(11), e309.

[17] Jenssen, T.-K., *et al*. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.,* **2001**, *28*, 21-28.

[18] NLM. 2003. Mesh: Medical subject headings. *www.nlm.nih.gov/mesh/*.

[19] Lindberg, D.A. *et al.* The unified medical language system. *Meth. Inform. Med.* **1993**, *32*(4), 281–291. *www.nlm.nih.gov/research/umls*.

[20] Craven, M.; Kumlien, J. Constructing biological knowledge bases by extracting information from text sources. *Proc. AAAI Conf. on Intelligent Systems for Molecular Biology (ISMB),* **1999**, 77–86.

[21] Ray, S.; Craven, M. Representing sentence structure in hidden Markov models for information extraction. *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI-01),* **2001**.

[22] Friedman, C. *et al*. Genies: A natural-language processing system for the extraction of molecular pathways from journal articles. *Proc. Int. Conf. on Intelligent Systems for Molecular Biology (ISMB),* **2001**; 74–82.

[23] Xiao, J.; Su, J.; Zhou, G. D.; Tan, C. L. Protein-Protein Interaction Extraction: A Supervised Learning Approach. *First International Symposium on Semantic Mining in Biomedicine (SMBM),* European Bioinformatics Institute, Hinxton, Cambridgeshire, UK, *10th -13th April*, **2005**.

[24] Yakushiji, A.; Miyao, Y.; Tateisi, Y.; Tsujii, J. Biomedical Information Extraction with Predicate-Argument Structure Patterns. *First International Symposium on Semantic Mining in Biomedicine (SMBM),* European Bioinformatics Institute, Hinxton, Cambridgeshire, UK, *10th -13th April*, **2005**.

[25] Yeh, A., *et al.* Background and overview for KDD Cup 2002 task 1: Information extraction from biomedical articles. *SIGKDD Explorations,* **2002**, *4*(2), 87–89.

[26] Blaschke, C.; Leon, E. A.; Krallinger, M.; Valencia, A.. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics,* **2005**, *6*(Suppl 1), 16.

[27] Gieger, C.; Hanisch, D.; Fluck, J.; Mevissen, H.-T.; Tresch, A.; Deneke, H. Using Text Mining Networks for the Context Specific Interpretation of Gene Expression Data. 2003 *Biometrical Journal* **2004**, *46,* Supplement Abstracts of the Joint Meeting of the IBS-DR and the DAE, 56.

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

[29] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf .Comp. Sci.*, **1988**, *28*, 31-36.

[30] IUPAC *http://www.iupac.org/general/about.html*.

[31] Bauerschmidt, S.; Gasteiger, J. Overcoming the Limitations of a Connetion Table Description: A Universal Representation of Chemical Species. *J. Chem. Inf .Comp. Sci.*, **1997**, *36*, 705-714.

[32] PubChem database. *http://pubchem.ncbi.nlm.nih.gov*.

[33] Negwer, M.; Scharnow, H.-G. *Organic-Chemical Drugs and Their Synonyms*. Wiley-Vch, **2001**.

[34] IUBMB: Enzyme Nomenclature. *Academic Press, San Diego,* **1992**, ISBN 0-12-227164-5, *http://www.chem.qmul.ac.uk/iubmb/.*

[35] Chemical Abstracts Services CAS, *http://www.cas.org/EO/regsys.html.*

[36] CrossFire Beilstein database, *http://www.mdl.com/products/knowledge/crossfire_beilstein.*

[37] Developmental Therapeutics Program NCI/NIH, *http://dtp.nci.nih.gov.*

[38] Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening, *JCIM*, **2005**, *45*, 177-182.

[39] Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the chemical semantic web through the use of InChI identifiers. *Org Biomol Chem*., **2005**, *3*(10): 1832-1834.

[40] Reference Sequence (RefSeq), *http://www.ncbi.nlm.nih.gov/RefSeq.*

[41] Dobson, C. M. Chemical space and biology. *Nature*, **2004**, *432*, 824-828.

[42] INN International Nonproprietary Names (INN). *http://www.who.int/medicines/oragnization/qsm/activities/qualityassurance/inn/orginn.shtml*.

[43] USAN United States Adopted Names (USAN). *http://www.ama-assn.org/ama/pub/category/2956.html*

[44] Resource Description Framework, *http://www.w3.org/RDF.*

[45] Name-to-Structure, ACD, *http://www.acdlabs.com/download/iupacname.html.*

[46] AutoNom, MDL, *http://www.mdl.com.*

[47] NameIt/DrawIt, Bio-Rad, *http://www.bio-rad.com.*

[48] Nomenclator/NamExpert, Chem4D, *http://www.cheminnovation.com.*

[49] nam2mol , OpenEye, *http://www.eyesopen.com*.

[50] Singh, S. B.; Hull, R. D.; Fluder, E.M. Text Influenced Molecular Indexing TIMI: A Literature Database Mining Approach that Handles Text and Chemistry, *J. Chem. Inf. Comp. Scim.,* **2003**, *43*, 743-752.

[51] Hanisch, D.; Fluck, J.; Mevissen, H.-T.; Zimmer, R. Playing biology's name game: identifying protein names in scientific text. *Pac. Symp. Biocomput.,* **2003**, 403-414.

[52] Brooksbank, C.; Cameron, G.; Thornton, J.: The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Research*, **2005**, *33*, 46-53. *http://www.ebi.ac.uk/chebi/.*

[53] Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; Cambridge, Massachusetts, The MIT Press, *Introduction to algorithms*, **1992**.

[54] DS Accord Cartridge, Accelrys, *http://www.accelrys.com.*

[55] Oracle Cartridge, CambridgeSoft, *http://www.cambridgesoft.com.*

[56] JChem Cartridge, ChemAxon, *http://www.jchem.com.*

[57] DayCart, Daylight, *http://www.daylight.com.*

[58] AUSPYX, TRIPOS, *http://www.tripos.com.*

[59] InfoChem Cartridge, InfoChem, *http://www.infochem.de.*

[60] MDL ISIS, MDL, *http://www.mdl.com.*

[61] Chemical Database Cartridge (MolCart), Molsoft, *http://www.molsoft.com.*

[62] Murray-Rust, P.; Rzepa, H. S.; Chemical Markup, XML, and the world wide web. 4. CML Schema, *JCICS,* **2003**, *43*, 757-772.

[63] Jain, A. K.; Duin R. P. W.; Mao, J. Statistical Pattern Recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.,* **2000**, *22*, 4-37.

[64] Bunke, H. *Handbook of Pattern Recognition and Computer Vision*, chapter 1.5. World Scientific Publishing. **1995**, 163-209.

[65] Brown, J. R.; Balmuth, J. R. Kekule: OCR – optical chemical (structure) recognition. *J. Chem. Inf. Comp. Sci.,* **1992**, *32*, 373-378.

[66] Ibison, P.; Jaquot, M.; Neville, A. G.; Simpson, R. W.; Tonnelier, C.; Venczel, T.; Johnson, A. P. Chemical Literature Data Extraction: The CLiDE Project. *J. Chem. Inf. Comp. Sci.,* **1993**, *33*, 338-344.

[67] RxList LLC. *http://www.rxlist.com.*

[68] ChemDraw, CambridgeSoft, *http://www.cambridgesoft.com/*

[69] Fey, T. Validation of chemical structure recognition software for 2d drawings and a following graphcials error curing. *Matster's thesis*, Fachhochschule Bonn-Rhein-Sieg, **2004**.

[70] Scale Vector Graphics (SVG). *http://www.w3.org/graphics/svg.*

[71] Messmer, B. T. Efficient Graph Matching Algorithms for Preprocessed Model Graphs. *PhD thesis*. Inst. of Comp. Science and Appl. Mathematics, University of Bern, **1996**.

[72] Maximale Arbeitsplatz Konzentration, MAK List, *http://www3.interscience.wiley.com/cgi-bin/mrwhome/104554790/HOME.*

[73] http://www.foodrisk.org/general_chemical_hazards_RA.cfm.

[74] http://www.emea.eu.int.

[75] http://toxnet.nlm.nih.gov.

[76] Berks, A. H. Current state of the art of Markush topological search systems, World Patent Information **2001**, 5-13.

[77] http://pubs.acs.org/cen/news/83/i17/print/8317notw1.html.