Information Extraction from Text for Improving Research on Small Molecules and Histone Modifications

Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Corinna Klein
geb. Kolářik
aus
Zittau

Bonn 2011

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. rer. nat. Martin Hofmann-Apitius
2. Gutachter: Prof. Dr. rer. nat. Holger Fröhlich

Tag der Promotion: 10. Juni 2011
Erscheinungsjahr: 2011

Abstract

The cumulative number of publications, in particular in the life sciences, requires efficient methods for the automated extraction of information and semantic information retrieval. The recognition and identification of information-carrying units in text – concept denominations and named entities – relevant to a certain domain is a fundamental step. The focus of this thesis lies on the recognition of chemical entities and the new biological named entity type histone modifications, which are both important in the field of drug discovery. As the emergence of new research fields as well as the discovery and generation of novel entities goes along with the coinage of new terms, the perpetual adaptation of respective named entity recognition approaches to new domains is an important step for information extraction. Two methodologies have been investigated in this concern: the state-of-the-art machine learning method, Conditional Random Fields (CRF), and an approximate string search method based on dictionaries. Recognition methods that rely on dictionaries are strongly dependent on the availability of entity terminology collections as well as on its quality. In the case of chemical entities the terminology is distributed over more than 7 publicly available data sources. The join of entries and accompanied terminology from selected resources enables the generation of a new dictionary comprising chemical named entities. Combined with the automatic processing of respective terminology – the dictionary curation – the recognition performance reached an F_1 measure of 0.54. That is an improvement by 29 % in comparison to the raw dictionary. The highest recall was achieved for the class of TRIVIAL-names with 0.79.

The recognition and identification of chemical named entities provides a prerequisite for the extraction of related pharmacological relevant information from literature data. Therefore, lexico-syntactic patterns were defined that support the automated extraction of hypernymic phrases comprising pharmacological function terminology related to chemical compounds. It was shown that 29-50 % of the automatically extracted terms can be proposed for novel functional annotation of chemical entities provided by the reference database DrugBank. Furthermore, they are a basis for building up concept hierarchies and ontologies or for extending existing ones. Successively, the pharmacological function and biological activity concepts obtained from text were included into a novel descriptor for chemical compounds. Its successful application for the prediction of pharmacological function of molecules and the extension of chemical classification schemes, such as the the Anatomical Therapeutic Chemical (ATC), is demonstrated.

In contrast to chemical entities, no comprehensive terminology resource has been available for histone modifications. Thus, histone modification concept terminology was primary recognized in text via CRFs with a F_1 measure of 0.86. Subsequent, linguistic variants of extracted histone modification terms were mapped to standard representations that were organized into a newly assembled histone modification hierarchy. The mapping was accomplished by a novel developed term mapping approach described in the thesis. The combination of term recognition and term variant resolution builds up a new procedure for the assembly of novel terminology collections. It supports the generation of a term list that is applicable in dictionary-based methods. For the recognition of histone modification in text it could be shown that the named entity recognition method based on dictionaries is superior to the used machine learning approach.

In conclusion, the present thesis provides techniques which enable an enhanced utilization of textual data, hence, supporting research in epigenomics and drug discovery.

Acknowledgments

Herewith, I would like to take the opportunity to thank Prof. Dr. Martin Hofmann-Apitius for giving me the opportunity to work on my thesis at the Bioinformatics department of the Fraunhofer Institute SCAI. Furthermore, I would like to thank Prof. Dr. Holger Fröhlich for his willingness to be the co-referent of the thesis. Special thanks go to Dr. Juliane Fluck, who introduced me to text mining, gave me strong support during my work, and critically reviewed the thesis.

I am very grateful to Theo Mevissen for his technical support, especially with ProMiner. I would like to thank Roman Klinger for providing his adapted Mallet-implementation of CRFs to me and for the many hours of good discussions. Furthermore, I appreciate the good cooperation with Harsha Gurulingappa, whose master thesis was supervised by me. I thank all other colleagues that accompanied me during my time at SCAI and thank the Bonn-Aachen International Center for Information Technology (B-IT) for the financial support of my thesis. Last but not least, special thanks are dedicated to my husband Adrian for his encouragement during the whole time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Research in medicine, biology, pharmacology, and toxicology deals with understanding the function of cellular and systemic processes, the mechanisms of disease development as well as the distribution and metabolization of chemicals, drugs, and nutraceuticals in cell cultures, model organisms and humans.

Pharmaceutical research specifically aims at identifying chemicals that well-directedly influence biological processes for treating medical conditions or alleviating disease symptoms. It is a complex, multistage process that includes many scientific disciplines and presupposes the analysis of a huge amount of heterogeneous, distributed information. The drug development steps span from hypothesis generation, target identification and validation to lead design, candidate support and clinical trials as well as the analysis and prediction of adverse events.

The retrieval, linkage, analysis, and interpretation of available information and experimental data as well as the deduction of mechanisms and rules is important for the generation of new hypotheses as well as for decision making. Likewise to the different fields that are involved in pharmaceutical research processes, the required information and data are scattered over heterogeneous databases and unstructured data resources, e.g. papers, reports, patents, blogs, and wikis.

The collection and storage of information on chemicals and pharmaceuticals has a long tradition, demonstrated by text resources, like the *'Beilstein's Handbuch der organischen Chemie'* started by F. K. Beilstein in 1880, which contains chemical, physical, pharmacological and physiological properties on organic compounds[1]. The advantage of databases that emerged from such resources is the possibility to query for certain structures, substructures or defined properties.

However, a restriction of databases is their data model, which in fact has to be extended for the inclusion of new data types, to reflect novel topics and trends by model evolution, which is a non-trivial task for complex data models [Chen et al., 1994]. Another primary challenge is the inclusion of new data sets, its curation (i.e. a quality checking) as well as the progressional annotation of present entities with new findings.

The rapid accumulation of new publications, which must be processed by human curators to extract data of new discoveries and to revise present ones, represents an additional difficulty for keeping databases with biological, pharmaceutical and chemical information up-to-date. Baumgartner et al. [2007] for instance predict that the annotation of biological entities, like proteins and genes of diverse organisms with terms from Gene Ontology[2]

[1] `http://www.beilstein-institut.de/index.php?id=111`

[2] www.geneontology.org

[Ashburner et al., 2000], will only be manageable with high efforts and will still take a long time from now on.

Apart from their advantages, databases alone cannot capture the richness of scientific information and argumentation contained in the literature. Complex assumptions, interpretations of findings and hypotheses usually are expressed and set into context by natural language. It is the basic medium to distribute ideas and recent findings, thus making it available to a broad readership. Hence, a high amount of the most current information obtained by laborious research does not reside in structured data sources, but in unstructured text. According to Hale [2005] only a small proportion of information is available in structured form manageable by database systems, whereas around 80 % is unstructured and written in natural language. Thus, the examination and analysis of a wide range of scientific literature, patents, adverse event reports, company reports, news, and patient generated content is a key element of science and research, especially in pharmaceutical research. For bench scientists, published data is the best source for interpreting their experiments and to keep up-to-date with the most current scientific knowledge. They provide contextual factors; especially for those topics that came up after the related databases were set up. In general, literature-based discovery has often been held out as a potential source of promising hypotheses [Zweigenbaum et al., 2007]. The study of literature enables the identification of novel facts, new connections between entities of interest and drives the generation of new ideas to be further explored and validated by experimentation.

However, the goal is hard to achieve by reading all documents as the amount of scientific literature is constantly increasing, especially in biomedicine and pharmacology-related fields. Research in this area shows a very high level of specialization and, consequently, the knowledge is often fragmented. As a result, written information is scattered over many journals from specialized subfields. Researchers can be aware of articles being published within their own community, but might not find connections to other related research results because they did not expect other links.

Thanks to the improvement and availability of computer systems and the world wide web it is becoming easier to store and distribute data; however it becomes harder to access and manage them [Claus and Underwood, 2002]. Meanwhile, an increasing number of articles are accessible via the worldwide web as electronic texts. As a consequence, scientists from academia and industry are faced with an overload of textual information available. Figure 1.1 exemplarily shows the steep increase of publications collected by the bibliographic database MEDLINE from 1950 to 2009.

The enormous growth of biomedical literature has urged the development of domain-specific efficient informatics tools to organize and support the analysis of this huge amount of unstructured data for accomplishing the discovery of so far unknown associations, hypotheses or trends. Furthermore, automated text utilization is needed to support the combination of information on biomedical entities residing in scientific literature with known facts in databases. This supports the annotation of entities in repositories with new properties and the use of as much information as possible for research.

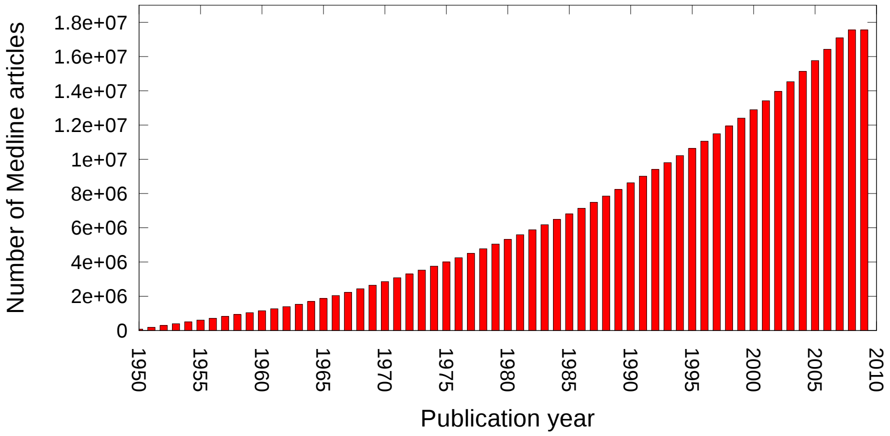


Figure 1.1: Cumulative number of published articles in MEDLINE from 1950 until the beginning of 2009.

1.1 Overview on the Biomedical Relevance and Information Resources of Chemical Entities and Histone Modifications

Chemical molecules are in the center of research in pharmacology and medicine which aim to understand the role of small-molecules in biological processes and to develop new pharmaceuticals for treating diseases. Information about biochemical reactions and interactions with targets, which are mainly proteins, as well as the mode of action of chemical compounds and their possible toxic effects are of high interest in the early stages of drug development. In later phases side and adverse drug effects (ADR) are investigated to prevent serious health problems caused by the utilization of pharmaceuticals. Interestingly, in the last years side effect information have become a trigger of innovative development for finding new applications of existing drugs, also called drug repositioning [Rikken and Vos, 1995, Grau and Serbedzija, 2005, Campillos et al., 2008]. They help to identify new off-side targets and to investigate novel modes of action.

The types of information produced in biomedical and pharmacological research are diverse. They span from numeric measured values obtained through experiments, like affinity or inhibition constants, to the descrition of the mode of action, side effects or ADR in form of natural language, such as *'dihydrofolate reductase inhibitor'*. Such terms are broadly applied for the annotation and classification of chemical compounds.

Information on chemical molecules is dispersed over diverse and dynamic resources. On one hand side there is a high amount of natural language data in form of scientific articles, patents, free text fields in data sources, etc. and on the other hand structured information resides in databases or classification systems. As research in chemistry and pharmacology has been and is still motivated by commercial interests, many chemical data sources are offered by private information and content providers or were developed in house of pharmaceutical companies. Thus, many available databases on chemical compounds, e.g. reviewed by [Jónsdóttir et al., 2005], are not freely accessible. Furthermore, access is often only provided on a per-item basis which hinders to obtain a data collection as a whole. A description on the historical development of this situation in chemistry is given by Murray-Rust et al. [2005]. In contrast, the community in biology and bioinformatics generated a high number of databases with the intention to allow access for everyone. Fortunately, with the generation of freely accessible databases such as KEGG [Goto et al., 1998], DrugBank [Wishart et al., 2006] or PubChem[3] as well as the ontology ChEBI [Degtyarenko et al., 2008], the situation in chemistry has started to change in the last decade. However, for getting a substantial overview on available data it is mandatory to collect and integrate the information on chemical compounds from diverse data sources. Different data types, formats and used terminology are great challenges that are related to this process. Furthermore, data quality is another important issue. Bradley [2008] for instance state that PubChem include data from suppliers without an extensive curation procedure. Therefore it accumulated structure and information errors.

One of the few initiatives that provides a free-to-access collection of compound data from across the web and repositories in form of a database is ChemSpider[4]. It *"aggregates chemical structures and their associated information into a single searchable repository and makes it available to everybody"*. Furthermore, with the help of the automatic recognition of chemical names in documents and web pages ChemSpider aggregates links to available data repositories. It provides diverse physicochemical properties shown in an online overview on the assembled data[5]. However, it is not dedicated to extract and aggregate mode of action descriptions from the growing amount of natural language data for the annotation of chemical compounds. Function annotation of chemical compounds is the addition of attributes which are usually descriptions of properties, like the biological function, a pharmacological effect of a compound or the membership to a certain structural class. Terms represent these properties, e.g. *'antiinflammatory agent'* and *'cyclooxygenase-2 inhibitor'*. They are provided as controlled vocabularies by chemical/pharmacological classification schemes, thesauri or ontologies, such as Anatomical Therapeutic Chemical (ATC) Classification System[6] and the United States Pharmacopeia (USP)[7], MeSH[8] or ChEBI[9] respectively. Furthermore, annotations of chemical as well as pharmaceutical compounds are collected in databases, such as DrugBank or PubChem. However, as providers of repositories put different effort in the curation of their data, they comprise different levels of annotation completeness and correctness. In addition to the structured annotation data a further important resource for finding new annotation information are scientific articles. For example, according to Agarwal and Searls [2008] a high number of new drug targets derive from novel biological discoveries first appearing in the scientific literature from academic sources. They further state that known targets are more frequently functionally characterized in some new way or associated with a disease process than targets are newly discovered. Thus, the exploitation of literature is important for making progress in pharmaceutical research. It is furthermore a valuable

[3] http://pubchem.ncbi.nlm.nih.gov/

[4] http://www.chemspider.com/

[5] http://www.iupac.org/publications/ci/2008/3001/ic_chemspider.html

[6] http://www.whocc.no/atc_ddd_index/

[7] http://www.usp.org/USPVerified/pharmaceuticalIngredients/

[8] www.nlm.nih.gov/mesh/

[9] www.ebi.ac.uk/chebi/

resource for generating new hypotheses, especially when new research fields emerge, such as epigenetics, which is introduced in the following section.

1.1.1 Introduction to Epigenetics and its Role in Biology, Medicine, and Pharmacology

Epigenetics gained growing interest in molecular biology as well as in medicine in the last years and investigates

> "stably heritable phenotypes resulting from changes in a chromosome without alterations in the DNA sequence" [Berger et al., 2009].

To explain the molecular background of epigenetic phenomena as well as to make its biomedical and pharmacological relevance clear, an excursus to epigenetic research and molecular biology is provided in the following paragraph.

Histone modifications – chemical modifications of proteins attached to chromatine, the methylation of cytosines of DNA strands and its regulating proteins are in the focus of this relatively young research field. Histone proteins form the core of nucleosomes – the chromatin unit – each consisting of four distinct histone protein dimers. As Figure 1.2 illustrates, DNA double strands are wrapped around this protein complex, whereas the histone tails protrude from the nucleosome core to the outside. Hence, its amino acids are accessible to enzymes covalently introducing different small chemical groups or small molecules or processing the amino acid itself.

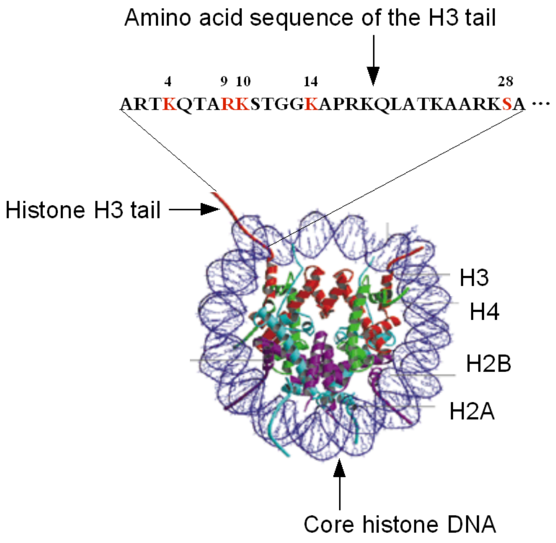

Amino acid sequence of the H3 tail
4
9 10
14
28
ARTKQTARKSTGGKAPRKQLATKAARKSA ···
Histone H3 tail
H3
H4
H2B
H2A
Core histone DNA

Figure 1.2: Nucleosome structure and amino acid sequence of the protein tail from histone H3. Amino acid positions that are often modified are depicted in red. H2A, H2B, H3, and H4 denote the four different histone proteins being part of the nucleosome core protein complex. (The figure was adapted from Marmorstein [2001].)

To date several histone modifying groups, molecules or transformation processes are known which are depicted in Table 1.1. The chemical modifications of amino acids change their physico-chemical properties and mark histone proteins for the recruiting of other proteins, thus participating in the formation of the different chromatin structure states – hetero- and euchromatin [Margueron et al., 2005]. Hence, the protein modifications alter the accessibility of DNA for the transcription machinery leading to either enhancing or silencing of the expression of genes. Exemplarily, the modification of the amino acid lysine with three methyl groups is shown in Figure 1.2. More than 70 sites for histone post-translational modifications (PTMs) have been reported [Taverna et al., 2007], whereas combinations of different histone modifications on one or different histones form a kind of code. An important feature of these marks is their ability to crosstalk, which is essential for transcription regulation. They often act in concert, and multiple feed-forward and feed-back mechanisms involving the same nucleosome or histone, or distinct nucleosomes and histones have been identified [Gräff and Mansuy, 2008]. Furthermore, it was found that DNA and histone methylation are coordinately regulated [Smallwood et al., 2007]. These modifications are stable, can be long-term and inherited over several cell divisions, making epigenetic regulation a key mechanism for cellular differentiation and cell fate decisions, also leading to stably heritable phenotypes [Szyf, 2007, Gräff and Mansuy, 2008, Berger et al., 2009]. However, they carry expression regulation functionalities [Jenuwein and Allis, 2001] which are not fully understood until now. Additionally, not all factors taking part in the regulation of the modifications itself are known.

Modification Types	Modification examples
Groups	acetyl, methyl, phosphate, adenine diphosphate (ADP) ribosyl, carbonyl, sumoyl
Molecules	biotin, ubiquitin
Process	proline isomerization, arginine deamination (i.e. citrulline generation)

Table 1.1: Examples of histone modifying groups, molecules, and processes [Margueron et al., 2005, Latham and Dent, 2007, Cuthbert et al., 2004, Nelson et al., 2006].

Meanwhile, the epigenetic marking of chromatin is recognized as potentially important biomarker of disease states and drug side effects, and provides a link between the environment and gene expression regulation [Herceg, 2007, Szyf, 2007]. It is becoming increasingly clear, that epigenetic mechanisms account for several diseases, like cancer [Ting et al., 2006], autoimmune diseases [Wilson, 2008], and neuropsychiatric disorders, as well as for side effects of drugs and cell aging [Santos-Rebouças and Pimentel, 2007, Feinberg, 2007, Lund and van Lohuizen, 2004, Tryndyak et al., 2006, Kavlock et al., 2008, Dang et al., 2009]. Lifestyle and diet might induce epigenetic changes that are most likely subtle and cumulative [Herceg, 2007] and can influence the susceptibility for diseases [Fraga et al., 2005]. Furthermore, epigenetic mechanisms contribute in a major way to the functioning of the brain, e.g. they are involved in memory and learning processes [Isles and Wilkinson, 2008, Gräff and Mansuy, 2008]. Several behavioral pathologies might be a consequence of psychosocial early life exposures, which altered epigenetic programming.

It is expected that signal transduction pathways, which are activated by cell-surface or intra-cellular receptors, are linked to epigenetic changes, connecting environmental or physiological events with a reprogramming of gene activity. Moreover, it is possible that epigenetic processes might override genetic polymorphisms at distinct points in life and in specific tissues exclusively. Therefore, they might be related to the susceptibility for certain diseases, especially for late onset diseases or medical conditions in subsequent generations [Kavlock et al., 2008]. For instance Anway et al. [2005] summarize the transgenerational effects of endocrine disruptors, like fungicides and pesticides in relation to male infertility caused by epigenetic changes. Also drugs with well-established mechanisms of action and therapeutic targets, like the antihypertensive hydralazine and the antiarrhythmic procainamide, were shown to affect DNA methylation and cause broad epigenetic reprogramming in T cells [Cornacchia et al., 1988]. Another well known example is the antiepileptic drug Valproic acid. For years it was considered to be a gamma-Aminobutyric acid (GABA) receptor stimulator, but was later found to be an Histone deacetylase (HDAC) inhibitor [Göttlicher et al., 2001]. Another drug causing epigenetic alterations is Tamoxifen which was found to be itself responsible for enhancing breast cancer cell lines to become Tamoxifen-resistant [Badia et al., 2007] and is described by Tryndyak et al. [2006] to be a potential hepatocarcinogen.

As the examples clearly show, unexpected environmental toxic and pharmacological agents might target the class of chromatin modifiers or influence signaling pathways which are highly responsive to drugs and toxic agents, thus affecting the long-term programming of the genome in diverse tissues [Szyf, 2007]. It leads to the conclusion that research in pharmacology has to contemplate potential hazards of drugs to the epigenome in the future.

Rising research interest for this field resulted in a steep increase of literature data in the last years. It is a fast growing resource for studying histone modification-related information, boosted by the development of two experimental methods, ChIP-chip [Ren et al., 2000] and Chip-Sequencing [Johnson et al., 2007]. They enabled the investigation of interactions between modified histones and DNA at a genome wide range, which is reflected by a steep increase of publications since 2000 shown in Figure 1.3.

Although there are databases providing epigenetic data, like the UCSC Genome Browser[10] [Koch et al., 2007], the ChromatinDB[11] [O'Connor and Wyrick, 2007], the Histone Database[12] [Marino-Ramírez et al., 2006], and ChromDB[13] [Gendler et al., 2008], MeInfoText[14] of [Fang et al., 2008] and PubMeth[15] [Ongenaert et al., 2008], however, in its current version, they do not support the analysis of histone modifications across species, nor are they related to studied cell or tissue types, phenotypes, diseases or chemicals. Hence, the highly topical and considerable context information on histone modifications resides in natural language text in the form of scientific publications.

However, no efforts have been put into the systematic analysis of the terminology used

[10] http://genome.ucsc.edu/

[11] http://www.bioinformatics2.wsu.edu/cgi-bin/ChromatinDB/cgi/visualize_select.pl

[12] http://genome.nhgri.nih.gov/histones/

[13] www.chromdb.org

[14] http://mit.lifescience.ntu.edu.tw/

[15] www.pubmeth.org

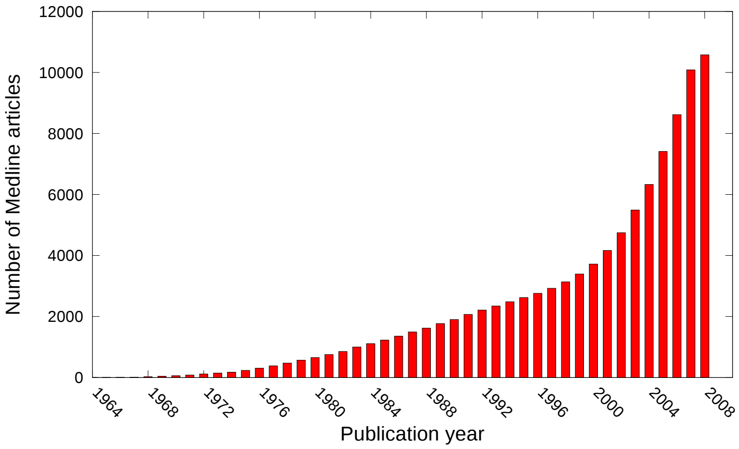


Figure 1.3: Cumulative number of published articles on histone modifications in MEDLINE from 1964 until June 2008.

for the description of histone modifications in the literature and no approach was developed so far to identify it in text.

1.2 Overview on Text Processing Methods

With the growing amount of natural language data accumulating in biomedicine, pharmacology, and chemistry, two fundamental needs are associated which have to be satisfied by automated methods:

- Information Retrieval (IR): The finding of relevant documents from large collections that satisfy an information need.
- Information Extraction (IE): The extraction of defined informative text parts usable for successive data mining approaches.

Whereas the first methods developed for information retrieval date back to the 50th, information extraction methods have been evolving since about two decades ago [Cardie, 1997, Sarawagi, 2008]. Considering the rising amount of scientific literature, patents, reports etc. it is important to have methods at hand which support the finding of relevant documents from the large pool of text. Methods and aproaches developed for Information Retrieval help to sift through natural language data and support the finding and ranking of relevant documents which correspond to a query. Therefore, the vector space model and the probabilistic model have been developed [Singhal, 2001], whereas text is represented by index terms. These could be single words, word stems or phrases that are obtained from text [Takenobu et al., 2000]. However, such approaches also bear limitations; they do not cope well with ambiguous terms and complexity [Hale, 2005]. Especially biomedical and chemical entities posses a high number of synonyms and provide a specific challenge for document retrieval tasks. Thus, Information Extraction techniques, especially the recognition of entity denominations and its mapping to one unique representation is carried out in support of other tasks and hence forms a part of a process pipeline. Extracted facts can be the data input for ontology or network construction and help to improve information retrieval systems or they are subjected to data mining algorithms [Cohen and Hunter, 2004, McNaught and J., 2005]. [Krallinger and Valencia, 2005] give an exemplary overview on information extraction and its linkage to information retrieval in the field of molecular biology. Information Extraction allows for the application of data mining techniques to textual data with the aim of generating new knowledge by finding unknown patterns, denoted as text mining [Hearst, 1999, Siefkes and Siniakov, 2005, Zweigenbaum et al., 2007]. Although there is this clear definition of text mining, the term is often not that strictly used by the community. It is rather utilized to describe the conglomeration of techniques adopted to textual data. However, text mining began to establish in the 90th as data mining applied to unstructured text in databases for knowledge discovery [Feldman and Dagan, 1995, Hotho et al., 2005].

The focus of this work lies on recognizing and identifying given named entity types in text as well as extracting new terminology based on methods of information extraction. It is a preliminary step for Information Retrieval and Information Extraction systems as well as for data mining approaches relying on textual information. Thus, an overview on the main challenges of Information Extraction is provided in the next section.

1.2.1 Introduction to Information Extraction

Human language is admirable for its richness and complexity, especially when describing interrelations of complex biological processes. Therefore, its formalization and making it manifest for computer processing is a big challenge.

Textual documents contain concrete data in unstructured form and thus cannot be passed directly to data mining methods for discovering general patterns. Therefore, unstructured textual data need to be processed before subjecting them to concrete applications by utilizing Information Extraction techniques, algorithms and methods performing two main tasks:

- Identification and extraction of facts of predefined types from natural language text, i.e. unstructured machine-readable documents [Riloff, 1999]. These can be entities, like *'aspirin'* and relationships between them, e.g. *'aspirin is an inhibitor of cyclooxygenase'*.
- Representation of the extracted facts in an appropriate structured form for future use, like data mining.

In general, IE cannot be used to generate new knowledge, but only to represent facts explicitly expressed in text in a more formal structure. Thus, IE methods can be considered as a first step, which then have to be followed up by data mining techniques to discover interesting relationships in the data. Usually, the extraction of specific facts depends on the user's needs. Typical subtasks of information extraction are:

- Terminology Extraction: It deals with the identification of relevant terms from text of a certain domain.
- Named Entity Recognition (NER): Is the recognition of entity names, like proteins, genes, chemical substances, diseases, etc.

IE type	Examples of published approaches and applications
Terminology extraction	Automatic recognition of multi-word terms Frantzi et al. [1998], Application: *Arrowsmith* [Smalheiser and Swanson, 1998]
Named Entity Recognition	Chemical entity recognition: *OSCAR3* [Corbett and Murray-Rust, 2006], *ChemFrag* Mack et al. [2004], protein name recognition: *ABNER* [Settles, 2005], *ProMiner* [Hanisch et al., 2003]
Relation Extraction	Gene-drug interaction finding: Chang and Altman [2004] (based on co-occurrence between entity types), *SemRep* [Ahlers et al., 2007] (considers semantic relations between entity types), protein-protein interaction finding: *iHOP* [Hoffmann and Valencia, 2004]

Table 1.2: Examples of Information Extraction approaches.

- Relationship Extraction: It deals with the identification of relationships between entities or terms.

In the 1980s IE was first established as an independent research field. At that time a number of academic and industrial research groups were working on the extraction of information from naval messages. Between 1987 and 1998 seven Message Understanding Conferences (MUCs) were launched by DARPA[16] for comparing the performance of IE systems. Since then information extraction has been experiencing rapid growth extending its applications to new domains and employing numerous new techniques [Siefkes and Siniakov, 2005].

In the last years the potential of Information Extraction methods that make natural language data more useful for research, commercial applications, etc. has been identified by the biomedical and pharmacology community. Thus, automated handling and analyzing textual data have already become integral part of pharmaceutical research [Roberts and Hayes, 2008].

Collecting data via Information Extraction techniques from text offers an opportunity to integrate many fragments of information, gathered by researchers from multiple fields of expertise, into a more complete picture exposing the interrelated roles of genes, proteins and chemical reactions in cells, tissues and organisms [Bekhuis, 2006]. The most basic use of extracted information is the direct population of a knowledge base. Furthermore, IE is carried out in support of other tasks, like information retrieval and hence forms a part of a process pipeline. Extracted facts can also be the data input for ontology or network construction or they are subjected to data mining algorithms [Cohen and Hunter, 2004, McNaught and J., 2005]. Examples of IE approaches and application of the methodologies are given in Table 1.2.

[16] Defense Advanced Research Projects Agency: `http://www.darpa.mil/`

1.2.1.1 Introduction to Named Entity Recognition

Most approaches which process natural language data basically rely on the recognition of the information carriers in text as a primary step. These are names of entities and concept denominations whose use is highly domain specific. The latter ones describe concepts of abstract nature and those of the real world. The term 'Named Entity' was coined for the Sixth Message Understanding Conference (MUC-6) and is described by:

> "A named entity is a phrase or a combination of phrases in a document that denotes a specific object or a group of objects" [Park and Kim, 2005].

The recognition of these text units is one of the most important preliminary steps in information extraction and basically supports information retrieval systems.

The problem of Named Entity Recognition (NER) was defined for the first time in the general-language domain in the context of the MUC [Grishman and Sundheim, 1996]. Initially, the main focus lied in the identification of person names, places, and organizations in large text corpora.

Meanwhile, the interest drifted also to other eager text producing fields, like the biomedical domain. Staab et al. [2002] estimated that more than 12 % of words found in biochemistry publications correspond to technical terms. These are genes, proteins, cell types, chemical substances, diseases, experimental methods, etc. Named entities and concepts used here have particular naming characteristics, like different numbers of words, capitalization, special characters, Roman numerals, etc. Their recognition provides a special challenge and became of major interest in bioinformatics within the last decade.

Several methods have been developed to support the recognition of named entities from numerous domains. The computational research aiming at automatically identifying named entities in texts forms a vast and heterogeneous pool of strategies, methods and representations [Nadeau et al., 2007, Krauthammer and Nenadic, 2004]. In the following a general overview on methods used for NER are classified by their basic underlying technique:

- **Rule-based approaches:** The earliest NER systems typically applied rule-based approaches [Cohen and Hunter, 2004, Nadeau et al., 2007]. In general, such systems consist of a set of term formation patterns using grammatical (e.g. Part-Of-Speech), syntactic (e.g. word precedence), lexical, morphological and orthographic features (e.g. capitalization) as well as domain knowledge in combination with dictionaries. They rely on a combination of regular expressions, heuristic and hand-crafted rules. However, the generation and maintenance of such rules is bound to high costs. Furthermore, rule-based NER systems lack the ability of portability, so that such approaches are often domain and language specific.

- **Dictionary dependent approaches:** They basically rely on domain-specific term lists, called dictionaries, that are identified with a string matching approach in text. The performance of the system is dependent on the comprehensiveness of the term list and the string matching algorithm.

- **Machine learning based approaches:** Here, the identification problem of entity names is converted into a classification problem. Certain properties of text tokens and inherent dependencies between them are utilized to generate a statistical model. The two challenges are the appropriate selection of a discriminating feature set and the detection of term boundaries, which are difficult to 'learn' [Krauthammer and Nenadic, 2004]. Usually, supervised learning is used for NER which involves a program that can learn to classify a given set of labeled examples. As annotated corpora have become available, Machine Learning (ML)-based approaches have attracted notice to NER research.

- **Hybrid approaches:** Most of the existing approaches use a combination of the basic methodologies and hence exploit their particular advantages. For instance, dictionaries and rules are also applied by machine learning methods.

When dealing with textual data a general problem is the prevalent use of various term and spelling variants for referring to the same real-world entity or to an abstract concept. A chemical entity for instance can be assigned to more than 100 different synonyms and spelling variants. Another challenge is the utilization of identical labels that may be related to different meanings. Often there is no one-to-one correspondence between concepts and terms which results in the problem of polysemes[17]. Term variations and ambiguous terms in text and in databases constitute an impediment for information extraction and retrieval. Information retrieval in databases is often dependent on the query terms used for the search. Smith [2003] called this the *Database Tower of Babel Problem* which hampers searching for information and putting them together into a larger system as ever more diverse groups are involved in sharing and translating ever more diverse information. Thus, the mapping of terms recognized in text to unambiguous references of abstract concepts or real world entities present in databases or ontologies, helps to aggregate different surface forms [Spasić et al., 2005]. This procedure is called named entity normalization (NEN) or reference resolution. Named entity recognition and normalization taken together can be referred to as concept identification [Cohen and Hunter, 2004, Zweigenbaum, 2008, Fundel, 2007]. It is crucial for the semantic interpretation of recognized terms and the integration of textual data [Jijkoun et al., 2008]. The correct identification of concepts in text is an important prerequisite for information retrieval for instance. Without normalization, different terms identified for the same concept would be treated as distinct items, which thus distorts follow up approaches and statistical analyses. An overview on state-of-the-art methods for Named Entity Recognition and term normalization is provided in Sections 3.1.2 and 3.1.3 of Part I.

1.2.2 Challenges of Chemical Named Entity and Histone Modification Term Recognition

Biological entities which came into the focus of Named Entity Recognition comprise for instance genes, proteins, and protein complexes, mutations and allele variants of genes (e.g. single nucleotide polymorphisms (SNPs)), and disease. In contrast, less efforts were spent on the finding of chemical names in documents. With regard to histone modification no approach has been described in the literature so far that deals with its detection in text.

[17] A polyseme is a word or phrase with multiple, related meanings. A word is judged to be polysemous if it has two senses of the word whose meanings are related.

As domain concept denominations and named entities differ in their characteristics it is fundamental to analyze their characteristics and reveal the main challenges one is faced to. For this reason the terminology utilized for chemical substances as well as histone modification are discussed within the following sections. Furthermore, a short overview on the historical development of chemical naming is given.

1.2.2.1 Overview on Terminology of Chemical Substances

In the early beginnings of chemistry – when it still was called alchemy – peculiar names have been given to investigated substances and compounds, like *'powder of algaroth'* or *'pompholix'* [Crosland, 2004]. With the foundation of modern chemistry a more rational naming system of chemical compounds was developed by de Moreveau, Berthollet, Fourcroy, and Lavoisier in the late 18th century [Kauffman, 1989, Crosland, 2004]. They proposed the first nomenclature, the *Méthode de nomenclature chimique*, and built the basis for some of the chemical names that are still in use, such as *'hydrogen'* or *'sodium chloride'*. Several decades later, in 1921, the first appointed commission for organic, inorganic and biochemical nomenclature – the International Union of Pure and Applied Chemistry (IUPAC) – was founded. The intention was to define a systematic and rational nomenclature with particular emphasis on organic chemistry. IUPAC Definitive Rules were published in 1957 and 1965 [Skolnik, 1976]. They were designed to systematically name a compound reflecting its structure, membership and behavior [Reyle, 2006]. It consists of an extensive rule set recommended for the description of a chemical structure by natural language, so that it is reconstructible. Nevertheless, variants of the nomenclature are also provided by Beilstein (*'Beilstein's Handbuch der Organischen Chemie'*)[18] and Chemical Abstract Service (CAS)[19], leading to variations in systematic names. Making things more complex, authors tend to apply the current rules non-conform, are sloppy or misinterpret the naming principles [Reyle, 2006, Eller, 2006, Brecher, 1999].

Besides, new names are permanently invented, especially a number of brand names for pharmaceuticals which are different for countries in which they are approved. The analgesic *'N-Acetyl-4-aminophenol'* for instance is named as *'Tylenol'* in the United States, *'PARALEN'* in Czech Republic, and known as *'Paracetamol'* in Germany. Names used for a concept establish in a community leading to habituation of people who have been keeping names out of date still in use. Different synonyms and term variants have been distributed for chemical structures, so that it is not unusual to find more than 100 names for one chemical substance (e.g. *'aspirin'*).

The following classification gives an overview on the main naming groups to which chemical terms can be assigned to:

- **Systematic names:** They reflect the information of the chemical structure, its membership and behavior, e.g. *'3-(3,4-dihydroxyphenyl)prop-2-enoic acid'*.
- **Trivial names:** They do not reflect the structure of the substance and cannot be recognized according to rules of any formal nomenclature system. Mostly they have a historical background and were derived from some notable property. Many trivial

[18] http://www.crossfirebeilstein.com/

[19] http://www.cas.org/expertise/index.html

names continue to be used because their systematic names are considered inconvenient for everyday use, e.g. *'caffeic acid'* utilized for *'3-(3,4-dihydroxyphenyl)prop-2-enoic acid'*.

- **Semisystematic names:** In such names at least one part is used in a systematic sense, e.g. in*'N-benzoylglycine'* the part *'benzoyl'* is systematic, whereas *'glycine'* is the trivial name for *'α-aminoacetic acid'*.
- **Registered trademark/brand names:** They are invented for exclusively identifying the brand owner as the commercial source of products, *'aspirin'*.
- **Acronyms and abbreviations:** They are used for convenience to get a short name, like *'L-DOPA'* for *'3,4-dihydroxy-L-phenylalanine'* or *'DTT'* used for *'Dithiothreitol'*.
- **Sum formula:** They consist of the elements contributing to a compound and the number of their occurrence, e.g. '$C_9H_8O_4$'.
- **Ordinary language names:** They mainly have historical origin and denote typical properties or the way of use, e.g. *'table salt'*.

Every chemical naming type described above has its own characteristics and can be linked to certain challenges which have to be considered when aiming at automated identification of chemical terminology in text.

Trivial names can be related to different concepts having diverse meanings. The name *'Bayer'* for instance can stand for the company or the drug *'aspirin'*. Systematic names are often long and have a complex structure and, even though there is a nomenclature recommendation provided by the IUPAC, there are large variations in how it can be and is applied. The two systematic names *'3-(3,4-Dihydroxyphenyl)-2-propenoic acid'* and *'3-(3,4-dihydroxyphenyl)prop-2-enoic acid'* are both used for *'caffeic acid'*. Additionally, the incorporation of whitespaces or its omission as well as different bracket positions within a name are relevant, since different spelling variants might be related to different structures, like *'2-chloro(ethylbenzene)'* and *'(2-chloroethyl)benzene'* [Brecher, 1999] depicted in Figure 1.4.

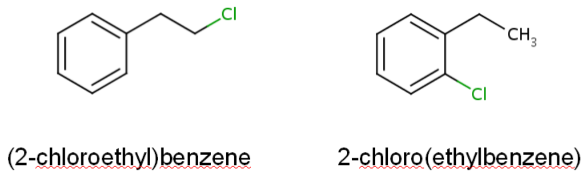


Figure 1.4: Structure depictions of the compounds *'2-chloro(ethylbenzene)'* and *'(2-chloroethyl)benzene'*.

Acronyms and abbreviations provide a challenge because the same combination of characters is often generated for different concepts so that they possess various meanings if used in different domains. Thus, *'DTT'*, introduced above, stands also for *'digital terrestrial television'*. Furthermore, several chemical structures have the same sum formula so that they are not unique, for instance '$C_9H_8O_4$' represents both *'aspirin'* and *'caffeic acid'*. Another source of diversity is a further author's free combination of chemical naming types that can regularly

Chemical name class	Term characteristics and challenges	Suitable NER approach
Systematic names	Spelling variants through different possibilities to apply IUPAC rules and that are partially the result of unexact use of term generation rules, long terms, names can be part of enumerations	Machine learning
Semisystematic names	Combination of systematic with trivial names/abbreviations, authors coin new terms, no rules exist	Machine learning
Trivial names	Limited number of short names	Dictionary-based
Registered trademark/brand names	Limited number of short names	Dictionary-based
Acronyms and abbreviations	Term ambiguity, coinage of new abbreviations without rules	Dictionary-based, Machine learning
Sum formula	Term ambiguity	Dictionary-based
Ordinary language names	Term ambiguity	Dictionary-based

Table 1.3: Overview on chemical name types, their term characteristics and techniques that come into consideration for chemical named entity recognition.

be found in the literature, like *'17-alpha-E'*. Here, an abbreviation and systematic name part are joined in one term and is used for *'17-alpha-Estrogen'*.

Names of chemical elements, substances, and compounds have been collected in freely available and commercial databases, such as DrugBank, HMDB [Wishart et al., 2007], KEGG or the CAS REGISTRY[20], The World Drug Index and CrossFire Beilstein database) respectively that could be used as terminology resource. Furthermore, ontologies like ChEBI and term hierarchies, such as MeSH[21] are a source of chemical terminology. Usually they comprise systematic, trivial, brand and trade names of a chemical as well as chemical formula, some abbreviations or a subset of them.

By virtue of the various term characteristics, the introduced chemical name classes bear different problems and challenges regarding the named entity finding in text. Thus, not every technique will be equally well dedicated. In principle, fundamental term properties which are crucial for NER are term variability, ambiguity, and the coverage of available terminology collections. Table 1.3 provides an overview on the name type classes, related term characteristics, and NER techniques that basically come into consideration.

A general challenge of Chemical Named Entity Recognition and Identification is the fact there is a huge chemical entity space. This can limit the applicability of dictionary-based

[20] http://www.cas.org/expertise/cascontent/registry/index.html

[21] http://www.nlm.nih.gov/mesh/

approaches since they are dependent on the comprehensiveness of terminology collections. However, term disambiguation and mapping to uniq representations (term identification) is straightforward with dictionary-based approaches which provides a clear advantage in comparison to machine learning approaches. In contrast, semisystematic names, terms that are newly coined by authors and terms that do not exactly follow IUPAC rules are expected to be better found by machine learning based methods due to their ability learn general properties of chemical entity's names. A literature survey on chemical entity recognition is provided in Section 3.1.2.1 of Part I.

1.2.2.2 Overview on Designators of Chemical Named Entities

As chemical entities can be assigned to many spelling variants, only the chemical structure is a unique and unambiguous representation of chemical molecules. Computer readable structure formats, such as the MOLfile introduced by Molecular Design Limited (MDL) [Dalby et al., 1992], have been invented that represent the connectivity between atoms of a molecule, charge, stereochemistry, etc. However, for indexing and ensuring uniqueness of molecules in a database designators have been designed with the aim to describe chemical molecules by short ASCII strings. These are Simplified Molecular Input Line Entry Specification (SMILES [Weininger, 1988]), InChI[22], and InChIKey[23], which encode the molecular structure. In contrast, the proprietary, but widely utilized CAS registry numbers[24] comprise no information on the chemical molecule. Since they are potentially usable for named entity normalization, they are described in more detail; the developer, the designator's characteristics, and an example are provided in the following.

CAS

- Developed by Chemical Abstract Service of the American Chemical Society
- Consists of three numbers separated by hyphens
- Encodes no structure information
- Is assigned to chemical entities in the order of admission to the CAS registry
- Example for *'aspirin'*: *50-78-2*

SMILES

- Developed by A. and D. Weininger
- Algorithmically generated, human-readable structure descriptions (short ASCII strings) of chemical molecules
- There are several equally valid SMILES for a molecule:
- Examples for *'aspirin'*:
  CC(=O)Oc1ccccc1C(=O)O,
  CC(=O)Oc1ccccc1C(O)=O,

[22] http://www.iupac.org/web/ins/2000-025-1-800
[23] http://www.inchi.info/inchikey_overview_en.html
[24] http://creationwiki.org/CAS_registry_number

CC(=O)OC1=CC=CC=C1C(=O)O

InChI (International Chemical Identifier)

- Developed by D. Tchekhovskoi, S. Stein and S. Heller at US National Institute of Standards and Technology
- Canonical identifier, algorithmically generated from structure-data files in Mol, SDF or CML format [Dalby et al., 1992, Murray-Rust et al., 1997]
- Chemical structures are expressed in terms of 6 layer types[25] (Example in Figure 1.5):
  - Main layer: Describes the bond connectivity for carbons and hydrogen in a chemical substance,
  - Electronic charge layer,
  - Stereochemical layer,
  - Isotopic layer,
  - Fixed-H layer: Represents one particular tautomer of a given structure. (For the definition of tautomers cf. Appendix A.1.)
  - Reconnected Layer: Allows to represent bonds between metals and carbon in a compound.
- InChI could be extremely long; it is not guaranteed that search engines will read and index them properly
- Example for *'aspirin'*: *InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H, 1H3,(H,11,12)*

InChIKey (hashed InChI)

- Developed by S. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi
- Condensed form of InChI with a fixed length of 27 characters[26]
- InChIKey facilitates web and database indexing and searching
- Example for *'aspirin'*: *BSYNRYMUTXBXSQ-UHFFFAOYSA-N*

Canonical descriptors in general have clear advantages over chemical names for data searching, indexing, and referencing. However, except CAS numbers, they are rarely used in text. This is for several reasons. As CAS numbers have been assigned to chemicals since 1957[27], they are widely applied for referencing chemical compounds in databases. However, CAS numbers do not encode chemical structure information, but are short strings. A drawback of CAS are mistypings and misapplications by incorrectly linking a name to a structure leading to the conclusion that they are not always reliable [Banville, 2006]. SMILES have been invented in 1988 to unambiguously describe the structure of chemical molecules.

[25] `http://wwmm.ch.cam.ac.uk/inchifaq/`

[26] The description was taken from: `http://www.inchi.info/inchikey_overview_en.html`

[27] According to statements from: `http://www.cas.org/expertise/cascontent/registry/regsys.html`

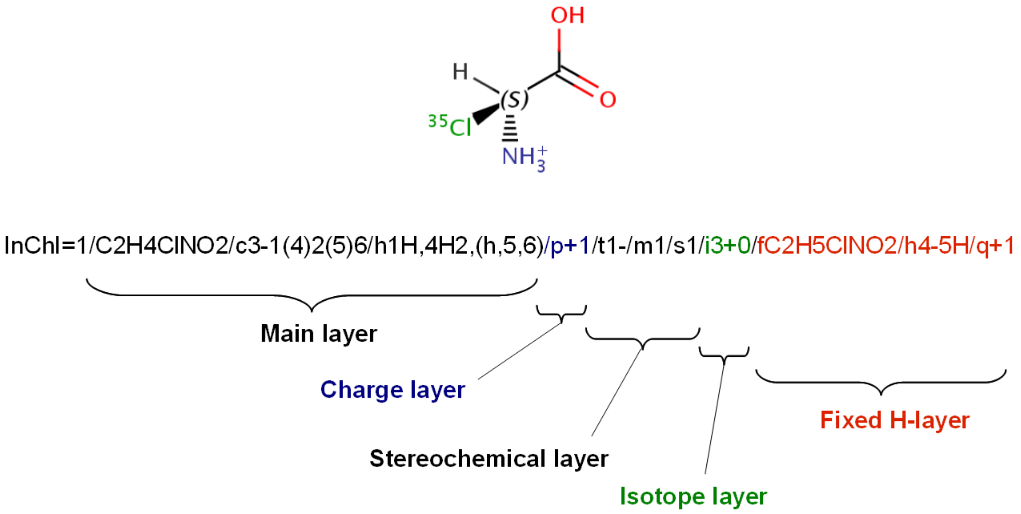


Figure 1.5: Structure and InChI identifier of the compound *'[35Cl]chloro-L-glycinium'*. All available information layers are marked and highlighted within the structure in the respective color. (Figure adapted from McNaught [2006].)

The representation of chemical compounds by InChI has been developed in 2000-2004 and InChIKeys in 2007 and thus are quite young designators. Both encode structure information, but are hardly interpretable by humans, whereas InChI can become quite long, providing a problem for indexing because of inserted line breaks, etc. Nevertheless, they are freely available and meanwhile they have been adopted by many databases as reference.

1.2.2.3 Overview on Histone Modification Terminology

Since 2004 an official nomenclature has been existing for describing histone modifications in text – called the 'Brno nomenclature', which was devised at the first meeting of the Epigenome Network of Excellence[28] [Nightingale et al., 2006]. An example term that corresponds to the official 'Brno nomenclature' firstly published by [Turner, 2005] is *'H3K9me3'*. *'H3'* stands for the protein *'histone 3'*, the letter *'K'* specifies the amino acid lysine and *'9'* its position within the protein sequence. Furthermore, words starting with *'trimethyl'* or *'me3'* explain that the lysine carries three methyl groups as chemical modification.

However, a primary analysis of histone modification terms, showed that the way how histone modifications are denominated in text is quite diverse and the application of the naming recommendation is not common. In general, this is a widespread habit also observable for the nomenclature application devised for other biomedical entities, like Single Nucleotide Polymorphisms (SNPs) or the use of the HUGO nomenclature for genes [Tamames and Valencia, 2006, Klinger et al., 2007]. Some typical examples of histone modifications as they can be found in scientific text are depicted in Table 1.4.

At a glance, the term list shows some of the diverse representation of one specific histone modification. Characteristically, there are histone type and modification descriptions par-

[28] `http://www.epigenome-noe.net/`

Histone modification term variants
H3K9me3 (Remark: Corresponds to the official nomenclature)
Me3-K9 H3
Me(3)-K9 H3
H3K9 tri-methylation
H3-K9 trimethylation
H3 Lys9 trimethylation
H3 tri-methylated at lysine 9
histone H3 trimethylated at lysine (K) 9
K9 trimethylation at histone H3
K9-trimethylated histone H3
tri-methylation of H3 at lysine residues K9
trimethylated H3K9
di- and trimethylated H3K9

Table 1.4: Selection of term variants of one histone modification type as they occur in scientific articles.

tially abbreviated and differing in hyphenation as well as word order. Furthermore, a term can also be part of an enumeration which includes two or even more modification types. An example is shown in the last line of Table1.4.

However, no approach has been developed so far that is able to recognize histone modification terms and to map them to uniq representative terms following the nomenclature.

Chapter 2

Problem Description and Goal

Function annotation of chemical entities and keeping them up-to-date is a challenging task. On one hand side there is a high number of chemical entities whose pharmacological or biological functions, however still incomplete, are stored in databases and classification systems. On the other side new findings and insights on the functions of entities, like descriptions about novel targets of chemical compounds and their interrelations continuously accumulate in published scientific articles, patents, reports, etc. Annotation terms related to chemical entities are an important source for getting a fast overview on their properties, e.g. for finding new uses of a chemical compound, to be able to classify and assign it to a given hierarchy or to predict function annotations of novel non-classified entities. Thus, the extraction of function descriptions on chemical entities from text sources and its inclusion into already available classification schemes and databases enriches and completes the knowledge about them. Nacher and Schwartz [2008] for instance realized that 138 drugs of DrugBank do not contain annotations in form of ATC identifiers and hence could not be integrated into their drug-therapy network study. It leads to the conclusion that pharmaceuticals are missing in ATC or were not annotated in DrugBank, which resulted in an incomplete network. It also makes clear that such studies highly depend on information of resources used, which might limit their value.

Therefore, the aim of this thesis is the development of a framework, which allows for harvesting biomedical and pharmaceutical property information from text and its combination with available structured annotation data. This supports the annotation of chemical entities and the extension of classification systems. The general idea of such a framework is depicted in Figure2.1. It includes two challenges that have to be tackled: the recognition of chemical named entities in natural language data and the extraction of pharmaceutical property information from text which is related to chemical compounds.

As chemical molecules can interfere with epigenetic processes it is important to investigate the influence of the chemical environment onto organisms. Most of this information is still provided in the form of natural language data as scientific literature only, which has been massively increaseing in the last years. However, not much work has been accomplished to make literature easier accessible for the application in epigenomic research. Only two research groups identified literature as valuable information resource to establish databases for epigenomics studies [Fang et al., 2008, Ongenaert et al., 2008]. Both use text mining for collecting information about DNA methylation, affected genes and cancer types. However, they omit to extract information about histone modifications from text, which play a key role in epigenetic mechanisms. No automated approach has been published so far that

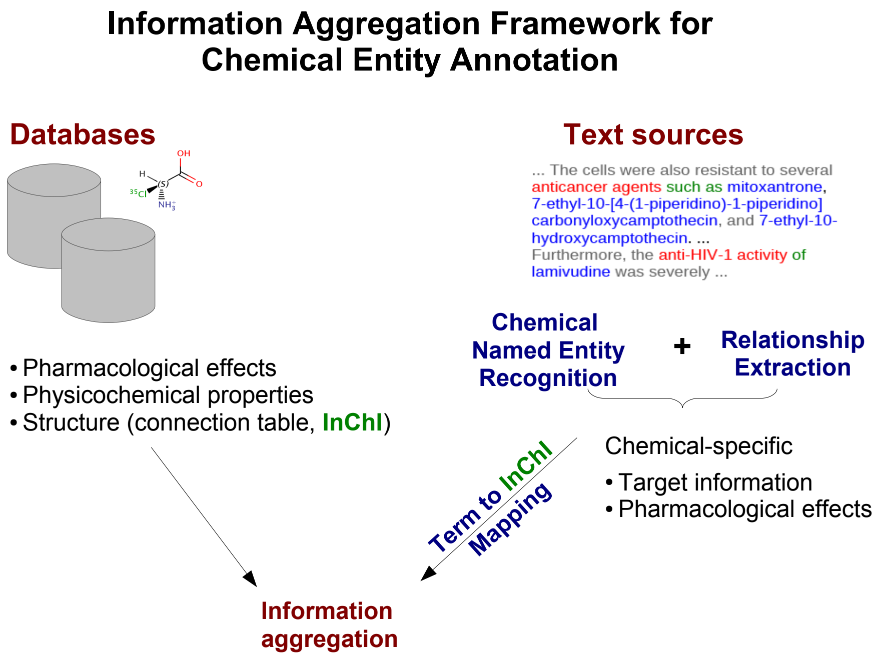

Information Aggregation Framework for
Chemical Entity Annotation
Databases
OH
H
(S)
O
35Cl
NH3+
• Pharmacological effects
• Physicochemical properties
• Structure (connection table, InChI)
Text sources
... The cells were also resistant to several anticancer agents such as mitoxantrone, 7-ethyl-10-[4-(1-piperidino)-1-piperidino] carbonyloxycamptothecin, and 7-ethyl-10-hydroxycamptothecin. ... Furthermore, the anti-HIV-1 activity of lamivudine was severely ...
Chemical Named Entity Recognition
+
Relationship Extraction
Chemical-specific
• Target information
• Pharmacological effects
Term to InChI Mapping
Information aggregation

Figure 2.1: Framework for the aggregation of information on chemical entities.

support the recognition of histone modifications descriptions in text and allow for its identification. To be able to harvest direct or indirect relations between chemical entities and histone modifications from text in the future, the challenge is to find histone modification descriptions in text. Therewith, literature data is made more accessible for information extraction approaches and better available for information retrieval tools.

2.1 Outline of the Thesis

The remaining part of the thesis is divided into two parts. Part I provides the reader with basic background information to understand the developed techniques described in the chapters of Part II. It includes Chapter 3 that gives a general introduction on Natural Language Processing techniques which is a prerequisite for information extraction. Beyond, Named Entity Recognition and term normalization methods on which the work is based on are explained as well as state-of-the-art literature overviews are provided. It briefly describes the generation and annotation of text corpora as well as measures used for evaluating information extraction methods. It is followed by the depiction of techniques utilized for predicting function annotations of entities as well as methods that support the extraction of entity related annotation information from text. Last, text data visualization and available approaches are shortly introduced.

Part II presents the developed methods, the obtained results and discusses its outcome in two separate chapters, which address the two given objectives. Chapter 4 describes the developed information aggregation framework. It is divided into Section 4.1 which investigates an approach for the recognition of chemical named entities. It includes the generation of a test corpus and a dictionary of chemical names integrated into the dictionary-based approach ProMiner. The dictionary processing steps are explained and the obtained results are discussed and compared to other approaches. Section 4.2 covers the work developed for extracting function annotation information, specifically pharmacological property information, on chemicals substances from text. In Section 4.3 the developed information aggregation framework is explained in detail and the obtained results are discussed. It includes two application scenarios that apply the extracted pharmacological concepts for a) the automated support of chemical compound annotation based on textual data and b) the extension of a therapeutic classification system by drug instances. Chapter 5 responds to the task of histone modification recognition in text. It describes the initial extraction of histone modification terminology from text for which the machine learning approach CRF was adapted. Through the newly developed term mapping strategy different modification representations form text were related to defined standard terms. This supported the generation of a histone modification-specific dictionary includable into ProMiner, resulting in a dictionary-based approach. The two NER approaches were evaluated on generated corpora and the results compared and discussed. Additionally, the design of a histone modification hierarchy is presented and several applications of the complete approach are depicted. The final Chapter 6 highlights and discusses the main achievements of this thesis and draws the conclusion for future prospectives.

Part I

Fundamentals

Chapter 3

Fundamentals on Applied and Developed Methods

3.1 Information Extraction Techniques

3.1.1 Overview on Methods applied in IE

Usually, information extraction involves some form of natural language processing (NLP). The following methods are the main components of information extraction systems [Cohen and Hunter, 2004]:

- Sentence Splitting
- Tokenization
- Part-of-Speech Tagging
- Named Entity Recognition and Normalization
- Phrase Chunking and Parsing.

Sentence splitting and tokenization are text preprocessing steps in almost all information extraction approaches. Constituting thereon, Part-of-Speech Tagging is the prerequisite of Phrase Chunking and Parsing and is also used in some approaches for Named Entity Recognition. A general description of these methodologies is given in the following paragraphs. Beyond, techniques applied in this work, i.e. Named Entity Recognition and Normalization are discussed in more detail in a separate Section 3.1.2. Phrase Chunking applied for information extraction is described in more detail in Section 3.2.2.

Sentence Splitting: Sentences are one of the most important elements of natural language. They are the smallest units for expressing completed thoughts or events in written documents. The correct recognition of sentence borders is therefore crucial for many IE approaches [Siefkes and Siniakov, 2005]. The detection of sentence boundaries is not trivial as the punctuation symbol '.' does not always occur at the end of sentences. It is ambiguous, as it often appears within entity names (organisms e.g. *'E.coli'*, proteins e.g. *'M.HsaI'*, chemicals e.g. *'SO4.2Na'*, etc.), abbreviations and decimals whose use is usually domain-dependent.

Approaches have been developed that are based on different methodologies. In general, most of them rely on extensive regular-expressions or hand-crafted rules. However, also machine learning (neural networks and decision trees, Conditional Random Fields) has been used for this task. An overview is given by [Palmer and Hearst, 1997, Katrin Tomanek and Hahn, 2007].

Tokenization: Another basic step in preprocessing text requires segmentation of the character string into its smallest units – tokens, which are words, punctuation marks and separators. This results in a sequence of tokens. In English language, word boundaries are normally indicated by white space. However, in scientific text tokenization typically requires typographical processing at the character level, to handle special characters and white space, upper case and lower case, superscripts and subscripts, and equivalence of Roman, Greek, and numerical suffixes [Siefkes and Siniakov, 2005, Krallinger et al., 2008]. Jiang and Zhai [2007] give a good overview on different tokenization methods used in the biomedical domain.

Part-of-Speech Tagging: Part-of-Speech (POS) indicates the role of words within a sentence. POS tagging is the assignment of a word to a noun, a verb or an adjective for instance. A common set of tags includes around forty categories (e.g. the Penn Treebank tagset[1] [Marcus et al., 1994b]), whereas much larger sets are known. POS tagging is a challenging problem since a word can have multiple parts of speech. The word *'antibiotic'* for instance can be a noun (e.g. in *'development of a new antibiotic'*) or an adjective (in *'antibiotic substance'*) depending on the context of the word.

Systems accomplishing this task are generally based on machine learning algorithms, such as Hidden Markov Models trained on manually POS-labeled corpora [Marcus et al., 1994b]. Some POS taggers also use manually compiled rules to judge ambiguous words; others are entirely rule-based, like the Brill's tagger [Brill, 1992, 1994]. Biomedical literature shows a slightly different POS distribution as compared to general English newswire texts [Cohen and Hunter, 2004, Krallinger et al., 2008]. This has motivated the implementation of specialized taggers optimized for the biomedical domain, such as the MedPost tagger [Smith et al., 2004] or dTagger [Divita et al., 2006].

3.1.2 Named Entity Recognition

3.1.2.1 Literature Survey on Biomedical and Chemical Named Entity Recognition

Gene and protein name recognition has been one of the most active fields in life science text mining since the late 1990th, which lead to a number of advanced applications and a growing number of publications in this scientific area. A good overview is provided by [Krallinger et al., 2008]. The predominant role of gene and protein name recognition in biomedical text mining is also illustrated by the fact that it was a central task of the two BioCreAtIvE assessments in 2003 and 2006 and of the JNLPBA (International Joint Workshop on Natural Language Processing in Biomedicine and its Applications)[2]. The best system applied to protein and gene name finding (gene mention task) at the challenge of BioCreAtIvE II

[1] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQP-HTMLDemo/PennTreebankTS.html

[2] http://research.nii.ac.jp/~collier/workshops/JNLPBA04st.htm

obtained an F_1 measure of 87.21 % [Smith et al., 2008]. Most of the successful systems use machine learning methods or dictionary-based approaches combined with approximate string matching.

Such an independent evaluation does not exist for the recognition of chemical named entities, however, an introduction to systems dealing with the recognition of this entity class in natural language text and their morphological analysis is given in the following paragraphs. Published approaches span from manually developed rule sets, grammar or dictionary-based approaches to machine learning based systems. The systems are differently specialized; some focus on finding only subclasses of chemical names, whereas others intend to recognize the complete bandwidth of chemical names. If available, performance measures are provided which are defined in Section 3.1.5.

Kemp and Lynch [1998] detect nongeneric chemical names in patents using handcrafted rules in combination with dictionaries which comprise chemical name fragments. They claim to correctly find 97.4 % of 14 855 specific chemical names in 70 patents from the International Patent Classification (IPC) class CO 7D. The false positive rate is reported to be 4.2 %.

Narayanaswamy et al. [2003] describe a manually developed set of rules relying upon lexical information, linguistic constraints of the English language and contextual information for the detection of six entity classes. These are proteins/gene, protein part, chemical, chemical name fragment, source and general biological term. The reason for choosing this approach is stated as the lack of an annotated corpus. The evaluation was done on a small hand-selected corpus containing 55 MEDLINE abstracts obtained by searching for acetylates, acetylated and acetylation. They found 158 chemical names from which 22 were ambiguous and classified into different classes and 13 chemical part names with two ambiguous terms. The F_1 measure for the first class is 90.86% (93.15% precision, 86.08% recall). The latter has an F_1 measure of 91.67% (100% precision, 84.62% recall).

Another rule-based system is the ChemFrag annotator of IBM which identifies complex tokens as organic chemical names [Mack et al., 2004]. It combines regular expression rules applied to recognize organic chemical name fragments with rules that assemble these fragments into longer descriptions. The rules are formal expressions, controlling the balance of parentheses, numbers and hyphens. In some of the rules a small dictionary of prefixes and suffixes is used. Mack et al. [2004] describe obtained results of an unpublished evaluation of the ChemFrag annotator on ten annotated patent documents to be 91 % in precision, 94 % in recall, and 92 % in F_1 measure.

The system EbiMed of Rebholz-Schuhmann et al. [2007] recognizes drug names using a drug dictionary compiled from MedlinePlus. However, they do not provide an evaluation of their method.

In parallel to this work Hettne et al. [2009] published a dictionary approach based on Peregrine [Schuemie et al., 2007], which is a string matching method developed for gene and protein name recognition. They use a combination of several terminology resources to obtain a chemical name dictionary. They report on a precision of 51 %, 49 % recall and 50 % F1 measure on the corpus of 100 MEDLINE abstracts published by Kolářik et al. [2008] and described in Section 4.1.1.

Wilbur et al. [1999] employ two approaches – chemical morpheme segments combined with heuristic and a score constituted of three measures and a Bayesian classifier applying character-based n-Grams – for discriminating between chemical names and non-chemical terms of a given term list from the Unified Medical Language System (UMLS) Methathesaurus[3]. They state to label 96.2 % of the chemical names and 97.0 % of the non-chemical terms correctly with the latter method.

Townsend et al. [2005] use a Naïve Bayes method applied to overlapping 4-Grams and context leading to a recall 70.4 %, precision 78.4 % and F_1 measure 74.3 % evaluated on a small corpus of seven articles randomly selected from *Organic and Biomolecular Chemistry 2003* and three documents from *Organic and Biomolecular Chemistry* [Townsend et al., 2004]. A further development lead to the open source program OSCAR3[4] (Open Source Chemistry Analysis Routines) [Corbett and Murray-Rust, 2006], which is the only software available to the academic community. Compared to Townsend et al. [2004] OSCAR3 additionally uses a modified Knesser-Ney smoothing [Chen and Goodman, 1996] to produce a refined 4-Gram model and implies an internal lexicon of chemical names initially populated from ChEBI. Furthermore, a set of rules is applied to group single chemical words for assemble multi-word chemical names. Chemical formulae are recognized with cascaded regular expressions. The system was evaluated on several different topical abstract corpora from MEDLINE. They achieved precision values between 64.1 % and 75.3 % as well as recall values between 69.1 % and 80.8 %.

In the work of Corbett et al. [2007] the toolkit LingPipe[5] was applied that is based on first-order Hidden Markov Models and n-Grams. They combined it with several dictionaries for the identification of chemical entities obtaining 73.5 % in recall, 75.3 % in precision and 74.4 % in F_1 measure.

The program developed by Sun et al. [2007] focuses on finding sum formula like '$CH_3(CH_2)_2OH$' in text using support vector machines and CRFs. Their best result was obtained with Support Vector Machines (SVMs) with recall of 90.36 %, precision of 94.64 % and an F_1 measure of 92.45 % on a randomly selected corpus of 200 chemistry publications from Royal Society of Chemistry.

Unfortunately, there exists no organized international assessment like BioCreAtIvE for the recognition of chemical entities in text. Hence, the comparison of available approaches turned out to be difficult because different corpora of different size have been applied for the evaluation.

A dictionary and a ML based technique have been adopted in this work for the recognition of named entities. Hence, they are explained in more detail in the next sections.

3.1.2.2 Dictionary-based NER Approaches

Available approaches that mainly rely on dictionaries are most often applied for identifying protein and gene names. The performance of dictionary-based methods depends on the one hand on the search algorithm and on the other hand on the quality and completeness of the dictionary. Both components are discussed in the following paragraphs.

[3] www.nlm.nih.gov/research/umls/

[4] http://oscar3-chem.sourceforge.net

[5] http://alias-i.com/lingpipe/

i	*id*$_i$		
1	DB06151	ST_1	InChI=1/C5H9NO3S/c1-3(7)6-4(2-10)5(8)9/h4,10H,2H2,1H3,(H,6,7)(H,8,9)/t4/m0/s1/f/h6,8H CC(=O)NC(CS)C(=O)O
		DBL_1	28939; C06809; D00221; . . .
		S_1	Acetylcysteine; ACC; Mucomyst; Acetadote; Fluimucil; Parvolex; Lysox; Mucolysin; (2R)-2-acetamido-3-sulfanylpro-panoic acid; . . .

2	DB04816	ST_2	InChI=1/C14H8O4/c15-9-5-1-3-7-11(9)14(18)12-8(13(7)17)4-2-6-10(12)16/ h1-6,15-16H
			C1=CC2=C(C(=C1)O)C(=O)C3=C(C2=O)C=CC=C3O

		DBL_2	3682; C10312; . . .
		S_2	Danthron; Altan; Antrapurol; Bancon; Chrysazin; Chrysazine; Criasazin; Danivac; Dantron; Diaquone; Dionone; Dorbane; Dorbanex; Dorbantyl; Duolax; Istan; Istin; Istizin; LTAN; Laxanorm; Laxanthreen; Laxipur; Laxipurin; Modane; Neokutin s; Pastomin; Prugol; Regulin; Roydan; Scatron d; Zwitsalax; 1,8-dihydroxyanthracene-9,10-dione; . . .
⋮	⋮	⋮	⋮

Table 3.1: Example of a dictionary based on DrugBank, usually incorporated in rule based Named Entity Recognition systems.
i: Object number,
id$_i$: Identifiers of central resource (DrugBank in this case),
ST_i: Structural identifiers,
DBL_i: Identifiers of other databases,
S_i: Set of synonyms.

Dictionary A dictionary is a collection of entity names from a certain domain, usually gathered from public domain-specific repositories. Thus, public databases provide a valuable term resource being routinely used by systems aiming at the identification of protein and gene names or disease terms, e.g. [Hanisch et al., 2005, Tsuruoka and ichi Tsujii, 2004, Fundel and Zimmer, 2006, Chun et al., 2006, Sasaki et al., 2008]. In such a dictionary all terms for a given concept are kept together. Usually, they are mapped to unique identifiers like database keys or other unique representations. Table 3.1 provides an example section of a chemical dictionary.

The problem with dictionary-based NER, however, is that dictionaries are seldom complete because of the existence of term variants and new names. The generation and addition of spelling variants is used to overcome this problem partially [Hanisch et al., 2005].

As databases are regularly updated by the curating organization, a dictionary-based NER system relying on these databases has to be automatically updated as well to incorporate new names and symbols. Otherwise it would go out of date.

String Matching Algorithms String matching can be divided into perfect and approximate string matching methods that work on character or token basis [Navarro, 2001]. Perfect string matching performs an exact text search for synonyms from a given term list against text. In comparison, approximate matching allows insertions, deletions or substitutions of single characters or tokens. Usually, similarity and distance metrics are used to score compared strings, e.g. by consideration of several editing operations, differently rated, for transforming one term into another. Cohen et al. [2003] give a good overview on mainly applied distance and similarity metrics.

Wang and Matthews [2008] studied the influence of exact string matching, six different similarity measures and rules onto the protein name recognition. They showed that exact string matching and classical similarity methods performed worse compared to an adapted TF-IDF[6], a measure usually used for information retrieval. Here the token order is not fixed and a fuzzy token comparison lead to acceptable results. An exact string matching method in combination with a prior expanded dictionary containing different spelling variants for identifying protein and gene names was developed by Fundel et al. [2005]. They used a postprocessing filtering via Support Vector Machines and achieved F_1 measures between 79 and 92.1 % at the BioCreAtIvE challenge for protein and gene names of three different organisms. Since approximate string matching allows for a fuzzy term search, most of the available NER approaches utilize this technique. Hence, term variants in text differing from terms in the dictionary to some extend can be detected by the systems. Systems based on approximate string matching have been provided by Tsuruoka and Tsujii [2003], Hanisch et al. [2003], Egorov et al. [2004], and Schuemie et al. [2007]. Tsuruoka and Tsujii [2003] used the edit distance as measure for calculating the similarity between two strings on character level. Arbitrary costs were defined for the individual edit operations that are dependent on the considered letter. To recognize a protein name in a given text, they performed a similarity calculation for every term contained in the dictionary and selected the term that was most similar. The F_1 measure performance on the GENIA corpus was 70.2 %. Hanisch et al. [2003] developed the protein and gene name recognizer ProMiner which also incorporates an approximate search algorithm, but compared to Tsuruoka and Tsujii [2003] it is based on tokens. An evaluation on the BioCreAtIvE I corpus resulted in an F_1 measure between 79 and 89.9 % [Hanisch et al., 2005] depending on the organism studied. Egorov et al. [2004] and Schuemie et al. [2007] developed a quite similar approach. Egorov et al. [2004] obtained 98 % precision and 88 % recall on a randomly selected MEDLINE corpus of 1000 abstracts. Schuemie et al. [2007] achieved 72 % precision and 75 % recall on the BioCreAtIvE II corpus.

As ProMiner was modified to recognize chemical entities in text, it is described in more detail below.

ProMiner ProMiner has been developed by Hanisch et al. [2003] to search multi-word terms in text using a term similarity function and context information. Their goal was to be able to efficiently process large corpora of text for recognizing protein and gene names.

The algorithm is not based on editing operations, but on the assessment of tokens making

[6] Term Frequency Inverse Document Frequency: It is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. In this case a word corresponds to a token and the document collection corresponds to all names in the dictionary.

Algorithm 1 Algorithm of ProMiner (adapted from Hanisch et al. [2003]).

```
Require: $S$ Set of all synonyms in the dictionary
Require: $T$ Set of all Tokens occurring in text
Require: $C = c_1, c_2, \dots, c_n \subseteq S$
Require: $\tau(c), c \in C$
Require: $\sigma(t), t \in T$
Require: $token(c), c \in C$
for each token $t_i \in T$ read from the abstract do
  for each synonym $s_j \in \sigma(t_i)$ do
    if ($s_j \notin C$) then
      $C = C \cup s_j$;
    end if
  end for
  for each candidate $c \in C$ do
    if ($t_i \in \tau(c)$) then
      Update match terms of $s_\alpha(c)$;
      $\tau(c) = \tau(c) \backslash t_i$;
    else
      Update mismatch terms of $s_\alpha(c)$;
      $s_\beta(c) += (1/token(c))$;
    end if
    if ($s_\beta(c) >$ Boundary threshold or Mismatch delimiter found then
      $C = C \backslash c$;
      if ($s_\alpha(c) >$ Acceptance threshold) then
        report $c$;
      end if
    end if
  end for
end for
```

up the candidate term that are conform or non-conform with a tokenized text snippet. In general, tokenized text is provided to the search algorithm which processes it token-wise. By this means, each token t_i is compared to synonym tokens of the dictionary that were generated the same way as text tokens. This comparison procedure results in a set of potential dictionary candidate synonyms $C = c_1, c_2, \dots, c_n$ that are similar to a given text snippet s. They are assessed by scoring functions, whereas two scores are defined for each c_j; a boundary and an acceptance score. They depend on the number of mismatching and matching tokens. Both scores are compared to respectively defined thresholds that are checked each time a token has been processed. They define the term acceptance, stop of term prolongation or rejection of a term. The procedure is shown in the Algorithm 1 below.

Hanisch et al. [2003] observed that tokens are differently important for the recognition of a protein or gene name. Therefore, each token of the synonyms in the dictionary is assigned to one of primarily defined token classes depicted in Table 3.2. This procedure relates a token either with a certain value, fixed for a specific token class or it defines the influence of a token onto the search behavior. Respective token class values define the token contribution to the final score. They were obtained from training examples by parameter optimization via robust linear programming.

ProMiner has been successfully used for the recognition of gene and protein names and obtained high performance results in BioCreAtIvE 1 and 2 [Hanisch et al., 2005, Fluck et al., 2006].

Class Name	Description	Token example	Example term
Modifier	Semantically modifying tokens	*'transporter' 'receptor', 'kinase'*	*'norepinephrine transporter'*
Specifier	Numbers and Greek letters	*'3', 'III', 'alpha'*	*'gold(III) fluoride', '3-carbamoyl- 5-methylisoxazole'*
Delimiter	Separator tokens	( ) . ;	*'6-(octadecylthio)purine'*
Non-descriptive	Annotating tokens	*'fragment'*	*'prothrombin fragment 1'*
Standard	Standard tokens	*'THF'*	

Table 3.2: Description of the main token classes used by ProMiner. Example tokens and terms are provided (adapted from Hanisch et al. [2003]).

Although the algorithm was primarily developed for the recognition of protein and gene names in text, it is generic, so that it can be adjusted do other domains as well. In principle there are four basic components that need to be considered when modifying the system. In the following they are briefly introduced:

- Domain-specific dictionary: It is the most important component on which ProMiner relies. The generation and incorporation of a raw dictionary is the basic step to adjust it to a new domain.

- Curation: Basically, resources like entity specific databases or ontologies used for dictionary generation have not been developed for compiling dictionaries applied by NER tools. Therefore, they lack spelling variants of terms being used by authors in text and hence cannot be found by the matching procedure. An automated expansion of the dictionary is crucial for the performance of the system. Furthermore, resources might not be well curated so that wrong synonyms enter the term list. That is why some synonyms have to be removed. Additionally, the meaning of terms can depend on their case form, like *'his'* that is a pronoun and *'HIS'* which is an abbreviation for the amino acid *'histidine'*. Therefore, some terms have to be marked to be treated during the identification phase in a special way.

- Tokenization: If the segmentation of text and synonyms in the dictionary into tokens is domain relevant, it provides a further possibility for the modification of the system.

- Disambiguation: A single term or name can be associated with several concepts from different domains and hence with different meanings. Occurrences of ambiguous terms need to be resolved to reduce false positive findings. ProMiner handles these cases by analyzing context information.

3.1.2.3 Machine Learning-based NER Approaches

When new entities are discovered (e.g. genes, proteins, diseases, etc.) or created (e.g. drugs), new names are coined as well. Systems that have the potential to find names, that are not contained in repositories and are thus not available to be included into a dictionary, are mostly based on Machine Learning.

In the context of Named Entity Recognition supervised techniques are used which rely on manually labeled training examples. Here, data are often represented as a sequence of tokenized text that is also called the observation or input sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$, $n \in \mathbb{N}$. Each given input sequence $\boldsymbol{x}$ is related to a label sequence $\boldsymbol{y} = (y_1, \ldots, y_n)$ (also called state sequence). Typically, the label sequence is encoded by a label alphabet $\mathcal{L} = \{I, O, B\}$ first introduced by Ramshaw and Marcus [1995]. Label B assigned to a token x_i denotes the beginning of a certain entity mention, $y_i = I$ represents token x_i as inside of it and $y_i = O$ means that x_i is a token which is outside an entity of interest and hence does not belong to it. An example of a token sequence and its corresponding label sequence is provided in Table 3.3.

By this means, the general task in ML-based NER is to find a label sequence $\boldsymbol{y}$ for an observation sequence $\boldsymbol{x}$. A number of different models has been proposed for assigning labels to the sequence of tokens in a sentence. Some models assume independence of the labels in the sequence, like decision trees [Quinlan, 1986] or Support Vector Machines (SVM) [Schölkopf and Smola, 2002]. However, the labels of adjacent tokens are seldom independent of each other in NER tasks. This has led to a number of different models that capture the dependency between the labels of contiguous tokens. These are Hidden Markov Models (HMM) [Rabiner and Juang, 1986], Maximum Entropy Markov Models (MEMM) [McCallum et al., 2000], and Conditional Random Fields (CRFs) [Lafferty et al., 2001, McCallum and Li, 2003], whereas CRFs have recently gained popularity [Cohen and Hunter, 2004, Nadeau et al., 2007, Sarawagi, 2008]. Conditional Random Fields are now established as the state-of-the-art ML methods for named entity recognition. CRFs have shown clear advantages over HMM and MEMM both theoretically and empirically. CRFs provide an advantage over HMMs as they exploit arbitrary feature sets along with the dependency in the labels of neighboring words. This allows them to overcome the independence assumptions made in the other models. Text tokens representing components of the input sequence $\boldsymbol{x}$ are described by several features representing characteristic attributes. An example subset of features used in this work is depicted in Table 3.4. The features have been assigned to different classes which depend on their characteristics and generation method. In CRFs the conditional probability of the label sequence can depend on arbitrary, non-independent features of the

$\boldsymbol{x}$	… for	both	H3	acetylation	and	H3K4	trimethylation	of …
$\boldsymbol{y}$	… O	O	B	I	O	B	I	O …

Table 3.3: Token sequence ($\boldsymbol{x}$) and label sequence ($\boldsymbol{y}$) for an example text snippet containing histone modification descriptions (highlighted in red).

Name	Explanation
Static morphol. features	Regular Expression
All Caps	[A-Z]+
Natural Number	[0-9]+
Alpha-Num	[A-Za-z0-9]+
Init Caps	[A-Z].*
Init Caps Alpha	[A-Z][a-z]*
Real Number	[-0-9]+[.,]+[0-9.,]+
Alpha-Num	[A-Za-z0-9]+
Roman	[ivxdlcm]+ or [IVXDLCM]+
Has Dash	.*-.*
Init Dash	-.*
End Dash	.*-
Punctuation	[„ ,.;:?!−+“ ”]
Autom. generated morphol. features	Autom. generation of a feature for every token:
Prefixes/Suffixes	Match that prefix or suffix
WordsAsClass	Match that token
Context	Is a token preceded or succeeded by:
Spaces	White space
In Brackets	Brackets

Table 3.4: Example features which are used as parameters of the CRF [McDonald et al., 2004, McDonald and Pereira, 2005, Klinger et al., 2008] are ordered by their classes and corresponding feature examples as well as their descriptions are given.

observation sequence, whereas the model does not need to take the distribution of those dependencies into account. In contrast, Maximum entropy Markov models (MEMMs) and other Markov models have a theoretical weakness, the 'label bias' problem [Lafferty et al., 2001]. As linear-chain CRFs were used to recognize histone modification terms in text in this work, they are described in more detail in the following paragraph.

Conditional Random Fields Linear chain Conditional Random Fields are a undirected probabilistic graphical model for computing the conditional probability $p(\boldsymbol{y}|\boldsymbol{x})$ of a possible label sequence $\boldsymbol{y}$ given the input sequence $\boldsymbol{x}$. In this paragraph the description of a special form of Conditional Random Fields – linear chain CRFs – which are appropriate to model sequential data, follows the explanation given in Klinger and Tomanek [2007] and Sutton and Mccallum [2006]. According to Sutton and Mccallum [2006] linear chain CRFs are defined as

$$p_\lambda(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \cdot \prod_{i=1}^{n} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j\Big(y_{i-1}, y_i, \boldsymbol{x}, i\Big)\right) \tag{3.1}$$

where n+1 is the length of the observation sequence, m the number of features, and $\lambda_j \in \mathbb{R}$. The weighting factors λ_j are model parameters and define the contribution of single features f_j to the entire model. The normalization to $[0, 1]$ is given by

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\mathcal{Y}} \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_j f_j\Big(y_{i-1}, y_i, \boldsymbol{x}, i\Big)\right) . \tag{3.2}$$

Here, $\mathcal{Y}$ is the set of all possible label sequences over which is summed up, so that a feasible probability is obtained. In this special case of linear-chain CRF well-known algorithms from the field of Hidden Markov Models like forward-backward propagation can be used to compute the normalization factor [Rabiner and Juang, 1986]. The feature functions f_j given in Equation 3.3 combine features of the considered token with properties of the label sequence. They represent attributes of text tokens in combination with every possible label transition[7]. In a linear chain CRF they have the general form

$$f_j\Big(y_{i-1}, y_i, \boldsymbol{x}, i\Big) = \begin{cases} 1 \text{ if } y_{i-1} \neq O \text{ and} \\ \quad y_i \ \neq O \text{ and} \\ \quad x_i \text{ has feature } m_j \\ 0\,, \end{cases} \tag{3.3}$$

where $i = 1, \ldots, n; n \in \mathbb{N}$ is the label for a token at position i in sequence $\boldsymbol{x}$, $j = 1, \ldots, m; m \in \mathbb{N}$ is the number of features. In case of this work, f_j exhibit Boolean values. This results in a feature vector representation of every token.

In order to get a system that provides specified labels to untagged observation sequences, a model has to be learned on given training data. The goal of model training is to estimate λ_j of the weight vector λ so that the probability of the output label sequence given the training data is maximized. Following Likelihood function given training data $\mathcal{T}$ is maximized:

$$\bar{\mathcal{L}}(\mathcal{T}) = \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{T}} \log p_\lambda(\boldsymbol{y}|\boldsymbol{x})\,, \tag{3.4}$$

which is done via maximum likelyhood estimation. The model training is described in Klinger and Tomanek [2007] and in Wallach [2002]. It can efficiently be performed using hill-climbing methods such as conjugate gradient or limited memory BFGS (L-BFGS) [Sha and Pereira, 2003].

If a model and documents to be labeled are given, the task is the determination of the most likely sequence of states $\boldsymbol{y*}$ for a given observation sequence $\boldsymbol{x}$. This means identifying the label sequence $\boldsymbol{y*}$

$$y* = \operatorname*{argmax}_{y\in\mathcal{Y}} p(\boldsymbol{y}|\boldsymbol{x}) \tag{3.5}$$

[7] In the I,O,B-format like mentioned above for the existence of one entity there are 8 possible label transitions: $B \to B, O \to O, I \to I, I \to O, I \to B, B \to O, B \to I$ and $O \to B$.

that maximizes the conditional probability. The most likely sequence is calculated using Viterbi's algorithm [Rabiner and Juang, 1986], a dynamic programming method.

There are several implementations of CRFs available that can be customized for example by modifying the feature set [Sarawagi, 2008] like MALLET developed by McCallum [2002].

3.1.3 Term Normalization

Term variations and ambiguous terms in text and in databases constitute an impediment for information extraction and retrieval, like query term dependent retrieval results in databases shown in [Kolářik et al., 2007].

Term variations originate from the ability of a natural language to denominate a single concept in a number of ways [Spasić et al., 2005]. Authors generate synonymous names due to e.g. different naming conventions, misspellings or use of acronyms. Jacquemin [1999] determined that one third of term occurrences in an English scientific corpus are term variants. They appear as orthographic, morphologic, syntactic, and lexico-semantic term variations or term abbreviations [Savary and Jacquemin, 2003, Nenadié et al., 2004]. Their description and examples for every type are depicted in Table 3.5.

Term variation type	Explanation	Term variation examples
Orthographic	Usage of hyphens and slashes, lower and upper cases, spelling variations	*H3K9 trimethylation*, *H3K9 tri-methylation*
Morphologic	Inflection (plural, singular forms), derivational transformations, genitive form	*histones*, *histone's*
Syntactic	Prepositional variants, term coordinations	*tri-methylation of H3K9*, *tri-methylation at H3K9*, *H3K9 tri-methylation*, *di- and tri-methylation of H3K9*
Lexico-semantic	Use of synonyms, which are interchangeable	*H3K9 tri-methylation*, *H3Lys9 tri-methylation*
Abbreviation and acronyms	Frequently used in technical sublanguages	*H3K9me3*

Table 3.5: Classification and description of term variation forms. Term variations are marked in red.

To enable the integration of textual data and improved information retrieval, synonyms have to be mapped onto one concept representation and ambiguity needs to be resolved.

Two general methodologies support the mapping of different concept and entity surface representations, which are discussed in the following.

3.1.3.1 Generation of Canonical Term Representatives

The mapping of different term surface forms to one canonical term representative is an option to standardize terms. A canonical term representative is a base form of individual term variations belonging to one meaning. It is obtainable by transformation of a term trough linguistic normalization, e.g. generating a singular form, removing dashes, etc. The utilization of such a standard term as concept and entity representation allows to integrate, compare or map different terminological and textual data from various resources, that can be text, ontologies (cf. its definition in Section 3.2.1), terminologies or databases.

Sarkar et al. [2003] investigated four strategies for their potential to map terms from Gene Ontology to terms of UMLS. They used exact string match, the generation of canonical term forms, the tool MetaMap (MMTx)[8] and Blast-based matching. It was shown that the term canonicalization approach performed best. In a second study Krauthammer and Nenadic [2004] generated canonical term forms of biomedical terms and successfully applied them for mapping different surface realizations belonging to one concept.

A tool which supports canonical term form generation is the Lexical Variant Generating program[9] developed by National Library of Medicine. It is a component of Lexical Tools related to UMLS and provides a series of commands that can be selected and combined to perform lexical transformations of terms [Aronson, 1994]. It involves lexical look-up in the SPECIALIST lexicon of UMLS as well as stripping and replacement functions for example, combined with algorithmic generation of an uninflected term form.

3.1.3.2 Mapping of Terms to Reference Identifiers

Due to synonymy and ambiguity several steps are required after named entity recognition to aggregate different surface term forms and to resolve the correct term meaning. Hence, term disambiguation is often necessary to reject the terms with wrong meanings early in the term identification pipeline, e.g. common English words. Additional filter steps reduce false positive terms. The final mapping to reference resources usually applies dictionaries compiled from these repositories. Similar to NE recognition the procedure has to tackle spelling and term variants, as utilized term collections often do not contain the complete set of synonyms of a given concept. Usually it is accomplished by the addition of spelling varied synonyms and/or application of mapping rules and similarity-based or fuzzy term matching methods.

Some named entity recognition methods imply normalization already in their general pipeline. The fundamental advantage of dictionary-based NE recognition approaches, like ProMiner, is their inherent feature directly allowing to normalize named entities in one step. This means when a term is found in text and disambiguated, it is a simple process to map it to unique identifiers that it represents. In contrast, Machine Learning-based NER approaches do not provide identification information of recognized terms. Hence,

[8] `http://ii-public.nlm.nih.gov/MMTx/docs.shtml`

[9] `http://www.nlm.nih.gov/research/umls/online%20learning/LEX_004.htm`

the entire normalization process has to be accomplished after NE recognition resulting in a two-step process. Although dictionaries are prevalently used in combination with fuzzy string mapping for NE normalization, some ML-based approaches have also been developed and tested.

Ben Wellner [2005] for instance applied CRFs in combination with TF-IDF for character-wise string comparison to map terms from text to UMLS concepts. They obtained an F_1 measure of 73 % while testing their approach on 34,296 lexical entries of UMLS.

Lim et al. [2007] applied a vector space model for finding similar terms in a term list leading to an F_1 measure of 70.7 % at the BioCreAtIvE II gene mention corpus.

The best systems at BioCreAtIvE II obtained F_1 measures of 81 % for the recognition of human proteins and genes and 92 % for yeast [Morgan et al., 2008].

Another tool that maps terms from biomedical text to concept identifiers of UMLS is MetaMap [Aronson, 2001]. It is a program developed at the National Library of Medicine which maps noun phrases identified by automatic term recognition in text to concepts. MetaMap uses a multi-level mapping strategy. First, a target term is analyzed for generating multiple variants, such as acronyms, synonyms and inflectional variants. These derivations of the original term are then mapped against concept names in the UMLS Metathesaurus and are ranked according to a similarity score.

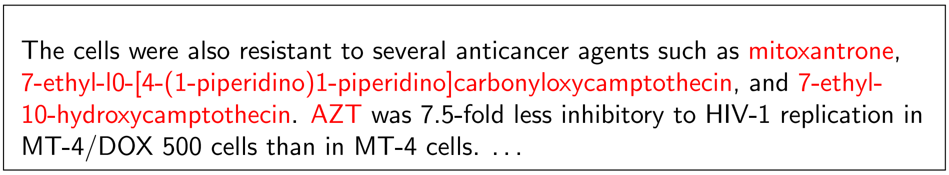
Chemical compounds provide an inherent property of uniqueness by their structure representation. Hence, several approaches have been developed to translate named entities to a structural representation.

One of the first approaches actively used as commercial product is *Name=Struct* from Brecher [1999]. It splits the names into meaningful fragments which are interpreted by a set of rules generating a structure representation from them. He claims to correctly transform 97 % of the parsable names from catalogs of not specified commercial chemical vendors, which were 72 % to 92 %, and 55 % of chemical names of the ChemFinder WebServer.

The system *CHEMorph* [Gerhard Kremer, 2006, Anstein et al., 2006] linguistically analyses systematic names that are based on the IUPAC nomenclature rules as well as on special nomenclature rules for sugar names. It generates SMILES strings from them and determines possible classes of the terms and is based on work of Reyle [2006]. *CHEMorph* was developed to detect synonymous entries as well as errors and inconsistencies in/or between databases. They claim to generate 93 % semantic analyses of 100 arbitrarily chosen names.

Corbett and Murray-Rust [2006] developed the system *OPSIN*, which is an Open Parser for Systematic IUPAC Nomenclature. It assigns chemical structures to complete systematic names by machine interpretation of systematic chemical names. They state to correctly transform 54.7 % of 8183 systematic names from the first 10,000 identifiers of PubChem.

All introduced methods involve the fragmentation of chemical terms and the analysis of the generated name components. Sets of rules are applied to constitute a chemical structure representation. However, the research groups used different terms for the evaluation which makes a final judgment of the systems difficult.



The cells were also resistant to several anticancer agents such as mitoxantrone, 7-ethyl-l0-[4-(1-piperidino)1-piperidino]carbonyloxycamptothecin, and 7-ethyl-10-hydroxycamptothecin. AZT was 7.5-fold less inhibitory to HIV-1 replication in MT-4/DOX 500 cells than in MT-4 cells. ...

Figure 3.1: Text passage example from Wang et al. [2003] in which chemical entities are annotated (highlighted in red).

3.1.4 Corpus Selection and Annotation

The collection and annotation of texts representing a certain language is the basis for information extraction processes. It is required for evaluating the performance of IE approaches and training of ML-based systems developed to recognize Named Entities and more complex IE challenges. Defined annotated corpora – gold standards – allow for system's comparison applied for specific tasks, like the ones announced by MUC, Text REtrieval Conference (TREC)[10] or the BioCreAtIvE challenge (Critical Assessment of Information Extraction in Biology)[11]. Furthermore, the change of an ML system's performance can be analyzed when using different parameters.

A corpus is chosen according to certain criteria meeting several demands that have been defined for specific IE tasks. In general, the text set should be balanced, representative for a certain sublanguage and recoverable. In the annotation process text snippet positions are manually marked with interpretative information, like certain defined entity classes or more complex structures, like relations. Leech [1993] defined several issues and corpus design maxims. Furthermore, he proposed the strategy of storing annotation and raw text separately also known as *Standoff annotation*. Figure 3.1 shows an example text section with annotated chemical entities highlighted in red, whereas Table 3.6 provides the corresponding *Standoff annotation*.

	Standoff annotation	
Term	Position	Entity type
mitoxantrone	69 – 81	Chem. entity
7-ethyl-l0-[4-(1-piperidino)1-piperidino]carbonyloxycamptothecin	83 – 146	Chem. entity
7-ethyl-10-hydroxycamptothecin	153 – 183	Chem. entity
AZT	185 – 188	Chem. entity

Table 3.6: Standoff annotation corresponding to the text passage in Figure 3.1. Chem. entity is the annotation type with which chemical names have been annotated.

[10] http://trec.nist.gov/

[11] http://biocreative.sourceforge.net/

Cohen et al. [2005a] have studied the most prominent corpora designed to promote biomedical text mining and analyzed their properties based on the number of applications that made use of them. They state that the utilization of annotated corpora is dependent on the distribution format and on structural and linguistic annotations. Further requirements are the publication of the annotation guidelines and inter-annotator agreement.

Several approaches have been developed for text annotation, like the open source tool WordFreak developed by Morton and LaCivita [2003] and MMAX[12], a commercial tool developed by EML Research GmbH.

However, the costs of producing annotated text data can be quite high. Especially ML systems require a fairly large number of both positive and negative examples for system training and testing.

The annotation of corpora is a tedious, time-consuming, and expensive task. Furthermore, as Sarawagi [2008] outlined, statistical learning techniques crucially depend on the training data being representative of the distribution on which the trained model is deployed. In general, training examples contribute statistical data to a learner, which in turn estimates several parameter values. The reduction of the amount of data to be labeled for learning and the support of choosing complying examples is of eminent value for the annotation process.

3.1.4.1 Active Learning

Active learning supports the limitation of human annotation effort by sample selection with the aim of obtaining a high system's performance. It is based on the assumption that labeled examples are not equally informative or equally easy to label. An informative example is one whose contribution to the statistics leads to a significant improvement of model parameter estimates. Engelson and Dagan [1996] state that this avoids redundant annotation of many examples that contribute roughly the same information to the learner.

Active learning is an iterative process which is composed of three main phases that are repeated until a stopping criterion is reached. A learning program examines many unlabeled examples and selects only those for labeling that are most informative for the learner at each stage of training. The phases consist of training, selective sampling and human annotation, described in Algorithm 2. The setting typically consists of a small set of labeled examples L and a large set of unlabeled examples U.

The stopping criterion can either be the number of iterations or a desired performance measure, whereas the model performance is evaluated on a test set in each iteration.

Various active learning algorithms have been developed, mainly differing in the method of assessing the informativity of new potential training instances. The two most popular active learning methods used in NLP are uncertainty-based sampling [Cohn et al., 1994] and query by committee [Seung et al., 1992]. In uncertainty-based learning, new instances are selected for annotation on which the classifier is least certain of their classification. The assumption is that instances which are harder to classify are more useful for training. In case of probabilistic models uncertainty of the classifier is commonly estimated using the entropy. For non-probabilistic ones, the classification margin is used, as in the case of support vector machines [Vlachos, 2008]. In query by committee, a body of classifiers is trained

[12] http://www.eml-research.de/english/research/nlp/download/mmax.php

Algorithm 2 Active Learning process (adapted from Vlachos [2008]).

```
Require: Small set of labeled data $L$
Require: Large set of unlabeled data $U$
Require: Model $M$ trained on $L$
repeat
    Apply the trained model classifier $M$ on $U$
    Rank the instances of $U$ according to a performance measure
    Manually annotate the top $b$ instances of $U$ and add them to $L$
    Train the model on the expanded $L$
until a stopping criterion is satisfied.
return a model $M$
```

on L and subsequently applied to the instances of U. Instances for which the classifiers yield the highest disagreement are considered to be the most informative. Common ways of estimating the disagreement are the vote-entropy [Argamon-Engelson and Dagan, 1999] and the Kullback-Leibler divergence [Pereira, 1993].

Active Learning has been applied to several problems in NLP, such as document classification, POS tagging, chunking, statistical parsing, and information extraction [Tomanek et al., 2007].

Active learning based on uncertainty sampling was applied in this work to extend an initially generated training corpus for the recognition of histone modifications in text. Similarly as in [Tomanek and Hahn, 2009] it was utilized to improve a primary learned CRF model. Therefore, an initial CRF system was applied as base learner to a large set of unlabeled data U, which was in this case MEDLINE. The obtained conditional probability $p(\boldsymbol{y}^*|\boldsymbol{x})$ of the most likely label sequence $\boldsymbol{y}^*$ (c.f. Equation 3.5) for an observation sequence $\boldsymbol{x}$ was utilized to determine the uncertainty q of the system, which is calculated by:

$$q = 1 - p(\boldsymbol{y}^*|\boldsymbol{x}) \tag{3.6}$$

The unlabeled data were ranked according to q. Those with the highest uncertainty values q were chosen for annotation and the extension of the training corpus. The convergence of the F_1 measure, an evaluation measure defined in the following section, was employed as stopping criterion.

3.1.5 Evaluation Measures

Different systems which provide the same solution for one task need to be compared with each other to ascertain the best suited one. Furthermore, an evaluation has to be conducted for measuring whether certain changes in a system have lead to an improvement in its performance. Hence, the output of Named Entity Recognition or information extraction systems is evaluated according to a gold standard which defines relevant and non relevant objects in a text corpus.

In the 80th van Rijsbergen [1979] developed a quality measure, the F_1 measure, for the evaluation of information retrieval approaches which is also used for NER and other information extraction tasks. It is predicated on following measures of the output comparison with the gold standard. Documents or entities that are relevant and are successfully found by the system are thus 'true positive' outputs. Documents/entities that are retrieved by the system but that actually are not contained in the gold standard are 'false positive' outputs. Instead, documents/entities that truly are relevant but that the system failed to identify are 'false negative' outputs. Table 3.7 gives another overview on the basic measures.

		Annotated data	
		Positive	Negative
System output	Positive	True positives (TP)	False positives (FP)
	Negative	False negative (FN)	True negatives (TN)

Table 3.7: Description of basic measures collected for the evaluation of IE approaches. (The figure was adapted from van Rijsbergen [1979].)

These basic measures are used for the calculation of precision and recall which are defined in the following paragraphs.

Precision is a measure of exactness and is calculated by dividing the number of true positives by the total number of outputs, which is the sum of true positives and false positives. The equation is given as:

$$Precision = \frac{TP}{TP + FP} \quad . \tag{3.7}$$

The recall is a measure of completeness and is calculated by taking the ratio of true positives by the total number of potential correct outputs, which is the sum of true positives and false negatives. The equation is given as:

$$Recall = \frac{TP}{TP + FN} \quad . \tag{3.8}$$

Recall and precision are usually considered conjoined whereas an optimal system obtains results with high values in both. However, often approaches can reach a high performance only in one measure, because rising or optimizing one is associated with the cost of decreasing the other measure. Which measure is aimed to maximize depends on the requirement to be fulfilled by the system. For instance on one hand side it could be important to develop an approach with high precision, accepting that some information is missed. On the other hand high recall rates could be more important than an optimal precision, even though there might be a higher number of false positive findings. An example scenario is a search for a certain topic by a company, whereas missing information could be highly relevant for product development or market monitoring and thus finally for its finances.

However, users of information extraction approaches are often interested in systems that identify a high number of available entities or relations with a low false positive rate. Thus, the weighted harmonic mean between recall and precision, the F_1 measure, is applied for the method's evaluation, which is defined by:

$$F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad where\ \beta = 1. \tag{3.9}$$

The weighting factor $\beta \geq 0$ can be used to shift the weight towards precision or recall, whereas F_β is balanced in both for β = 1.

Another measure – the accuracy – represents the ratio of correct outputs to the total number of cases evaluated. It measures the fraction of correct answers, i.e. true positives and true negatives, with respect to the total number of test cases:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad . \tag{3.10}$$

Accuracy is often used to evaluate classifier predictions [Kotsiantis, 2007]. However, there is the accuracy paradox indicating that classifiers or systems may reach a high accuracy, but show at the same time only a low precision. Hence, to evaluate information extraction system's the metrics precision, recall, and F_1 measure should be favored.

3.2 Function Annotation of Entities

The subsequent sections deal with techniques addressing the challenges and problems of entity annotation and are organized as follows: At first, the characteristics of ontologies and their advantage for utilization in annotation is explained in Section 3.2.1. It is followed by Section 3.2.2, which introduces fundamental methodologies and approaches that support the identification of entity annotations and new property information in text resources utilizable as automated annotation collection and as basis for the definition of new annotation classes. Methods that make use of already annotated chemical compounds for propagating or predicting certain function annotations/properties to non-annotated chemical entities are outlined in Section 3.2.3.

3.2.1 Impact of Ontologies for Function Annotation and Data Management

Basically, Ontology has been a branch of philosophy since about 2000 years. It is the science that describes the reality by a classification of entities and deals with relations which hold between entities belonging to a certain domain and distinct domains of science [Smith, 2003]. Biemann [2005] gives a nice historical overview on ontology in philosophy. In the context of information science ontologies provide systems with a re-usable machine-readable definition of relevant information by representing the domain knowledge in a formal way [B. Yildiz, 2007, Mizoguchi and Ikeda, 1996]. Guarino [1998] describes an ontology prevalently used in Artificial Intelligence as:

> "an engineering artefact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words. . . . In the simplest case, an ontology describes a hierarchy of concepts related by subsumption relationships; in more sophisticated cases, suitable axioms are added in order to express other relationships between concepts and to constrain their intended interpretation."

This is in contrast to the philosophical sense of an ontology, which is a particular system of categories accounting for a certain vision of the world, independently on a particular language [Guarino, 1998].

Ontologies applied in the context of information science serve as metadata schemas, providing a controlled vocabulary of concepts, each with explicitly defined and machine-processable semantics. They are crucial for the engineering, management, organization and representation of knowledge, the modeling, integration, retrieval and extraction of information as well as for database design, language engineering, agent-based systems design for example, and object-oriented analysis [Guarino, 1998, Silvescu et al., 2001].

The main components of an ontology are concepts, their definitions, relations, and axioms. The combination of an ontology with associated instances, that are 'things' represented by a concept, is a knowledge base [Stevens et al., 2000].

The use of a controlled vocabulary from an ontology supports annotation processing by machines for their automated comparison, analysis, and propagation as well as improved information sharing when applied in different data sources. The GOA project aiming at uniformly annotating proteins of species in the focus of research with Gene Ontology terms[13] is a prominent example [Camon et al., 2003]. Furthermore, the hierarchical organization of ontologies predestine their application in data retrieval systems. It allows for semantic search of information at different levels of granularity. For example a query of a bibliographic or fact database with the chemical family concept *'azides'*, returns all entries containing compounds that belong to this chemical family. Beyond, ontologies provide defined concept denominations, that, when used for indexing and search, avoids the problem of search term-dependent retrieval results.

Thus, ontologies obtained increasing attention in the computer science community within the last two decades, especially in Bioinformatics, Artificial Intelligence, Computational Linguistics, and Database Theory.

3.2.2 Information Extraction for Supporting Function Annotation of Entities

Entity classification systems, ontologies or free text fields of repositories provide function annotations of entities. The latter ones are for instance the *Function* field from the protein database Swiss-Prot[14] [Bairoch and Apweiler, 1997] and *Pharmacology* as well as *Mechanism of Action* embodying pharmacological effects of chemical entities from DrugBank.

Usually, the excerption of new entity annotations is often done manually by database curators. Therefore they read high amounts of articles to fill in predefined forms with annotation information of interest. Though, as Baumgartner et al. [2007] state, manual work will not be sufficient to annotate huge numbers of biologically entities. Although they analyzed the annotation coverage of entities in genomic and protein databases and predicted its prospective development, the problem is a general one. Thus, textual data represent a bottleneck for entity annotation because of the data overload. From this follows that automated support is needed to assist the annotation process for instance by selecting relevant documents and paragraphs or by finding new entity annotations in text. This can basically be aided through Information Extraction methodologies.

The following sections describe the impact of ontologies for Information Extraction on one hand and the role of Information Extraction techniques for the generation of ontologies on the other hand.

[13] www.geneontology.org

[14] `http://www.expasy.ch/sprot/`

3.2.2.1 Impact of Ontology for Information Extraction

Ontologies attained strong attention for Information Extraction. Systems relying on them are generally divided into ontology-based and ontology-driven information extraction approaches [Spasić et al., 2005].

In the first case they serve as a source of vocabulary for natural language processing systems, e.g. GoPubMed [Doms and Schroeder, 2005]. Although collecting names is not the function of ontologies, most of the biomedical and chemical ones provide lists of names for entities they comprise. However, their application for Named Entity Recognition is limited, because concept denominations provided by ontologies often differ from terms available in text. Furthermore, they are often long multi-word names which differ from terms available in text. This is especially the case for Gene Ontology making it difficult to recognize its terminology in text [Blaschke et al., 2005].

In ontology-driven information extraction systems ontologies are actively used to guide the analysis of textual data for extracting factual knowledge to instantiate one or several predefined extraction forms [Guarino, 1998, Spasić et al., 2005, Nedellec and Nazarenko, 2006]. The forms to fill represent parts of the ontology, e.g. concepts and relationships that model a gene regulation network in which proteins interact positively or negatively with genes [Nedellec and Nazarenko, 2006]. An example IE approach based on ontologies is OpenDMAP [Hunter et al., 2008] which was constructed for harnessing instances of protein transport events, protein-protein interactions and the expression of a gene in a particular cell type.

The assembly of adequate ontologies for certain domains or subdomains or the extension of existing ones is a bottleneck for many applications. Especially the idea of the Semantic Web revealed the development and application need of ontologies to automatedly support information structuring and preparation for knowledge representation of this fast growing information resource [Maedche and Staab, 2001].

3.2.2.2 Ontology Learning

The manual design of ontologies by knowledge engineers and domain experts is time and labor intensive. The process spans from knowledge acquisition and identification of the domain's key concepts to its encoding in some formal language, like the obo or owl format [Stevens et al., 2000]. To support ontology construction and speed up the process, the definition of new concepts and relations to be integrated into an ontology has to be assisted or the finding of instances which correlate to concepts. Information Extraction from text is a promising approach for knowledge acquisition [Nedellec and Nazarenko, 2006]. It is based on the fact that natural language texts from specific domains comprise the domain's concepts and their relations.

Methods developed for learning ontologies and its instantiation from large natural language resources rely on the identification of concept descriptions and instances, i.e. noun phrases or named entities of a given domain. The subsequent analysis of the concept distribution or their co-occurrence in combination with clustering methods as well as the utilization of relation extraction methods support the generation of a structured concept output [Nedellec and Nazarenko, 2006, Biemann, 2005]. Mizoguchi and Ikeda [1996] and Biemann [2005] give an overview on these techniques and methods.

However, as Biemann [2005] claims, learning ontologies completely from text instead of manually creating them is problematic. He summarizes that none of the automated methods used today are good enough for creating semantic resources of any kind in a completely unsupervised fashion. Furthermore, a text corpus is biased and might only reflect a limited concept and relation space of a certain domain [Nedellec and Nazarenko, 2006], so that important parts for the ontology application could be missing. Thus, automated methods, like IE techniques applied to extract concept relationships in text as well as instances, can only support ontology construction.

Information Extraction combined with Named Entity Recognition can meet the challenge of finding annotation information on entities in large amounts of text. Hence, a fundamental technique underlying this work that has been employed by several groups for term extraction and ontology learning is explained in the subsequent section.

Extraction of Hypernymic Phrases Predominant semantic relations in ontologies are the taxonomic relationship *is-a* of hypernym/hyponym pairs and meronymy, the *part-of* relation. The first one can be interpreted either as an instance-class relation or as a generalization relation between two classes, whereas the latter one is only a relation between classes.

Hearst [1992] found out, that many text genres frequently contain phrases describing taxonomic relationships between noun phrases. In general, phrases following this constitution, relate two or more noun phrases, some semantically specific (hyponym) and others more general (hypernym) by a taxonomic relationship. An example for such a construct is *'Adinazolam is a benzodiazepine derivative'*, whereas *'Adinazolam'* constitutes the hyponym and *'benzodiazepine derivative'* the general term – the hypernym. Fiszman et al. [2003b] also demonstrated in a study that the coverage of hypernymic propositions in biomedical text, like MEDLINE, is promising. The extraction of such relations would support ontology extension and its population with instances as well as entity annotation.

The identification of hypernymic phrases in text is a multi-level process. A prerequisite before extracting complete hypernymic phrases from text is the recognition of concept descriptions and instances, which are the basic information carriers: Named Entities and terminological units that are noun phrases [Bourigault, 1992, Siefkes and Siniakov, 2005]. The noun phrase definition can be found in Manning and Schütze [1999]; examples are provided in Figure 3.2. As NER was discussed in previous sections, methods for noun phrase recognition and extraction are introduced at this point.

According to NER the general task of recognizing the major sentence constituents like noun phrases, is the finding of their boundaries. This is achieved by partial decomposition of the sentence structure through a local analysis of sentence fragments (also called chunks) named as shallow parsing or alternatively chunking, which is in contrast to full parsing dealing with the structure analysis of an entire sentence.

To accomplish this task, several techniques have been developed to find non-recursive noun phrases in text spanning from machine learning to pattern- or rule-based approaches. ML approaches which have been specifically set up for base noun phrase recognition are

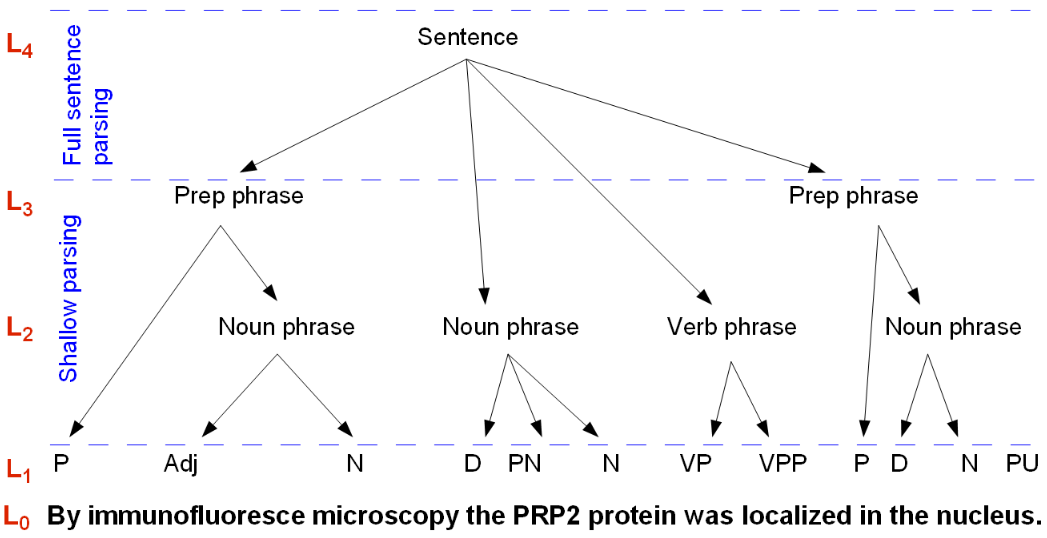


Figure 3.2: Example of a parsing tree with the different indicated levels L_i. L_1 depicts part-of-speech and $L_2 - L_4$ provide the different parsing levels. The Figure was adapted from Craven and Kumlien [1999].
D: determiner, N: noun, PN: proper noun, Adj: adjective, P: preposition, VP: past tense of be, VPP: verb past participle, PU: punctuation

based on Support Vector Machines, like the YAMCHA tagger [Kudoh and Matsumoto, 2000], Conditional Random Fields [Sha and Pereira, 2003], MEMMs [McCallum et al., 2000, Sha and Pereira, 2003] or on transformation-based learning [Brill, 1994], like the tagger TBL from Ramshaw and Marcus [1995]. Usually, they were trained on annotated newswire corpora, e.g. provided by the Conference on Computational Natural Language Learning (CoNLL[15]) or the Penn Treebank corpus of Wall Street Journal text [Marcus et al., 1994a]. They induce statistical models from lexical and token feature information as well as from automatically assigned part-of-speech classes or from one of both [Ramshaw and Marcus, 1995, Wermter et al., 2005]. They reached F_1 measures between 86-94 % dependent on the utilized corpus.

Rule- or pattern-based methods utilize sets of syntactic patterns defined by regular expressions, regular grammars or rule sets [Zweigenbaum, 2008]. For this, text has to undergo a syntactic analysis process in order to determine part-of-speech tag sequences corresponding to the given token sequences.

Dagan and Church [1994] and Justeson and Katz [1995] developed syntactic patterns defined by regular expressions which depends on documents tagged with part-of-speech, though they do not report on performance. A further approach for identifying concept denominations in text which relies on patterns was developed by Frantzi [1997]. She focused on defining a measure to identify candidate terms from domain texts in general and does not provide the performance of the system. Another noun phrase chunker – *Analytics* –

[15] http://www.cnts.ua.ac.be/conll2000/

developed by TEMIS[16] is based on a set of hand-crafted finite-state grammar rules and extracts nouns, proper names, and noun phrases. It has been shown in Wermter et al. [2005] that it obtains a good F_1 measure performance of $\sim$ 91 % in recognizing base noun phrases in biomedical domain texts (Genia corpus) which is comparable to machine learning based methods, like YAMCHA or TBL ($\sim$ 89 % and $\sim$ 86 % respectively).

To identify hypernymic relations completely, the simplest method would be to extract and statistically analyze co-occurring entities or concepts in text. It relies on the hypothesis that entities which are repeatedly mentioned together are somehow related. Co-occurrence can provide a huge amount of related entities, but embodies no information about the quality and direction of an existing relation. Additionally, approaches based on this method may identify a high number of false positive relations as well, i.e. relations of another type or unrelated concepts and entities, like shown in the following example: *'... chemical A is not a protein B activator ...'*. Hence, more sophisticated methods should be used to extract specifically related concepts like rule- and pattern-based extraction. The underlying basis of this approach is that sentences or phrases conforming exactly to a pattern or a rule, express the predefined relationship(s) between the sentence entities. Skusa et al. [2005] give a good overview on relation extraction methods. Basic linguistic structures relevant for the recognition of taxonomic relationships is described in the following paragraph.

Hearst [1992] identified a set of lexico-syntactic patterns that are easily recognizable, occur frequently in text and across genre boundaries. That is why hypernymic propositions are also called Hearst phrases or patterns.

In general, three syntactic structures encode hypernymic propositions. These are phrases which consist of noun phrases connected by verbs. Furthermore, there are appositive structures, and nominal modifications of nouns [Cimiano et al., 2005, Hearst, 1992, Rindflesch and Fiszman, 2003]. Some patterns for each class are introduced within the following paragraph, whereas NP_i represent noun phrases. $NP_{1,...,n}$ correspond to the hypernym of NP_0, whereas their relationship is reflexive and transitive, but not symmetric.

In propositions involving verbs the most frequent occurring verb is a form of *'be'*.

- Propositions involving verbs:

  NP_1 is (a | an) NP_0

  NP_1, NP_2, ..., and NP_n are NP_0

The example hypernymic proposition *'Adinazolam is a benzodiazepine derivative'* matches the first pattern. *'Adinazolam'* corresponds to NP_1 and is a pharmaceutical. The term *'benzodiazepine derivative'* complies with NP_0.

In appositive structures, two noun phrases must be contiguous conjointed by commas, parentheses, or lexical items, like *'including'*, *'such as'*, and *'especially'*.

- Proposition of appositive structure:

  NP_0 such as NP_1, NP_2, ..., NP_{n1} (and | or) NP_n

  NP_0 (including | especially | like) NP_1

[16] http://www.temis.com/

$NP_1, NP_2, \ldots, NP_n$ (and | or) other NP_0

In nominal modifications the two taxonomically related concepts of the hypernymic proposition are represented in a single noun phrase.

- Nominal modification:

  $NP_0 NP_1$

  $NP_1 NP_0$

In these cases, the head noun may represent either the hypernym or the hyponym, while the modifying noun represents the other [Fiszman et al., 2003b].

Several groups concerned with the recognition of hypernymic propositions in text pursuing different goals and interests. Approaches that incorporate such patterns for the extraction of Hearst phrases from are introduced in the following section.

Literature Overview on Hearst Phrase Extraction Hearst [1992] applied pattern-based relation recognition to general texts to find terms and expressions that are not defined in Machine Readable Dictionaries. Since the set of entries within such a dictionary is fixed, the use of text for building up large lexicons for natural language processing was considered advantageous. Her intention was to automatically acquire hyponymic lexical relationships between two or more noun phrases from unrestricted, domain independent text. Finally, the extracted hierarchical related terms were compared with WordNet[17] [Miller, 1990], a hand-built lexical thesaurus. It was shown that the approach achieves promising results for augmenting and verifying existing lexicons like this.

Fiszman et al. [2003b] developed an approach called SemRep to interpret hypernymic propositions in MEDLINE articles. Their approach is composed of a two-step process. First they identify syntactic structures that potentially encode hypernymic propositions with a module named SemSpec (Semantic Specification). Subsequently, identified syntactic arguments are matched by MetaMap to concepts in the Metathesaurus of UMLS. These concepts are then subjected to semantic validation, whereas the system includes the semantic groups Disorders, Procedures and Chemicals and Drugs. They report on 46 % recall and 78 % precision for the identification of treatment propositions. However, no application of SemRep was extensively discussed in their work.

Cimiano et al. [2005] extracted hypernymic phrases by matching regular expressions over part-of-speech tags. They intended to induce concept hierarchies from text collections, WordNet and the World Wide Web by applying an agglomerative hierarchical clustering algorithm which is guided by hypernymic/hyponymic term pairs. A manual evaluation of the learned taxonomic relations within the hierarchy revealed a precision of 65.66 % which is a better result compared to the one that was obtained by a pure clustering method tested.

Although the introduced works were developed with different basic intentions, they show the value of extracting hypernymic propositions from text and their subsequent use

[17] `http://wordnet.princeton.edu/`

in diverse fields. They laid the basis for developing an approach to extract chemical entity related information from text.

3.2.3 Function Prediction of Chemical Entities

The prediction of properties like the pharmacological activity or the toxicity of chemical compounds is an important technique applied in computer-aided drug design. It is for instance utilized to virtually select a subset of a chemical compound library that is tested in wet-lab experiments. This can reduce costs and increase hit rates for lead discovery.

Function prediction has been the main subject of structure-activity relationship studies, which tries to correlate molecular structure to biological properties, pharmacological activity or toxic effects [Wilson, 1982, Winkler, 2002, Selassie, 2008]. It aims to find models derived from training data which can be used to predict respective properties of new compounds. Another actively utilized method is ligand-based virtual screening [Bajorath, 2002, 2001]. This method is utilized for finding molecules which have similar or better activities compared to compounds with known biological or pharmacological effects. Here structural similarity between molecules is utilized for the virtual search of a large compound collection in a database and to rank molecules according to its structural overlap.

Both SAR and virtual screening support the finding of new functions for compounds with unknown biological activity toxicity. They rely on molecule's inherent properties, i.e. the structure and physico-chemical property information. Therefore chemical compounds are represented by descriptors, so called molecular fingerprints. They capture a broad range of molecular characteristics like physico-chemical properties and chemical structure. They typically encode 2-dimensional (2D) and/or 3-dimensional (3D) features of fragments or the complete molecule as a vector of binary values. The variables indicate the presence or absence of certain substructures, topological properties, etc. of a molecule. Examples are the publicly available MACCS keys [McGregor and Pallai, 1997], the BCI fingerprint [Barnard and Downs, 1997], the Daylight fingerprint or 3D pharmacophore fingerprints [Brown and Martin, 1996].

Several similarity metrics, statistical methods, and supervised machine learning algorithms have been developed to comply with this task. These are for instance regression analysis, logic-based classifiers, perceptron-based techniques, statistical learning algorithms, instance-based learning or Support Vector Machines. [Sen and Srivastava, 1992, Kotsiantis, 2007, Selassie, 2008] give a good overview on their basic principles, the main algorithms as well as their advances, and problems.

In the case of virtual screening one of the most prominent similarity measure used for comparing the structural similarity between two molecules is the Tanimoto coefficient Tc [Willett et al., 1998, Reddy et al., 2007]. Therefore two molecules A and B are represented by vectors $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$, whereas the variables a_n and b_n denote the binary values set 'on' for respective features. The Tanimoto coefficient is defined by the following equation:

$$Tc(A,B) = \frac{\sum_{i=0}^{n} a_i b_i}{\sum_{i=0}^{n} a_i^2 + \sum_{i=0}^{n} b_i^2 - \sum_{i=0}^{n} a_i b_i} \tag{3.11}$$

The expectation in ligand-based virtual screening is that compounds exhibiting a similarity above a certain threshold of Tc possess a certain biological activity of the reference molecules. In general a cutoff of $\geqq$ 0.85 Tanimoto similarity is used as rule of thumb, because it is expected that 80 % of these compounds show the required biological activity [Patterson et al., 1996]. However, studies e.g. by Martin et al. [2002], Bajorath [2001] revealed that this is not a binding constraint.

Models obtained by regression analysis which is often used for SAR studies are validated through several measures, the correlation coefficient r, the standard deviation, Leave-One-Out cross validation q^2, and the $F\ value$ (for its definition cf. [Selassie, 2008]). The first three ones provide a quality measures of the fit of the model and constitute the variance in the data, whereas the F-value is used as statistical significance of the regression. Models exhibiting r and $q^2 \geqq 0.90$ as well as high values of $F\ value$ are usually considered to be related to a high predictive power. However, as Kubinyi [2004] and Golbraikh and Tropsha [2002] revealed, a model validated as highly predictive on the training set is not necessarily related to a high predictivity on testing sets.

The main factor relevant for both SAR and ligand-based virtual screening is the chosen molecule descriptor. It influences the performance and hence needs to contain adequate information that is relevant for the given problem [Winkler, 2002, Bajorath, 2001].

As a new descriptor for predicting classes of the drug classification schemes like ATC was developed in this work, the literature was searched for a method with which it could be compared with. The found available approach accomplishing this task is described in the following section.

3.2.3.1 Related Work – Class Prediction of Chemical Compounds

A method – called SuperPred – that predicts pharmacological function as class labels of the ATC classification scheme for compounds has been published by Dunkel et al. [2008]. It is based on chemical structure similarity and is described in more detail, because it was used for comparison with the developed approach explained in (cf. Chapter 4.2). SuperPred applies SMILES string representations of chemical structures as input and generates a ranked list of ATC classes for a given chemical compound. It relies on a basic dataset of 6300 compounds that are related to ATC class annotations and whose structures are represented by structural fingerprints. They comprise a combination of physico-chemical and substructure properties. The main principle behind the operation of SuperPred is that structurally similar compounds exhibit similar biological activity. For comparing the structural similarity between two molecules the Tanimoto coefficient was used. The results are given in terms of decreasing Tc values, providing structurally most similar compounds of the basic data set as well as their ATC classes. They state that SuperPred achieved an accuracy of 80.6 % for the fraction of compounds which exhibit a Tanimoto coefficient of > 0.85.

3.2.3.2 Classification Methods

Four classification methods that were used in this work for ATC class prediction of chemical compounds are briefly introduced in the following section. Its fundamental principles and the corresponding main parameters are described. For more information the reader is referred to further descriptions provided by the cited literature.

k-Nearest Neighbor (k-NN) is an instance-based classifier relying on similarity between instances. Those that have similar properties generally exist in close proximity within a dataset [Cover and Hart, 1967, Kotsiantis, 2007].

Similarity between instances is determined by relative distance metrics like Euclidean or Manhattan distance [Fielding, 2007, Kotsiantis, 2007]. In the ideal case, the metric minimizes the distance between two similarly classified instances, while maximizing the distance between instances of different classes. Hence, the Nearest Neighbor classifier assigns a novel object to the class of training examples to which it is closest in the feature space. k-NN locates the k nearest instances to the query instance and determines its class by identifying the single most frequent class label. Although, the k-nearest neighbor algorithm is sensitive to the local structure of the data, it is simple and produces remarkably low classification errors [Kotsiantis, 2007].

Decision Trees are trees which classify instances by sorting them on basis of attribute values. Each node in a decision tree represents an attribute, its edges form the decision making function [Kotsiantis, 2007].

To build a tree, the training data is repeatedly partitioned [Quinlan, 1986] based on an attribute value test. That attribute is selected in every step that best separates the training data according to a measure, which is the information gain of an attribute A with values $A_1, A_2, \dots A_v, v \in \mathcal{N}$:

$$gain(A) = I(p, n) - E(A)\,, \tag{3.12}$$

$I(p, n)$ is the entropy denoted as

$$I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \;\; and \tag{3.13}$$

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i)\,. \tag{3.14}$$

Here, n and p characterize the number of negative and positive examples for an attribute A in the training data.

The classification of a test instance is performed by moving from the root along the branches to a leaf node, that is assigned to a class label.

The most important parameter which influences the performance of the classifier and can avoid overfitting is tree pruning. Thereby noise is reduced by neglecting uninformative nodes and hence making better decisions is enabled [Mansour, 1997]. Furthermore, this also reduces the computational time and complexity.

The Naïve Bayes classifier is a probabilistic classifier and specialized form of Naïve Bayesian networks [John and Langley, 1995]. The Naïve Bayes classifier is based on Bayes' theorem

$$p(y|\boldsymbol{x}) = \frac{p(y)p(\boldsymbol{x}|y)}{p(\boldsymbol{x})}\,, \tag{3.15}$$

where $\boldsymbol{x}$ is a vector of random variables denoting the observed attribute values and y is a random variable denoting the class of an instance. The classifier is founded on two assumptions:

(a) the feature variables are conditionally independent given the class and

(b) it posits that no hidden or latent attributes influence the classification process [John and Langley, 1995].

A further common assumption often made is to consider the values of numeric attributes normally distributed within each class. Under the conditional independence assumption of the feature variables one obtains

$$p(\boldsymbol{x}|y) = \prod_{i=1}^{n} p(x_i|y)\,. \tag{3.16}$$

The denominator $p(\boldsymbol{x})$ of Equation 3.15 is not important for classification as it can be considered a normalization factor. Thus, normalization is achieved by the sum of $p(\boldsymbol{x}|y)$ over all classes giving the value one.

Given a test case $\boldsymbol{x}$, the Naïve Bayes classifier computes the probability of each class y given the vector of observed values and predicts the most probable class.

Support Vector Machines (SVM) were developed by Vapnik [1995] to solve binary classification problems [Schölkopf and Smola, 2002]. Thereby a hyperplane determined that separates two data classes with labels $\{1, -1\}$ by maximizing the possible distance between the separating hyperplane

$$\{\langle w, x\rangle + b = 0\} \tag{3.17}$$

and the instances on either side of a margin m. Here, x is the data vector, w is the normal vector orthogonal to the hyperplane, and b is the bias. Those hyperplane is determined, which minimizes $\frac{1}{2}||w||^2$ such that side condition

$$y_i(\langle w, x_i\rangle + b) \geq 1 \tag{3.18}$$

holds, given the training data, whereas y_i is the class label. The optimal separating hyperplane is defined through support vectors that are training instances closest to its boundaries

[Burges, 1998]. The values for w and b are determined through optimization methods, such as quadratic programming. The decision function used for assigning a class label to new data x has the form:

$$f_w(x) = \operatorname{sgn}(\langle wx \rangle + b)\,. \tag{3.19}$$

Often, data that have to be classified are not clearly separable, such that no hyperplane exists. A solution for this problem is to allow some misclassification. Thus, slack variables $\xi_i \geq 0$ are introduced giving a so called soft margin hyperplane:

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \;\; \forall \; 1 \leq i \leq n. \tag{3.20}$$

In this case, $\frac{1}{2}||w||^2 + \mathcal{C} \sum_{i=1}^{n} \xi_i$ is minimized. Here $\mathcal{C}$ is a positive constant, which is an error penalty chosen by the user, whereas a large C corresponds to a high penalty.
To find a hyperplane for non-linearly separable data, a solution would be to map them into a higher dimensional feature space $\mathcal{M}$. However, as such a mapping is computationally intensive and the illustration of the obtained separation could become very complex in the low-dimensional space, it is not feasible. Thus, a kernel trick is used, which allows to avoid the mapping into a high dimensional space and the calculation of the scalar products. Therefore, a kernel function

$$\mathcal{K}(x_i, x_j) := \langle \Phi(x_i), \Phi(x_j) \rangle \;, \tag{3.21}$$

is introduced, replacing the scalar product, whereas $\Phi : \mathcal{X} \to \mathcal{F}$ is the mapping and $\mathcal{F}, \langle \cdot, \cdot \rangle$ is the scalar product space, also called feature space. There are several commonly utilized kernel functions, such as polynomial or radial basis functions (RBF) [Burges, 1998].

3.3 Data Visualization

Visual representation of data translate them into a visible form which can highlight important features and enables rapid insight into complex data [Thomas and Cook, 2005]. Since the sense of sight is the dominant one for humans, visual data representation is essential for the analytical reasoning process, especially when dealing with large text corpora.

For making the above discussed information extraction techniques applicable and thus becoming more useful to biomedical researchers, the form of presenting text mining results is one of the major challenges of biomedical text mining [Cohen et al., 2005a]. As Hotho et al. [2005] illustrate, many of the graphical visualization approaches that have been developed for text mining purposes are based on methods which span from explorative data analysis and information visualization to visual data mining. Their aim is to integrate the human in the data exploration process by visual data analysis. The enabling to navigate thousands of documents can improve and simplify the use of literature and provides the capability to better and faster gain insight into massive data as well as extract relevant patterns or information and come up with new hypotheses [Keim, 2002].

According to Cohen et al. [2005a] and Zweigenbaum [2008] tools, provided to biomedical researchers who can benefit from the increasing amount of textual data, should address the following issues:

- Usability,
- Visualization type,
- Layout and visual cues,
- Navigation,
- Interactivity, and
- Retrieval speed.

Several web applications have been developed in the last years for supporting visualization and exploration of large text corpora, like Textpresso[18] [Müller et al., 2004], GoPubMed[19] [Doms and Schroeder, 2005], iHOP[20] [Hoffmann and Valencia, 2004], EBIMed[21] [Rebholz-Schuhmann et al., 2007], Whatizit[22] [Rebholz-Schuhmann et al., 2008], and SCAIView[23] [Gattermayer, 2007, Hofmann-Apitius et al., 2008]. They basically rely on the recognition of biomedical Named Entities and concepts, like proteins, genes, chemicals, Gene Ontology or MeSH terms in text documents, which are highlighted in the document view by most of the web applications. They enable document search and rank the obtained documents according to relevance or provide a statistical analysis of entities from several predefined entity classes. However, most of them are restricted to MEDLINE as document resource and cannot be customized to user defined entities in general.

Another tool AliBaba[24] [Plake et al., 2006] provides a graph representation of extracted entities that are linked by relation expressions identified in text and statistical measures. Gattermayer [2007] provides a detailed feature comparison of most available tools.

Since entities, obtained with Named Entity Recognition techniques extended and constituted in this work, were applied to augment the in-house developed web application SCAIView [Gattermayer, 2007], it is described in more detail at this point. SCAIView has been developed for enabling enhanced information retrieval and viewing entity as well as concept names recognized by Named Entity Recognition in a document corpus.

It is a knowledge discovery system that encompasses the capability of syntactic and semantic search[25]. The inclusion of a tree representing terminology hierarchy and ontology concepts of the biomedical domain allows for complex queries by a combination of query terms with the in- or exclusion of concepts of the provided hierarchies and ontologies. SCAIView builds on named entity recognition results of ProMiner and CRF approaches as

[18] http://www.textpresso.org/

[19] www.gopubmed.org

[20] http://www.ihop-net.org/UniPub/iHOP/

[21] http://www.ebi.ac.uk/Rebholz-srv/ebimed/

[22] http://www.ebi.ac.uk/webservices/whatizit/info.jsf

[23] http://scai.fraunhofer.de/scaiview.html?&L=1

[24] http://alibaba.informatik.hu-berlin.de/

[25] Syntactic search uses words or multi-words phrases that occur in documents and queries as atomic element in document and query representations. Semantic search is based on semantic analysis of documents through natural language processing techniques and retrieving documents by matching these semantic representations [Giunchiglia et al., 2008].

well as text indexing technology from Lucene, which is a high-performance Information Retrieval library developed for the generation and searching of text indexes by the apache foundation[26]. Identified named entities are added to the index as well. A front-end provides a graphic depiction, which currently provides two different visualization possibilities in form of distinct views onto the corpus tagged with recognized entities:

- **Entity view:** It gives information on the found entities x_i, like the ranking by Relative Entropy 3.22 (also known as Kullback-Leibler divergence [Kullback and Leibler, 1951]), the number of occurrence within the corpus, short entity descriptions for some entity classes, and links to external information resources. Thus, the web application augments identified and highlighted entities with additional available information.

$$RelativeEntropy(x_i) = p_1(x_i) \cdot \lg \frac{p_1(x_i)}{p_2(x_i)} \quad , with \tag{3.22}$$

$$p_1(x_i) = \frac{f_{x_is}}{F_s} \quad , and \tag{3.23}$$

$$p_2(x_i) = \frac{f_{x_ic}}{F_c} \quad , where \tag{3.24}$$

  f_{x_is} is the frequency of a single entity x_i of entity class z in a defined document subcorpus s,

  f_{x_ic} is the frequency of a single entity x_i of entity class z in the complete document corpus c,

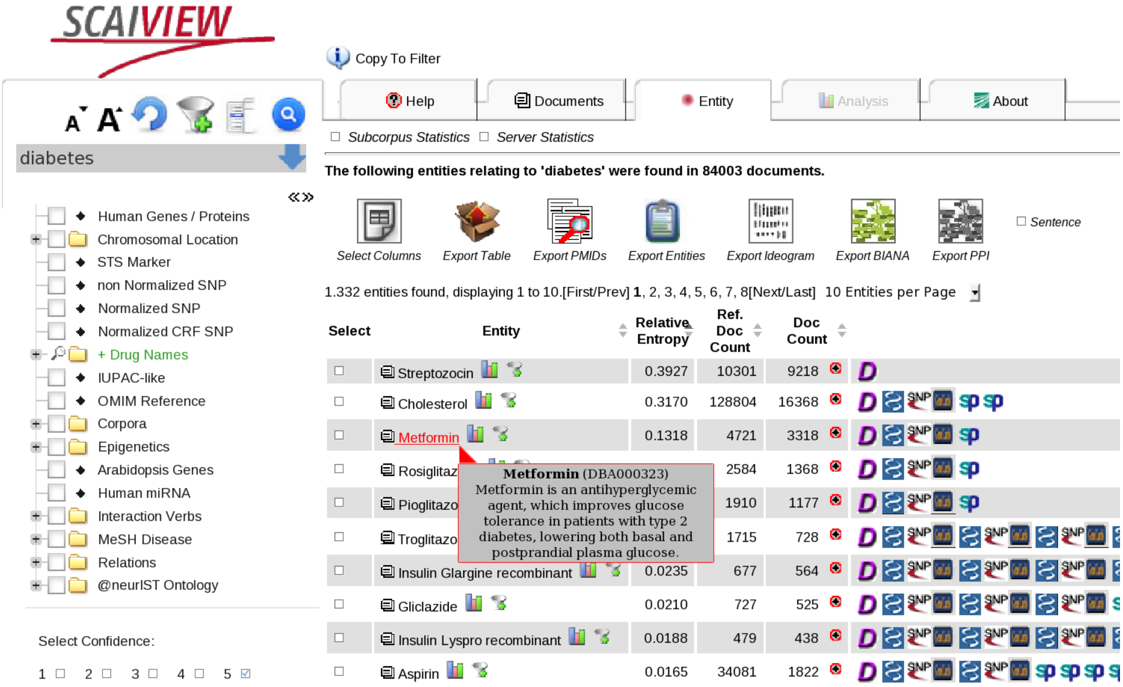
  $F_s = \sum_i^n f_{x_is}$ is the frequency of all entities $x_i \ldots x_n$ of entity class z in a defined document subcorpus s, and

  $F_c = \sum_i^n f_{x_ic}$ is the frequency of all entities $x_i \ldots x_n$ of class z in the complete document corpus c.

- **Document view:** It shows the subcorpus related to a distinct selected entity with the possibility to highlight all available biomedical entity types in the text. It helps to get a fast overview on the main entities and hence the theme of the documents.

Example of the entity and document view highlighting the NER results via SCAIView are illustrated by Figures 3.3 and 3.4.

[26] http://lucene.apache.org/



□ Subcorpus Statistics □ Server Statistics

□ Human Genes / Proteins
□ Chromosomal Location
□ STS Marker
□ non Normalized SNP
□ Normalized SNP
□ Normalized CRF SNP
+ Drug Names
□ IUPAC-like
□ OMIM Reference
□ Corpora
□ Epigenetics
□ Arabidopsis Genes
□ Human miRNA
□ Interaction Verbs
□ MeSH Disease
□ Relations
□ @neurIST Ontology

Select Confidence:

1 □ 2 □ 3 □ 4 □ 5 ☑

The following entities relating to 'diabetes' were found in 84003 documents.

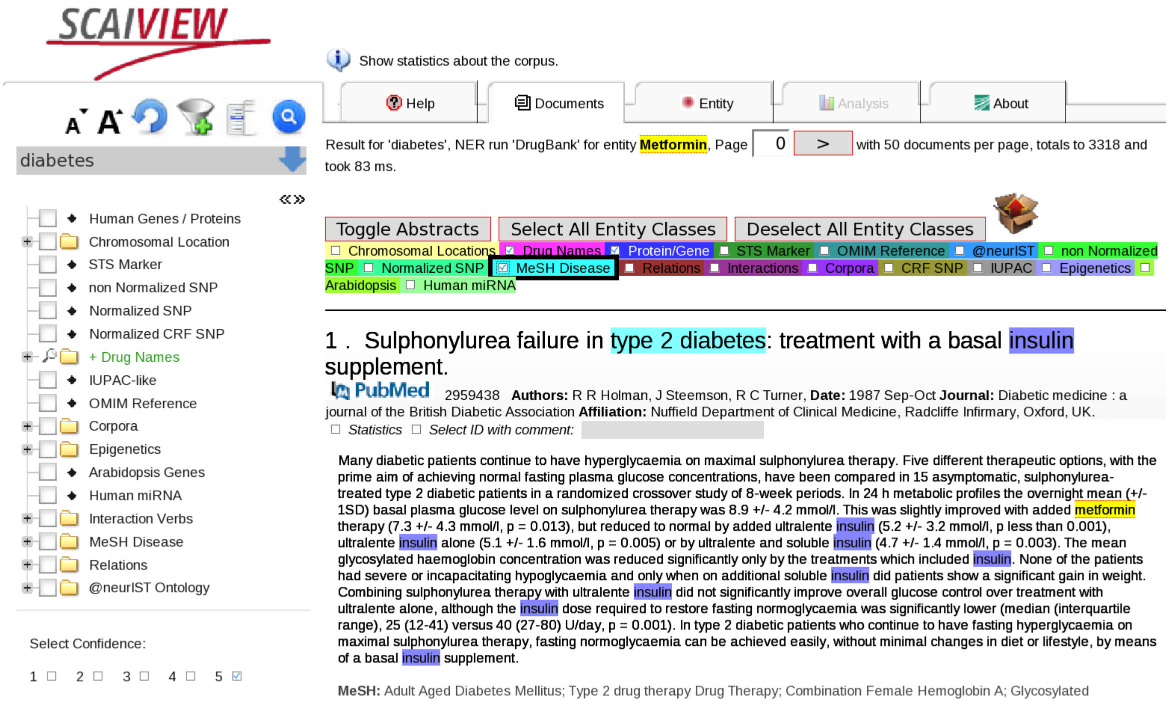
□ Sentence

1.332 entities found, displaying 1 to 10.[First/Prev] **1**, 2, 3, 4, 5, 6, 7, 8[Next/Last] 10 Entities per Page

Select	Entity	Relative Entropy	Ref. Doc Count	Doc Count
□	Streptozocin	0.3927	10301	9218
□	Cholesterol	0.3170	128804	16368
□	Metformin	0.1318	4721	3318
□	Rosiglitaz		2584	1368
□	Pioglitazo		1910	1177
□	Troglitazo		1715	728
□	Insulin Glargine recombinant	0.0235	677	564
□	Gliclazide	0.0210	727	525
□	Insulin Lyspro recombinant	0.0188	479	438
□	Aspirin	0.0165	34081	1822

Metformin (DBA000323)
Metformin is an antihyperglycemic agent, which improves glucose tolerance in patients with type 2 diabetes, lowering both basal and postprandial plasma glucose.

Figure 3.3: Entity view of SCAIView (screenshot). It shows a ranked list of drugs which are prevalently related to the queried disease *'diabetes'*.


SCAIVIEW
diabetes
«»
Human Genes / Proteins
Chromosomal Location
STS Marker
non Normalized SNP
Normalized SNP
Normalized CRF SNP
+ Drug Names
IUPAC-like
OMIM Reference
Corpora
Epigenetics
Arabidopsis Genes
Human miRNA
Interaction Verbs
MeSH Disease
Relations
@neurIST Ontology
Select Confidence:
1 □ 2 □ 3 □ 4 □ 5 ☑
Show statistics about the corpus.
Help
Documents
Entity
Analysis
About
Toggle Abstracts
Select All Entity Classes
Deselect All Entity Classes
□ Chromosomal Locations ☑ Drug Names ☑ Protein/Gene □ STS Marker □ OMIM Reference □ @neurIST □ non Normalized SNP □ Normalized SNP ☑ MeSH Disease □ Relations □ Interactions □ Corpora □ CRF SNP □ IUPAC □ Epigenetics □ Arabidopsis □ Human miRNA
PubMed
□ Statistics □ Select ID with comment:

Result for 'diabetes', NER run 'DrugBank' for entity **Metformin**, Page 0 > with 50 documents per page, totals to 3318 and took 83 ms.

1 . Sulphonylurea failure in type 2 diabetes: treatment with a basal insulin supplement.

2959438 **Authors:** R R Holman, J Steemson, R C Turner, **Date:** 1987 Sep-Oct **Journal:** Diabetic medicine : a journal of the British Diabetic Association **Affiliation:** Nuffield Department of Clinical Medicine, Radcliffe Infirmary, Oxford, UK.

Many diabetic patients continue to have hyperglycaemia on maximal sulphonylurea therapy. Five different therapeutic options, with the prime aim of achieving normal fasting plasma glucose concentrations, have been compared in 15 asymptomatic, sulphonylurea-treated type 2 diabetic patients in a randomized crossover study of 8-week periods. In 24 h metabolic profiles the overnight mean (+/-1SD) basal plasma glucose level on sulphonylurea therapy was 8.9 +/- 4.2 mmol/l. This was slightly improved with added metformin therapy (7.3 +/- 4.3 mmol/l, p = 0.013), but reduced to normal by added ultralente insulin (5.2 +/- 3.2 mmol/l, p less than 0.001), ultralente insulin alone (5.1 +/- 1.6 mmol/l, p = 0.005) or by ultralente and soluble insulin (4.7 +/- 1.4 mmol/l, p = 0.003). The mean glycosylated haemoglobin concentration was reduced significantly only by the treatments which included insulin. None of the patients had severe or incapacitating hypoglycaemia and only when on additional soluble insulin did patients show a significant gain in weight. Combining sulphonylurea therapy with ultralente insulin did not significantly improve overall glucose control over treatment with ultralente alone, although the insulin dose required to restore fasting normoglycaemia was significantly lower (median (interquartile range), 25 (12-41) versus 40 (27-80) U/day, p = 0.001). In type 2 diabetic patients who continue to have fasting hyperglycaemia on maximal sulphonylurea therapy, fasting normoglycaemia can be achieved easily, without minimal changes in diet or lifestyle, by means of a basal insulin supplement.

MeSH: Adult Aged Diabetes Mellitus; Type 2 drug therapy Drug Therapy; Combination Female Hemoglobin A; Glycosylated

Figure 3.4: Document view of SCAIView (screenshot). It shows the found documents that are related to the drug *'Metformin'* (highlighted in yellow) and *'diabetes'*. Genes and proteins as well as drugs and terms of the MeSH disease hierarchy part are highlighted in dark blue and light blue respectively.

Part II

Developed Systems and Implementations

Chapter 4

Building a Framework for the Information Aggregation of Chemical Entities

The first aim of this work addresses the generation of a framework for the aggregation of function annotation information on chemical entities from unstructured natural language text and structured resources. It relies onto two subtasks: the recognition of chemical named entities and the extaction of biomedical and pharmaceutical property information from text, which is combined with available structured annotation data. The approaches developed for both subtasks are individually covered in Section 4.1 and Section4.2. The final developed framework is described in Section 4.3, which includes the application of the aggregated function information.

4.1 Recognition of Chemical Named Entities in Text

Dictionary-based Named Entity Recognition methods introduced in Section 3.1.2.2 provide the strong advantage that normalization of found terms can be done in a straightforward way (as discussed in Section 3.1.3.2). The mapping of term representations of chemical molecules to ontologies or databases thus enables its linking to chemical structure representations and further information, like its biological effects, targets or physicochemical properties. Hence, the potential of applying a dictionary-based system for finding chemical named entities in text becomes clear. An important task for such an approach is the generation and of a domain-specific dictionary. It can be combined with an automated pre-processing step, the curation of synonyms contained in the dictionary which removes non-chemical terms, generates spelling variants, etc. This process supports the regular automated update of dictionaries and the maintenance of the quality of the system with manageable manual efforts. The post-processing of the obtained results is the third component which disambiguates found entities. All issues have been covered in the subsequently described studies preparing an approximate string matching method, like ProMiner, to be applicable for Chemical Named Entity Recognition.

4.1.1 Generation of an Evaluation Text Corpus and Annotation of Chemical Entity Classes

One of the basic steps when entering a new domain for that named entity recognition has to be developed, is the collection of a domain-specific text corpus. It is a foundation to get an overview how authors use the domain-specific terminology in text and to which main subclasses they belong to. The annotation of named entities is essential for the development of Named Entity Recognition methods, in particular for the training of machine learning based approaches and for the evaluation and comparison of the results of different tools.

In difference to the biomedical domain, where public contests, like BioCreAtIvE, lead to a number of corpora publicly available with several annotated biological entities (for an overview cf. Cohen et al. [2005b]), there is no corpus publicly available to establish and evaluate chemical NER systems. Several publications report on annotated training and evaluation sets generated for the development of their approaches [Corbett et al., 2007, Kemp and Lynch, 1998, Elena M. Zamora, 1984]. They used patents or articles from MEDLINE, the European Patent Office and journals of the American Chemical Society, hindering the distribution of the annotated corpora due to publisher conventions. Therefore, a new text corpus has been annotated with chemical entities.

MEDLINE was chosen as document resource. The corpus was assembled by R. Klinger performing following steps: A system based on CRFs for detecting IUPAC names described in [Klinger et al., 2008] was applied to select titles and abstracts from MEDLINE. The constraint was that they contain at least one IUPAC entity. The basic assumption for this choice was if that abstracts comprise IUPAC names, also other chemical terms can be expected in these abstracts. The selected corpus contains 67 abstracts with chemical terms functioning as positive examples and 39 documents lacking any chemical entities so that they encompass the negative text examples. This procedure resulted in the corpus CHEM-EVAL, which is composed of 106 documents (title and abstract) from MEDLINE and contains 31,791 tokens.

The corpus was annotated employing the tool WordFreak[1]. According to a primary term analysis, chemical named entities were assigned to seven classes defined with respect to morphological name properties and semantics: TRIVIAL, ABB, SUM, IUPAC, PART, MODIFIER, and FAMILY. An overview on the classes, their general definition and examples is given in Table 4.1.

In the following, the annotator's definition of the classes is described in more detail. The separation between TRIVIAL and IUPAC names is based on term length. Chemical terms consisting of only one word are classified as TRIVIAL, even if they are IUPAC names officially. Systematic multi word and semi-systematic names are always annotated as IUPAC. This includes names that imply only a IUPAC-like part and an abbreviation, like *'17-alpha-E'*, where 'E' stands for *'estrogen'*. The requirement is that the name needs to be complete. It also incloses names describing a radioactive labeling of a compound, e.g. *'3H-testosterone'*. Although it does not strictly follow the definition of IUPAC, the distinction was chosen to match the machine learning needs; the discrimination between tokens belonging to IUPAC and non-IUPAC terms. Furthermore, descriptions of substance classes which could be used as a base for building various chemical derivates and analogs were annotated as IUPAC and not assigned to the class FAMILY (e.g. *'1,4-dihydronaphthoquinones'*). Terms were only assigned to the class FAMILY if they describe well defined chemical families (e.g. *'glucocorticoid'*), excluding pharmacological classes (e.g. *'anti-inflammatory drug'*). Partial chemical names like *'17beta-'* occurring separately in text have been annotated as PART. The objective of this

[1] http://wordfreak.sourceforge.net/

Annotation class	Description	Examples
TRIVIAL	trivial and brand names	*'aspirin'*, *'estrogen'*
ABB	abbreviations and acronyms	*'TPA'*
SUM	sum formula, atoms, SMILES, InChI	*'KOH'*, '$C_1H_{22}N_2NiO_7$'
IUPAC	IUPAC names, IUPAC-like names, systematic, semi-systematic names	*'1-hexoxy-4-methyl-hexane'*
PART	partial IUPAC class names	*'17beta-'*
FAMILY	chemical family names	*'disaccharide'*
MODIFIER	names modifying the meaning of the chemical name	*'analogs'*, *'7-substituted'*

Table 4.1: Classes defined for the annotation of the CHEM-EVAL corpus, their description and example names.

entity class is to use it in future for the resolution of IUPAC name enumerations like *'2- and 3-methylhexane'*. In general, chemical names were not tagged if they are part of other entities like proteins e.g. *'3-ketosphinganine'* in *'3-ketosphinganine reductase'*.

The corpus annotated by two annotators, the author and Juliane Fluck which comprises 1343 annotated entities in total. The distribution of the annotated chemical entities over the classes is shown in Figure 4.1. As entity class MODIFIER is of no relevance for the present approach it is omitted in all following analyses of this work.

As Figure 4.1 illustrates, the main portion of the annotated entities belongs to the classes TRIVIAL (31.8 %) and IUPAC (29.0 %). They are followed by the other classes with a large distance. A generalization of the entity distribution for all documents in MEDLINE is not possible, because CHEM-EVAL is only a small corpus. Nevertheless, it gives a rough estimation which chemical entity types occur most probably in titles and abstracts of MEDLINE.

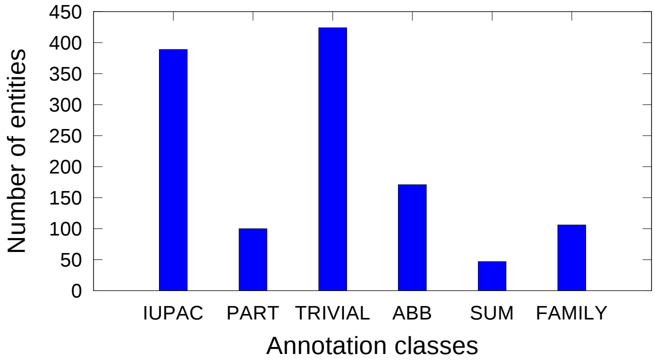


Figure 4.1: Chemical entity distribution over the defined chemical annotation classes on the annotated CHEM-EVAL corpus.

4.1.2 Generation of a Chemical Named Entity Dictionary

Definition of the Chemical Dictionary Content A chemical named entity recognizer should be able to identify names of chemical substances as well as chemical substance families. The identification of substructure and chemical class names helps to retrieve documents containing findings that subsume several chemical compounds.

Biopolymers like proteins and gene sequences, also chemical compounds in the sense of their definition, have been excluded from the assortment for a chemical dictionary. They are already comprised in protein and gene dictionaries of the available precursory version of ProMiner.

Terminology Resource Analysis Similar to gene and protein resources, databases for chemical substances do not only hold structural, chemical and physico-chemical or biological information. Repository providers collect denominations of chemical substances as well such that a mapping of chemical names to structures is given and normalization of found terms in text is possible. Following resources have been taken into consideration as terminology providers:

- Commercial Databases
  - **CrossFire Beilstein database**[2]
  - **CAS REGISTRY^{SM}** [3]
  - **The World Drug Index (WDI)** [4]
- Freely available Databases
  - **Kyoto Encyclopedia of Genes and Genomes (KEGG)**[5] [Goto et al., 1998]

[2] http://www.info.crossfiredatabases.com/

[3] http://www.cas.org/expertise/cascontent/registry/regsys.html

[4] http://www.daylight.com/products/wdi.html

[5] http://www.genome.jp/kegg/

  - **PubChem**[6]
  - **DrugBank**[7] [Wishart et al., 2006]
  - **Human Metabolome Database (HMDB)**[8] [Wishart et al., 2007].
- Thesauri and Ontologies
  - **Medical Subject Headings (MeSH)**[9]
  - **Chemical Entities of Biological Interest (ChEBI)**[10] [Brooksbank et al., 2005, Degtyarenko et al., 2008]

The considered data sources have been generated by the suppliers with different focus and hence provide different subsets of the entire chemical compound space. CAS, Beilstein, WDI, KEGG, DrugBank and HMDB are specialized on small molecules, polymers, drugs and organism's metabolites, whereas the thesauri and ontologies also collect chemical structure families and side groups of chemical molecules. More detailed descriptions of the resources can be found in Appendix A.2.1.

There are several requirements to be drawn on term resources used for dictionary-based NER approaches, like availability, number of provided entities and synonyms, low ambiguity of terms, etc. Given that entity normalization is a straight forward procedure and an inherent advantage of dictionary-based NER methods, the selection of dedicated references uniquely identifying the entities is another important issue for the choice of resources.

Several analyses have been performed to determine resources that can serve as terminology resource for compiling a chemical dictionary. The analysis is related to resource versions downloaded in January 2008. At first, the overall available number of entities provided by eight resources is considered. A general overview on the number of entries contained in the analyzed repositories is given in Table 4.2.

The data clearly show that the commercial databases stand out from the public resources in the number of available entries. Unfortunately, the commercial operators do not allow for using their databases to extract entity names and to establish a dictionary from them. Thus, only terms from freely available databases and ontologies can be extracted and utilized for generating dictionaries applied in NER systems. Therefore, the databases DrugBank, KEGG drug and KEGG compound, HMDB, PubChem as well as the ChEBI ontology, the MeSH hierarchy (referred to as MeSH-T), and the supplementary file of MeSH (referred to as MeSH-C) were considered for this work.

4.1.2.1 Raw Dictionary Generation and Performance Analysis

Dictionaries applicable by ProMiner have been created from all of the freely available resources to perform further analyses. Table 4.3 illustrates which data fields have been

[6] `http://pubchem.ncbi.nlm.nih.gov/`

[7] `www.drugbank.ca`

[8] `http://www.hmdb.ca/`

[9] `http://www.nlm.nih.gov/mesh/`

[10] `http://www.ebi.ac.uk/chebi/`

	Resource	Number of entries
Commercial	CrossFire Beilstein	10 mill
	CAS	33 mill
	World Drug Index	80,000
Public	PubChem-C; PubChem-S	18,4 mill; 36,8 mill
	MeSH-T	8,617
	MeSH-C	175,136
	ChEBI	15,562
	KEGG (K-C; K-D)	21,498 (15,033; 6,834)
	DrugBank	4,764
	HMDB	2,968

Table 4.2: Total number of entities contained in chemical information resources; Data from Jan 2008
(PubChem-C: PubChem Compound;
PubChem-S: PubChem Substance;
K-C: KEGG-Compound;
K-D: KEGG-Drug;
MeSH-T: subtree D of the main MeSH hierarchy;
MeSH-C: Supplementary Concept Records of MeSH)

extracted from the respective resources. Furthermore, the size of the dictionaries, the total amount of synonyms, and the calculated average synonym number per entity are provided.

It turns out, PubChem is the dictionary with the most objects, but the average synonym number is the lowest compared to the other dictionaries. Only those PubChem substance entries have been considered for dictionary generation which have at least one synonym and a link to PubChem compound. The second largest one is the supplementary file of MeSH (MeSH-C). However, it contains a low synonym average number as well. Contrarily, HMDB and DrugBank, the smallest databases, excel the others in that respect.

In a further study the actual number of synonyms per entry have been counted for every given resource. The distribution of the synonym count is depicted in the graphs of Figure 4.2. Many entries of PubChem, MeSH-C and MeSH-T, as well as DrugBank and HMDB contain a high amount of synonyms. This makes these resources very valuable, because many of the used synonyms in literature are potentially covered.

Only considering the number of entries and its provided synonyms is not enough. It is expected that some of the repositories partially overlap in entities and its synonyms, but will also be complementary to some extend. An example of entries overlapping in synonyms is documented in Table 4.4 for *'aspirin'*. It shows the number of its synonyms and examples of non-overlapping synonyms provided by the different resources.

Database	Fields used	Dictionary size	Synonym number	D
DrugBank	Name, Synonyms, Brand Names, Brand Mixtures, Chemical IUPAC Name, Chemical Formula, Isomeric SMILES, Canonical SMILES, CAS Registry Number	4,765	63,737	13.4
HMDB	Name, Common Name, Synonyms, Chemical IUPAC Name, Isomeric SMILES, Canonical SMILES, InChI Identifier, CAS Registry Number	3,007	41,576	13.8
KEGG-D	NAME, FORMULA	7,367	25,122	3.4
KEGG-C	NAME, FORMULA	15,172	43,976	2.9
MeSH-T	MH, ENTRY, N1, RN	8,617	79,615	9.2
MeSH-C	NM, N1, RN, SY	179,832	534,955	3.0
ChEBI	name, synonym, xref (if CAS is available)	19,935	86,768	4.4
PubChem	<PC-Substance_synonyms_E>	5,339,322	7,323,992	1.4

Table 4.3: Database/ontology fields used for the extraction of chemical names to be included in separate dictionaries. The dictionary size and number of contained synonyms is given respectively. D is the average number of synonyms per entity.

Test of Single Resource Dictionaries on the Corpus CHEM-EVAL Obviously, the considered resources have been developed for different purposes. A concluding question is which resources contribute in which extend to the six defined annotation classes of major interest. For instance, it can be expected that MeSH-T and ChEBI dictionaries will mainly match entities from the class FAMILY, because of the hierarchical chemical domain representation of the underlying resources. To get a preliminary impression of the class coverage of the dictionaries for the CHEM-EVAL corpus, a simple string matching approach has been applied as baseline experiment. It furthermore helps to identify and classify synonyms of the dictionaries that lead to false positive matches. Therefore, no pre-processing or curation of the dictionaries was performed, which means that no names were removed, added or changed. Following constraints were used for all searches:

- All synonyms were searched with a simple case insensitive string search, hyphens were ignored.

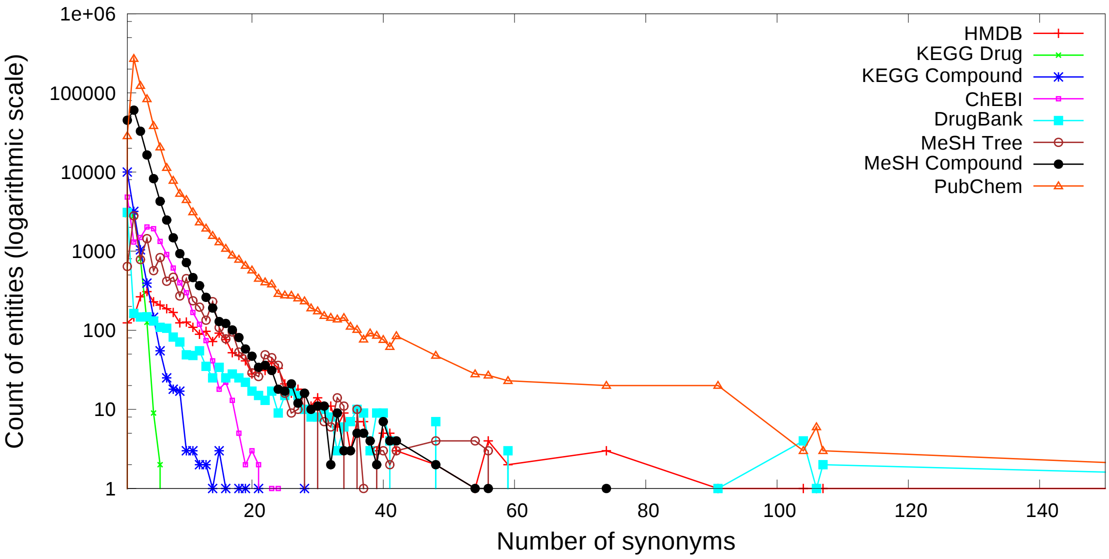


Figure 4.2: Plot of the synonym count distribution for the analyzed repositories.

Resource	Number of synonyms	Non-overlapping synonym examples
KEGG-C	7	—
KEGG-D	5	—
ChEBI	27	acide 2-(acétyloxy) benzoïque
DrugBank	> 100	Kyselina acetylsalicylova
HMDB	63	—
MeSH-T	20	—
PubChem	> 100	CCRIS 3243

Table 4.4: Overlap example *'aspirin'*: The resources, number of synonyms and examples of non-overlapping synonyms are provided.

- No control of the correct association of the found names to the corresponding entry was performed.

Result Analysis of the Simple String Matching The results obtained with a simple search strategy and uncurated dictionaries can be considered as a baseline and give only a rough estimate of the coverage of different sources. Furthermore, they help to choose terminology resources and to reflect which efforts have to be invested in dictionary curation and the adaption of the search strategy.

Table 4.5 provides recall and precision for every specific dictionary and the six annotated chemical named entity classes TRIVIAL, ABB, SUM, IUPAC, PART, and FAMILY (see also [Kolářik et al., 2008]). Furthermore, the identification results of every dictionary were combined (named as Concatenated Dictionary Results in the following) to get an impression of the potential optimal recall. The obtained recall and precision are presented in the right column.

- **Analysis of the Results obtained for Dictionaries**: The first row in Table 4.5 depicts precision and recall when the single class entities were analyzed altogether. It can be observed that precision and recall are not very high for all dictionaries. This is even true when the Concatenated Dictionary Results are analyzed. The highest recall when considering single dictionaries was obtained with the PubChem dictionary, identifying 33 % of all entries, followed by the ChEBI and MeSH-T dictionary (both 27 %). The combination of all results enhances the recall to 49 %, but decreases the precision to 13 %. The precision of 13 – 59 % is low due to several recognition problems and non-chemical synonyms. Table 4.6 gives a summarized overview on the identified recognition problem classes and lists some examples of false positives terms.

  The highest precision rates were achieved by the KEGG-D dictionary (59 %), followed by MeSH-C (44 %). In contrast, ChEBI and PubChem produced the lowest precision of 13 % and 15 % respectively. This is because ChEBI embodies many unspecific terms like *'groups'* or *'inhibitors'* by virtue of its hierarchical structure. Additionally, it comprises pharmaceutical property terms (e.g. *'enzyme inhibitors'* or *'adrenergic agonist'*) that have been excluded from the term class FAMILY by definition. In PubChem a lot of short names, common English words and single characters are responsible for false positives, these are for instance *'Serial'*, *'and'*, *'for'*, *'at'*, *'all'*, *'mg'*, *'reach'*. These terms belong to the false positive classes 1, 3 and 8 shown in Table 4.6.

  Many other names unspecificly occur in both resources. These are for instance one-character terms (e.g. *'D'*, *'J'*) and short names or abbreviations whose meaning is dependent on the spelling case (cf. false positives belonging to classes 7 and 8). This means in particular that if they are written in lowercase they match common English words, but if they are capitalized or written uppercase they represent a name of a pharmaceutics or a valid abbreviation.

  A further problem arises when chemical names are an integral part of another entity name, like a protein, which is shown by class 6 in Table 4.6. This would lead to the retrieval of false positive documents, while only being interested in the chemical and not in the protein.

- **Analysis of results obtained for the single chemical entity classes:** The analysis of the recall for every single annotation class shows that terms belonging to the TRIVIAL class could be found with the highest recall. The use of the PubChem dictionary identified 66 %, followed by MeSH-T with 64 % and KEGG-C with 57 %. The combination of the results by concatenation lead to a promising recall of 88 %. Considering the recognition of FAMILY names, the ChEBI and MeSH-T dictionary obtained the highest value (both 42 %). This result is not unexpected, because only those two resources contain general chemical group and family terms in their hierarchy. Sum formula were only recognized to a certain degree by the ChEBI, PubChem (both 31 %), and KEGG-C dictionary (12 %). The recognition rate of the ABB class has to be carefully regarded, because abbreviations are often short names, sometimes only one character long and therefore highly ambiguous. Entities belonging to class IUPAC have been recognized with a low recall by all tested dictionaries. It shows that IUPAC terms are either sparsely covered by the dictionaries or the simple search strategy is not capable to identify them. The bad result for class PART is inevitable, since only full names of chemicals lie in the interest of databases. Nevertheless, some are contained in ChEBI denominating chemical groups. However, terms of this kind, for instance *'diethyl'* or *'benzoyl'*, are part of chemical entity names like *'diethyl-N-[2-fluoro-4-(prop-2-ynylamino)benzoyl]-L-glutamate'* and increase the rate of false positive partial matches. Therefore, strategies to avoid such problems have to be integrated into the recognition approach.

In summary, from the obtained results it can be concluded, the overall recall of a simple search strategy using individual uncurated dictionaries is low. The results show that only the combination of all dictionaries leads to an acceptable recall rate for chemical named entities belonging to the classes TRIVIAL and FAMILY.

Curation processes applied to the dictionaries are necessary to achieve a higher precision performance. Competitively good results obtained by ProMiner for gene and protein name recognition at BioCreAtIvE 2007 lead to the assumption that the precision could be highly enhanced through dictionary curation and more elaborate named entity recognition techniques. This will be described in Section 4.1.2.2.

Class	PubChem	ChEBI	MeSH_C	MeSH_T	HMDB	KEGG_C	KEGG_D	DrugBank	Combined
ALL (1206)	*0.15* 0.33	*0.13* 0.27	*0.44* 0.10	*0.34* 0.27	*0.21* 0.16	*0.30* 0.24	*0.59* 0.12	*0.33* 0.13	*0.13* 0.49
IUPAC (391)	0.16	0.08	0.09	0.05	0.06	0.07	0.03	0.01	0.23
PART (92)	0.04	0.13	0.00	0.00	0.04	0.05	0.00	0.00	0.13
SUM (49)	0.31	0.31	0.04	0.00	0.00	0.12	0.00	0.00	0.31
TRIV (414)	0.66	0.52	0.18	0.64	0.36	0.57	0.35	0.40	0.88
ABB (161)	0.49	0.23	0.09	0.2	0.15	0.15	0.03	0.03	0.58
FAM (99)	0.18	0.42	0.05	0.42	0.08	0.19	0.17	0.00	0.71

Table 4.5: Comparison of the entities recognized in the evaluation corpus with dictionaries based on the analyzed resources. All annotation classes are considered. (The total number of the annotated entities per class are given in brackets.) Precision (slanted) and recall are given for an exact match of an entity.

	Classification	False positives
1)	Common English single words	*'neutrino'*, *'baseline'*, *'murine'*, *'probe'*, *'voltage'*, *'selective'*
2)	Colloquial terms	*'Sweet Stuff'*, *'Green Gold'* used for *'Cocaine'*
3)	Single characters	*'P'*, *'A'*, *'E'*, *'R'*
4)	Partial match	*'methyl'* in *'2-methyl-1-propanol'*
5)	Non-chemical terms	e.g. pharmacological class terms: *'enzyme inhibitors'*, *'antihypertensive agent'*, proteins: *'proteinase 3'*
6)	Chemical name as part of another entity name, e.g. a protein name	*'norepinephrine'* within *'norepinephrine transporter'*
7)	Chemical name if searched case insensitive matches common words with another meaning	*'proven'* was identified through synonym *'Proven'*
8)	Abbreviation matches common words	*'his'* was identified through synonym *'HIS'*
9)	Ambiguous abbreviations	*'CAP'* is a short form of the chemical entity *'Chloramphenicol'* as well as an abbreviation of a disease (*'community-acquired pneumonia'*), a protein, a pathway name [Prada et al., 2006], the modified nucleotide at the 5´end of messenger RNA, etc.

Table 4.6: Classification of false positive identified names using a simple search strategy.

4.1.2.2 Improvement of the Dictionary Quality by Curation

Before the curation process is described in detail, an overview on the dictionary generation process is provided. The workflow is illustrated in Figure 4.3, which shows the flow from the raw data to the raw dictionary that is processed by the curation procedure. The curated dictionary is then included into ProMiner. The main influential parameters/settings of the curation and recognition process are included. Furthermore, the general structure of the Named Entity Recognition result is shown. Dictionary curation implies processing

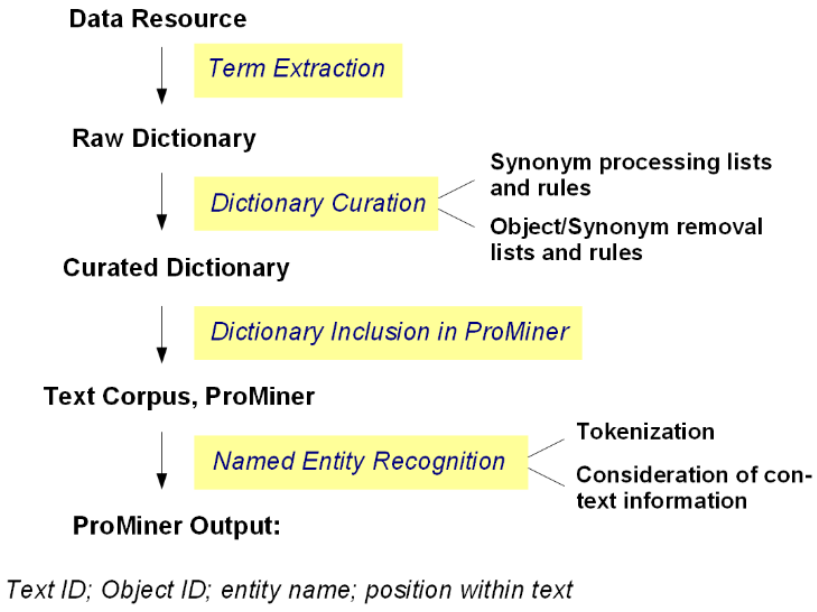

Data Resource
Term Extraction
Raw Dictionary
Dictionary Curation
Synonym processing lists and rules
Object/Synonym removal lists and rules
Curated Dictionary
Dictionary Inclusion in ProMiner
Text Corpus, ProMiner
Named Entity Recognition
Tokenization
Consideration of context information
ProMiner Output:
Text ID; Object ID; entity name; position within text

Figure 4.3: Workflow of the dictionary generation and processing. The single dictionary generation steps and its use within the approximate string matching approach ProMiner are illustrated. Finally, the general structure of the ProMiner output is given.

steps aiming at the improvement of the dictionary quality which leads to an increase in the NER approach performance. Thus, most of the recognition problems described in the previous section can be solved by removing some terms or complete entries. Additionally, a decrease in false positive findings can be achieved by defining a certain treatment for specific synonym classes. Basically, it can be divided into two steps:

1.) **Automated curation:** This is an automated preprocessing of the raw dictionary supporting periodic updates of the dictionary. It is divided into two steps:

   - Automated expansion and deletion of synonyms or objects
   - Automated classification of synonyms

   The information about changed entries and classified synonyms is stored in control files which are accessible to curators in a readable format.

2.) **Manual curation:** This means a manual check of the identification results after a ProMiner run on a large text corpus like complete MEDLINE and leads to:

   - Manual expansion and deletion of synonyms or objects
   - Manual classification of synonyms

   The automatedly generated curation control files are manually extended by this procedure.

Automated Curation Databases are regularly extended and updated and require automated curation processes to support periodical updating of the dictionaries without user intervention. Serving this demand, ProMiner contains a curation module, which was originally developed for the gene and protein domain Hanisch [2005]. Its generic design allows to adapt it to new domains, like chemical named entities. In the following the general concepts of the raw dictionary preprocessing are described in more detail:

- **Automated Synonym and Object Processing**
  - **Automated synonym expansion and processing:** This aims at generating and adding spelling variants to extend the entries. The reason for this is the observation that authors use synonyms and spelling variants of chemical names in text which are not covered by the applied terminology resources e.g. *'SantalolA'* added as variant of *'Santalol A'* or*'1 Naphtol'* added for *'1Naphtol'*. Another observation was that synonyms of pharmaceuticals provided by databases contain the extension of the producing company or the assignment to nomenclature classifications of nonproprietary names, like the International Nonproprietary Names (INN)[11] or United States Adopted Names (USAN)[12]. Usually they are provided in brackets following the chemical name, like *'torasemide (INN)'* or *'torsemide (USAN)'*. However, such name extensions do not occur in scientific texts. Since ProMiner assumes that these tokens of a synonym are a valid parts of it, they have to be removed from synonyms, otherwise they could not be found in text. Such name extensions were defined by automatic generated term lists from synonyms provided by the introduced databases that were subsequently manually filtered and applied for automated synonym processing. This resulted in a term list of 1672 words.
  - **Automated synonym and object deletion:** The utilized terminology from chemical databases, the MeSH hierarchy, and the ChEBI ontology contain entries, like proteins, which have to be excluded from the dictionary by definition. Such entries (compare with Table 4.6 class 5) have been automatedly removed by looking up key words and suffixes specifically characterizing wrong synonyms. Key words of proteins are for instance *'protein'*, *'receptor'* or the suffix *'ase'*, a typical ending of an enyzme name. Proteins not possessing such term properties (e.g. *'TNF alpha'*), have been identified by ProMiner applying the human protein and gene search within the chemical raw dictionary. Pharmacological class terms were removed which contain keywords like *'agonist'* or *'blocker'*. Beyond, terminology resources contain names (cf. class 2) which are for instance colloquial language terms. Since they lead to false positive matches, such synonyms need to be removed as well. To identify these terms in the raw dictionary, curator-defined term lists are used by the curation module. Additionally, single lowercase words (Table 4.6 class 1) which occur in a user specified common English word list are deleted as well. Filter checking the synonym length have been applied to delete

[11] http://www.who.int/medicines/services/inn/en/index.html

[12] http://www.ama-assn.org/ama/pub/about-ama/our-people/coalitions-consortiums/united-states-adopted-names-council.shtml

synonyms comprising only one or two characters or numbers and hence treating false positives of class 3.

- **Automated Classification of Synonyms**

  This procedure assigns terms to predefined classes whereby they are marked to be specifically treated during the term matching procedure. These classes are:

  - **Case-sensitive synonyms:** These target ambiguous terms whose meaning is dependent on the case, concerning terms of classes 7 and 8 described in Table 4.6. They have to be searched in a case-sensitive manner. The automated assignment to class Case-sensitive synonyms is a rule-based process including two main steps:

    1. If a lowercase equivalent of a non-lowercase synonym from the raw dictionary is found in a user defined common English word list (e.g. synonym: *'Proven'* and list entry: *'proven'*), the synonym is added to a defined list.
    2. Application of regular expressions to synonyms like [A-Z]{3-4} which matches synonyms like *'ADA'*.

  - **Questional:** Synonyms, especially short ones described by classes 9 and also 8, like *'CAP'*, can have several meanings. This particular one is an abbreviation for *'Chloramphenicol'*, but could also be a name for a pathway, the description of the modified nucleotide at the 5´end of mRNA or a headdress. Therefore, such terms should only be accepted as match, if a further synonym of the respective entity is identified in the same document as well. This is encoded by a further control file used by the curation process.
  - **Synonyms having a specific structure:** There are synonyms that obtain a specific treatment during tokenization.

    * Some synonyms consist of numbers and delimiters at which text strings are normally split during tokenization. CAS numbers are this terms for instance, e.g. *'50-78-2'* which would be separated into *'50'*, *'78'* and *'2'*. As single numbers would match to many occurring terms in text resulting in many false positives, they are defined to be left together.

  - **Standard synonyms:** All other synonyms are members of this class and are detected in a case-insensitive manner.

The generated synonym handling information is stored in curator alterable term lists for the two non-standard synonym classes.

Manual Curation If adapting the curation process to a new domain or if databases are massively extended by new entries, the generated domain-specific rules for automated entry extension, deletion and synonym classification might not cover new problem cases. Therefore, the automatedly curated dictionary is applied in a ProMiner search on a large corpus like MEDLINE, whereas the recognition result is visualized and checked by the author thereafter. Especially recognized synonyms occurring with a high frequency in a large corpus can be an indicator for problem cases. Hence, they may point to ambiguous synonyms often used in different domains.

Only manual curation can help to uncover new problem terms which were not contained in a small test corpus like CHEM-EVAL. It leads to a manual extension of the dictionary and control lists corresponding to the introduced classes and an adjustment of the classification rules. New synonyms are included into the dictionary which have been found in text, like *'POB'* for *'Phenoxybenzamine'* or the company code *'SDZ WAG 994'* for *'N-Cyclohexyl-2`-O-methyladenosine'* The file containing *Questional* names, for example, was extended by 547 new entries.

4.1.2.3 Adjusting the Approximate String Matching to the Chemical Domain

Named entities of different domains exhibit a variable structure in terms of its composition (multiword terms) and types of special symbols used within a name. Therefore, tokenization of the dictionary terminology as well as text can have an influence on the recognition performance. Furthermore, as biological and chemical entities are physically related to each other, this property is also reflected by their names. Authors use nested terms for describing proteins or pathways e.g. protein *'androgen receptor'* where *'androgen'* is a chemical entity by itself. Hence, chemical names can be integral part of other entity names, also shown by the false positive terms of class 5 in Table 4.6. Here, the protein – mostly a receptor or an enzyme – is described and not the chemical substance itself. If such terms would be integrated into a succeeding term frequency analysis of a text corpus used for document retrieval for instance, misleading results could be obtained when there is only interest in documents describing the chemical and not the protein. Hence, these chemical terms should not be found by the term search procedure because they are part of a different entity class.

- **Modulation of the Tokenization:** Many chemical named entities are composed of several words and comprise commas, white spaces, hyphens, apostrophes and different types of brackets. Thus, compared to terms of other domains they provide a specific structure that has to be considered for tokenization. As the tokenization of a string sequence depends on the types of delimiters used for separation, different settings can influence the recognition results.

  The approximate string matching tool ProMiner includes a tokenizer which transforms text as well as the synonyms into a sequence of tokens. By default the string splitting takes place at brackets, punctuation marks, and white spaces. For chemical entity recognition the hyphen has been introduced additionally to test its influence on the system's performance.

  In a tokenization variation study the curation parameters were left constant, whereas the tokenization performed with different sets of delimiters. In the first study the default delimiter set was used. In a second one the hyphen was added to the delimiter set. The analysis was performed by applying the DrugBank dictionary within ProMiner. The obtained results were evaluated on the test corpus CHEM-EVAL.

  The yielded results studied in dependence of different tokenization settings did not show a marginal difference in the recognition performance. Utilizing the hyphen as additional delimiter in the tokenization improved the precision only by about 1 % compared to the default tokenization settings (data not shown). This result was obtained at the cost of a recall drop by 1 %. Preferring a high recall, tokenization default settings were used in all following chemical named entity recognition studies.

- **Consideration of Context Information:**

  The approximate search algorithm of ProMiner provides a possibility to reject such partial term matches. It analyses the context of a recognized entity name and is able to look for specific tokens (usually words) occurring nearby an entity name, often directly succeeding it. Such context tokens change the meaning of a term are defined as forbidden tokens by the token class *Modifier*. Hence, a curator-defined term list is used by the search machinery that contains single words corresponding to tokens generated by the tokenization process. If tokens of this class are found in or nearby a candidate term in text, but no other similar term of the dictionary contains them, it is not accepted as a hit.

  To create such a curator-defined token list, an available *Modifier* list used for protein and gene name detection from ProMiner was analyzed for its application in the chemical domain. 28 terms like *'motif'* or *'domain'* important for proteins and genes, but not for chemicals were manually removed. Additionally, the list was manually and automatedly extended by further tokens modifying the meaning of a chemical entity name, like *'pathway'* in *'CAP pathway'*. However, most of the added tokens are words specifying enzymes or enzyme classes. They were derived from the enzyme database Brenda[13] and comprise about 1400 head words of enzyme names, e.g. *'oxidase'* and *'hydrolase'*. The impact of the inclusion of context information was evaluated together with the dictionary curation which is shown in the following section.

4.1.2.4 Evaluation of the Dictionary Curation and the Approximate String Matching

For studying the impact of the curation and the adjusted approximate string matching, single dictionaries from all studied resources were evaluated on CHEM-EVAL. They were left uncurated or were processed by the curation pipeline and then included into ProMiner. Figure 4.4 provides precision and recall for every single case.

It is obvious that the dictionary curation and an approximate string matching procedure adjusted to the chemical terminology characteristics results in a considerable improvement of the precision without a major loss in recall. Regarding precision, it could be demonstrated that the curation process results in high quality dictionaries. The highest increase of precision was obtained for ChEBI with 50 %, followed by DrugBank with 48 % and HMDB with 47 %. The lowest improvement was observable for KEGG-D with 22 % and PubChem with 28 %. In case of KEGG-D a relative high precision was already obtained with the uncurated dictionary, whereas the precision with the raw dictionary of PubChem is low. The analysis of the false positives revealed that terms or entries corresponding to all error classes described in Table 4.6 were removed or successfully tackled by the curation process, except those which belong to class 4. This means that general English words, one- and two-character terms, and

[13] www.brenda-enzymes.info

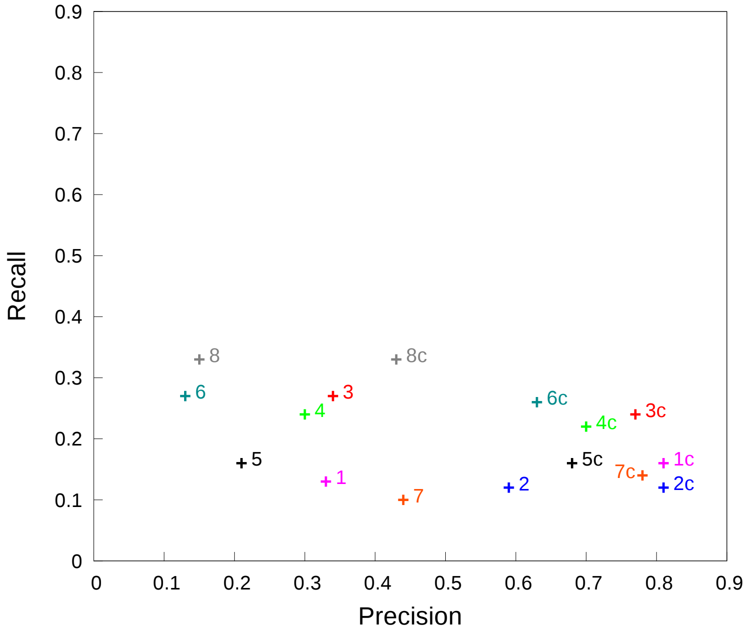


Figure 4.4: Impact of the curation process on the single dictionaries evaluated on CHEMEVAL. Precision and recall are provided for the ProMiner search with dictionaries either uncurated (indicated by a colored number) and curated (indicated by a colored number + ‘c’):
1: uncurated DrugBank; 1c: curated DrugBank
2: uncurated KEGG-D; 2c: curated KEGG-D
3: uncurated MeSH-T; 3c: curated MeSH-T
4: uncurated KEGG-C; 4c: curated KEGG-C
5: uncurated HMDB; **5c**: curated HMDB
6: uncurated ChEBI; 6c: curated ChEBI
7: uncurated MeSH-C; 7c: curated MeSH-C
8: uncurated PubChem; 8c: curated PubChem

colloqial names as well as objects describing non-chemical entities were removed. Terms that have different meaning when searched case insensitive cause false positive matches were set as case sensitive and ambiguous abbreviations were resolved. Furthermore, the context-adapted search allowed the filtering of proteins which comprise a chemical entity as part of their names.

The remaining false positives are caused by partial matches of synonyms in chemical names and correspond to error class 4. An example is *‘pyran-4-one’* which was found as part of *‘2-morpholin-4-yl-6-thianthren-1-yl-***pyran-4-one’** or *‘testosterone’* found as partial match of both *‘[1 beta-3H]***testosterone’** and *‘3H-***testosterone’**. Such partial matches are responsible for false positive as well as false negative NER results at the same time.

The low decrease in recall in general and even a slight improvement of 3 % for DrugBank and MeSH-C shows that the curation procedure does not remove many valid synonyms of chemical entities. Thus, it does not lead to many further false negative results.

A good recognition of named entities in text is a basic step in making natural language accessible for further information extraction methods, data analysis and information retrieval. However, the recognition alone is not sufficient to serve NER successive approaches. Especially the performance of relation extraction methodologies or information retrieval systems that enable semantic search are dependent on the mapping of different synonyms of an entity to one representation.

The merging of the single dictionaries which results in one combined chemical name list is described in the following sections.

4.1.2.5 Generation of a Final Chemical Dictionary

Combination of Single Curated Chemical Resource Dictionaries To support information extraction methodologies and information retrieval approaches that allow for semantic search, the mapping of different synonyms which belong to one chemical entity to one representation becomes very important. It allows to collect and extract all occurrences of an entity from a given text corpus in one step. An unresolved synonym mapping of entities that origin from different resources becomes apparent especially in information retrieval tasks. This is illustrated by Table 4.7 which shows differences in the number of MEDLINE articles that were obtained for the entity *'Aspirin'* when single resource dictionaries were used in ProMiner. These results are compared to the article number gained with an exemplarily manual unified list of *'Aspirin'* synonyms generated from all resources.

As can be seen, the dictionary KEGG-D with the least number of synonyms yielded the fewest number of articles. However, to obtain a high number of articles it is important to include an almost complete set of synonyms that represent an entity. Since this situation is not given with the single dictionaries, the retrieved article sets differ in dependence of the dictionary used for the article search. As result, several names corresponding to a chemical entity point to different identifiers, because they are derived from various data resources. Thus, they are considered as different entities by the approach, although they represent the same chemical entity. This leads to the problem that the analysis of the partially redundant and non-overlapping NER results for *'Aspirin'* which point to different resource identifiers is made laborious. Thus, to allow that every obtained article on a certain chemical entity is included within an information retrieval or extraction task, they have to be pooled ex post by a good strategy. In contrast, if a synonym list was used that is a combination of synonyms from all resources, more articles were obtained at once in comparison to all single results. This is reflected by the last row in Table 4.7. This example and the analysis of the raw and curated single dictionaries in Section 4.1.2.1 demonstrate, there is no single repository which can be particularly considered as a standard resource for the generation of a chemical name dictionary. A combination of the single dictionaries would lead to a high recall, especially for the chemical entity classes TRIVIAL and FAMILY. Thus, merging of corresponding dictionary entries of every chemical entity from the different resources in advance would remove redundant synonyms and decrease the number of entries, thus reducing the size of the final dictionary. Furthermore, the join of objects from resources that do not provide a link to a structure with those that are related to a structure representation, like the InChI identifier, supports term-structure normalization.

Resource	Resource identifier	Number of synonyms	Obtained articles from MEDLINE
KEGG-D	D00109	5	77,448
MeSH-T	D001241	20	77,686
KEGG-C	C01405	7	78,213
CHEBI	15365	27	78,252
HMDB	HMDB01879	63	78,582
PubChem	148573	156	79,031
DrugBank	DB00945	134	80,408
All combined	D00109; 15365; C01405; D001241; HMDB01879; DB00945; 148573	233 (joined and unified)	91,164

Table 4.7: Resource identifiers, number of synonyms, and number of MEDLINE articles obtained for *'Aspirin'* either using single dictionaries from the analyzed chemical data sources in ProMiner or a combination of its synonyms. The Resources are ordered by the obtained number of MEDLINE articles. The last row provides the article number for an exemplarily manual combined synonym list used in ProMiner which is linked with the normalized InChI identifier InChI=1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12).

Selection of the Resources to be Merged A prerequisite to the dictionary merging is to identify the resources to be joined. Additionally, strategies had to be established to map entities from several repositories onto each other which leads to one common dictionary for chemical named entities. To test whether every single dictionary, especially the large one PubChem, should to be considered for the merging process to make up the final dictionary, two analyses were conducted.

At first the synonym overlap of the single resources with PubChem was investigated. The results are shown in Figure 4.5. It was discovered that the synonym overlap of most single resources, except for the resources MeSH-T and MeSH-C, with PubChem is high. The high coverage of DrugBank and ChEBI is related to the inclusion of structures from DrugBank and ChEBI into PubChem. In contrast, MeSH-T provides a complimentary resource to PubChem since it contains chemical structure family denominations.

In the second study different concatenations of simple ProMiner outputs obtained with single uncurated or curated dictionaries were conducted and evaluated on the corpus CHEM-EVAL. Therefore, ProMiner results were successively added so that the impact of single added dictionaries onto recall and precision could be tested. At first those single ProMiner results of dictionaries were combined which yielded a high precision on CHEM-EVAL by their single evaluation described in Section 4.1.2.4 (cf. Figure 4.4). Hence, the initial result combination consists of DrugBank and KEGG-D. Step by step it was augmented with further ProMiner outputs of remaining dictionaries. Figure 4.6 shows recall and precision for the

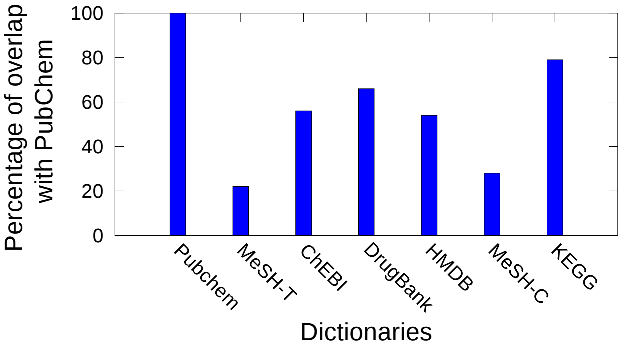


Figure 4.5: The figure shows the synonym overlap between the uncurated PubChem dictionary and uncurated dictionaries generated from DrugBank, HMDB, combined KEGG, MeSH-T, MeSH-C, and ChEBI. (The figure was adapted from [Kolářik and Klinger, 2008]).

different generated ProMiner result concatenations.

It could be shown that the addition of single ProMiner results obtained with single dictionaries improves the recall in every step. This is true for uncurated and curated dictionaries. Already the concatenation of the single results of DrugBank and KEGG-D increased the recall without main loss in precision. However, every following combination with ProMiner results of additional dictionaries decreases the precision. Especially the largest dictionary PubChem introduces the highest decrease in precision, followed by ChEBI. Even though PubChem contains most objects, its impact to the overall recall is only marginal in comparison to the combination result of all other dictionaries (compare result points 15 with 16 and 15c with 16c in Figure 4.6). Furthermore, only a small number of synonyms is related to the chemical entities in PubChem on average. It was observed that PubChem contains wrong synonym assignments to chemical entities. The latter fact would lead to wrong normalizations of terms and lowers the quality of a Named Entity Recognition System. Additionally, the size of the PubChem dictionary makes it difficult to handle. This leads to the conclusion that PubChem is not necessarily needed for the final chemical entity dictionary. Figure 4.6 also shows that the addition of the results obtained with the dictionary from database HMDB does not highly contribute to the recall. However, considering applications that are operated after Named Entity Recognition, the inclusion of HMDB into the final dictionary is valuable since this database provides additional information on chemical entities, like Gene Ontology annotations.

According to the considerations, following resources were considered for the generation of a combined dictionary: DrugBank, KEGG-D, KEGG-C, HMDB, ChEBI, MeSH-T and MeSH-C. They can be assigned to three basic semantic entity classes:

- Pharmaceuticals and nutraceuticals: DrugBank, KEGG-D, and MeSH-C
- Small molecules and metabolites: KEGG-C, HMDB, ChEBI, and MeSH-C
- Chemical family classes: MeSH-T and ChEBI.

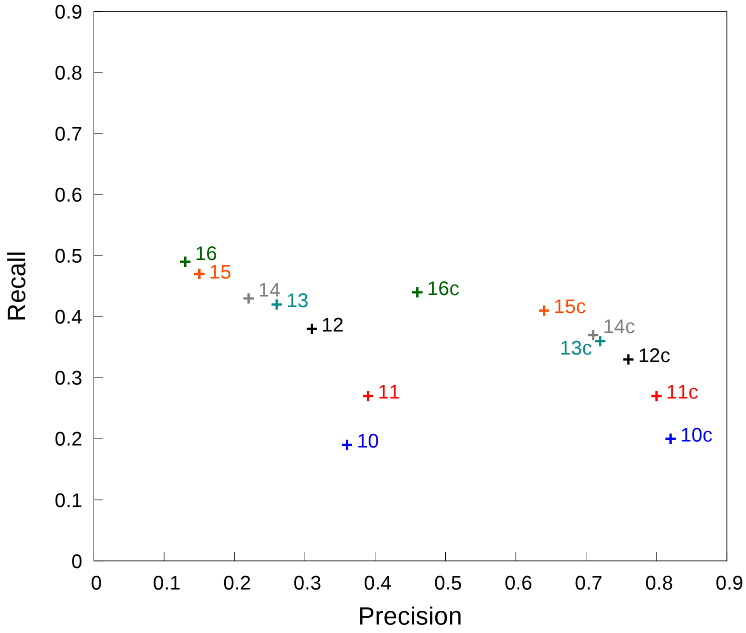


Figure 4.6: Impact of different concatenations of ProMiner results from single uncurated and curated dictionaries onto entity recognition. The evaluation was conducted on the corpus CHEM-EVAL, precision and recall are provided.
10: uncurated DrugBank, KEGG-D; 10c: curated DrugBank, KEGG-D
11: uncurated 10 + MeSH-C; 11c: curated 10 + MeSH-C
12: uncurated 11 + MeSH-T; **12c**: curated 11 + MeSH-T
13: uncurated 12 + KEGG-C 13c: curated 12 + KEGG-C
14: uncurated 13 + HMDB; 14c: curated 13 + HMDB
15: uncurated 14 + ChEBI; 15c: curated 14 + ChEBI
16: uncurated 15 + PubChem; 16c: curated 15 + PubChem

After choosing the resources a workflow had to be established for generating a chemical dictionary and hence finally adapting ProMiner to the chemical domain. It is introduced in short within the next section.

Merging Workflow Conception To obtain ProMiner$_{\text{Chem}}$, a ProMiner version that allows for Named Entity Recognition of chemical entities, dictionary generation and processing steps were combined to a workflow. The complete workflow, is drafted in Figure 4.7. It starts with the generation of the raw dictionaries by extracting the chemical terminology from the data repositories. This is followed by the curation of the dictionaries. Both processes were already described in the previous Section 4.1.2.2. The last step, which depicts the merging procedure, is used to combine selected single dictionaries. It is explained in the following paragraphs.

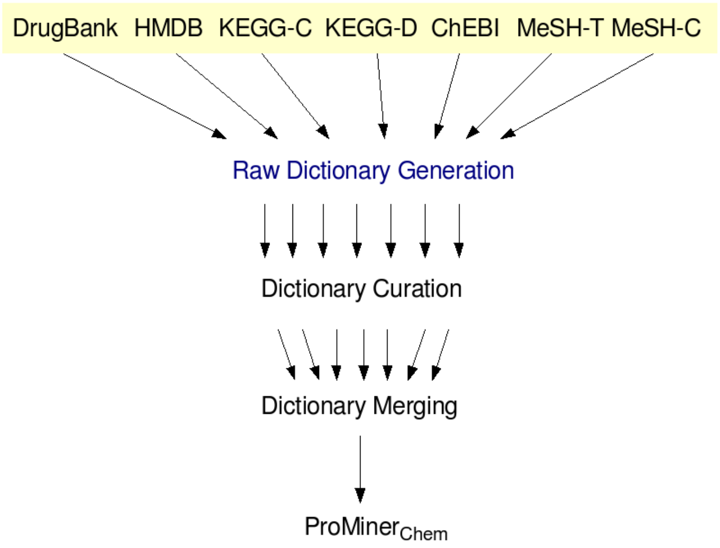


Figure 4.7: Workflow conception for the dictionary merging. The row highlighted in yellow shows the utilized resources.

Dictionary Merging The most suitable and simple approach would be to merge all dictionary entities which point to the same unique identifier commonly representing a chemical structure that is provided by all utilized resources.

InChI and CAS identifiers, introduced in Section 1.2.2.1, are uniquely related to chemical structure information. CAS is indirectly linked to a structure representation and InChI is a defined string representative of a chemical structure and hence unique. In contrast, InChIKey is not yet prevalently used in chemical data resources. In the case of SMILES several different string variants can be generated for one chemical compound so that it is hard to map them onto each other. Thus, InChI and CAS identifiers were considered to serve as denominators through which entries of different databases are mappable and can normalized by the relation to a chemical structure.

A concluding step is to analyze the coverage of both InChI and CAS identifiers in the discussed resources. The number of entities linked to InChI and CAS identifiers and its portion in comparison to the total entity number is provided in Table 4.8 respectively.

What can be seen is: MeSH-T and MeSH-C only refer to CAS identifiers, but to a low amount (~34 % and ~33 %). The low CAS-coverage of MeSH-T is consequential, because it is a hierarchy of chemical classes, mainly providing high level concepts. The entries of the other resources are linked to both identifiers, whereas InChI is more present than CAS. The coverage of entities linked to InChI is between ~54 % and ~99 %. The lowest fraction was obtained for ChEBI. This could be explained by the high number of chemical class names occurring in this ontology that cannot be linked to an InChI identifier. The addition of all objects from the single curated dictionaries and the common analysis of the InChI and CAS identifier number provides a general overview on the linkage of entities with these two identifiers. A corresponding statistics is provided in Table 4.9.

It points to the fact that CAS is the most prevalent identifier and hence needs to be

	InChI	CAS	Entry No.
KEGG comb.	17,507 (77.64 %)	13,814 (61.26 %)	22,548
DrugBank	4,487 (94.17 %)	2,225 (46.69 %)	4,765
HMDB	2,982 (99.17 %)	2,648 (88.06 %)	3,007
ChEBI	10,717 (53.76 %)	6,069 (30.44 %)	19,935
MeSH-T	—	2,936 (34.07 %)	8,617
MeSH-C	—	58,675 (32.63 %)	179,831

Table 4.8: Overview on the linkage of the entities to the identifiers InChI and CAS in the analyzed raw data sources. For KEGG the respective values of the drug and compound subdatabases were unified.

Total object number	172,992
Objects related to CAS	82,768 (47.85 %)
Objects related to InChI	35,153 (20.32 %)
Objects related to both InChI and CAS	23,146 (13.38 %)
Objects not related to InChI and CAS	78,370 (45.30 %)

Table 4.9: Statistics on the object number of the overall dictionary (after dictionary curation) and the linkage coverage of InChI and CAS.

considered for merging. About 27 % less objects are linked to InChI. Far less (only ~13 %) are related to both, CAS and InChI identifiers. In contrast, a high number of entries carry none of the two discussed identifiers, especially those from MeSH-T, MeSH-C, and ChEBI. The only solution to merge those, is to comprise their synonyms.

The developed merging strategies used for the join of corresponding objects, its challenges and obtained results are described within the following sections.

The analyses described in the former sections have lead the basis for the development of a merging strategy. Three merging steps have been developed utilizing the two identifier types InChI, CAS, and, if not available, entity synonyms. Primarily, they have been separately implemented, analyzed, and optimized. In the following every single merging strategy is described in detail.

CAS Merging The simplest process is the merging by CAS numbers. As the single information levels of CAS identifiers do not encode some chemical information, only a prefect string matching can be applied to this type of identifier. This ensures that only those objects are joined that are related to the same chemical structure. Hence, all entries that comprise an identical CAS identifier have been merged.

InChI Merging Available InChI strings linked to entities of the selected databases for merging have been compared by string matching. All entries of the resources possessing an identical InChI have been merged, leading to a combined synonym and database reference list.

As introduced in Section 1.2.2.1 the InChI string is segmented into layers that encode different information on the compound's structure, like charge, stereochemistry or certain isomeric variants of tautomeric structures. In the case of tautomers the InChI transformation program[14], developed under the auspices of the International Union of Pure and Applied Chemistry (IUPAC), provides different InChI generation options. There is a normalization procedure called *'Mobile H Perception'* when checked during the InChI creation the same InChI identifier is generated for different tautomeric isomers. When it is not used, InChI represents a specific tautomer by the presentation of a Fixed H-layer. According to this it was observed, that the used resources provide different InChI identifiers for a tautomeric compound that vary in the provided layers. An example structure represented by different InChI identifiers is depicted in Figure 4.8 for the compound *'oxalacetic acid'*. For the given example it turned out that ChEBI provides the InChI with a Fixed H-layer, whereas KEGG and HMDB omit it. It is assumed that this is caused by differences in the use of the InChI generation program which transforms a structure representation into an InChI string. Since these differences in InChI occurred very often for tautomeric structures obtained from different resources, the Fixed H-layer was removed from all available InChI identifiers by a preprocessing step. It was performed before the InChI comparison and entry merging was done.

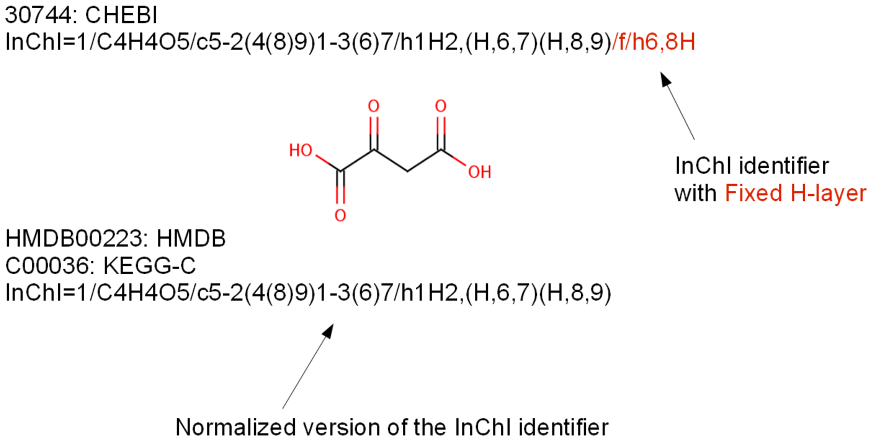

30744: CHEBI
InChI=1/C4H4O5/c5-2(4(8)9)1-3(6)7/h1H2,(H,6,7)(H,8,9)/f/h6,8H
O
O
HO
OH
O
InChI identifier
with Fixed H-layer
HMDB00223: HMDB
C00036: KEGG-C
InChI=1/C4H4O5/c5-2(4(8)9)1-3(6)7/h1H2,(H,6,7)(H,8,9)
Normalized version of the InChI identifier

Figure 4.8: An example of varying InChI identifiers provided by different resources for the substance *'oxalacetic acid'* is shown. Respectively, the corresponding database identifiers are given. The difference between the normalized InChI version and the non-normalized InChI is depicted by the 'Fixed-H layer', which is marked in red.

Synonym Merging Entries containing no InChI or CAS identifiers can only be merged if they contain a certain number of overlapping synonyms. Furthermore, there could be

[14] http://www.iupac.org/inchi/release102final.html

	Database ID	Synonyms
Entry 1	D004874 MeSH-T	Ergonovine Maleate, Ergotrate, Ergonovine, Ergometrine Maleate, Ergobasin, Ergometrine, Ergometrin, Bedford Brand of Ergonovine Maleate
Entry 2	D01163 KEGG-D	Ergometrine maleate, Ergometrine, Ergonovine maleate, $C_{19}H_{23}N_3O_2.C_4H_4O_4$
Entry 3	C07543 KEGG-C	Ergonovine, Ergometrine, $C_{19}H_{23}N_3O_2$
Entry 4	DB01253 DrugBank	Ergobasine, Ergotrate, Ergometrine, Ergotrate maleate, $C_{19}H_{23}N_3O_2$

Table 4.10: Entries of *'Ergometrine'* from different resources which overlap in their synonyms (respective synonyms are underlined). Respective structures from the given resources are depicted in Figure 4.9. Terms marked in blue are specifically highlighted for the illustration of the synonym problem described in Section 4.1.2.5.

entities which are linked in one resource to an InChI and/or CAS identifier, but do not carry any of those in another one.

The basic idea is when objects overlap in their synonyms to a certain extend, they are likely related to the same chemical compound and hence can be joined, thus reducing redundant information within the dictionary. An example is provided in Table 4.10. It shows entries of *'Ergometrine'* from the four resources MeSH, KEGG-D, KEGG-D, and DrugBank which overlap in its synonyms.

Synonyms occurring in different dictionary entries have been identified using a specific function of ProMiner, that automatically finds ambiguous synonyms. They were collected, whereas the information on their occurrence in multiple objects was stored. Furthermore, they were used to calculate the overlaps of dictionary objects in its synonyms. The fraction of shared synonyms was taken to define a precondition, the measure M and a constraint C that restricts the merging:

- **Precondition:** Objects that overlap by a certain set of ambiguous synonyms are collected, grouped and ordered by the number of their synonyms. The one with most synonyms is defined as standard entity E_S to which all other entities within one group $E_1 \ldots E_n$ are compared to. Subsequent, the fraction of overlapping synonyms F_O between E_S and E_n is determined. Thereby sum formulae are not expedient to be considered for the overlap calculation, because they are not unique by nature and hence are highly ambiguous (e.g. *'$C_8H_5O_4$'* is related to three different chemical entities – *'2-carboxybenzoate'*, *'3-carboxybenzoate'*, and *'4-carboxybenzoate'*). Table 4.11 shows an example group of four overlapping objects.

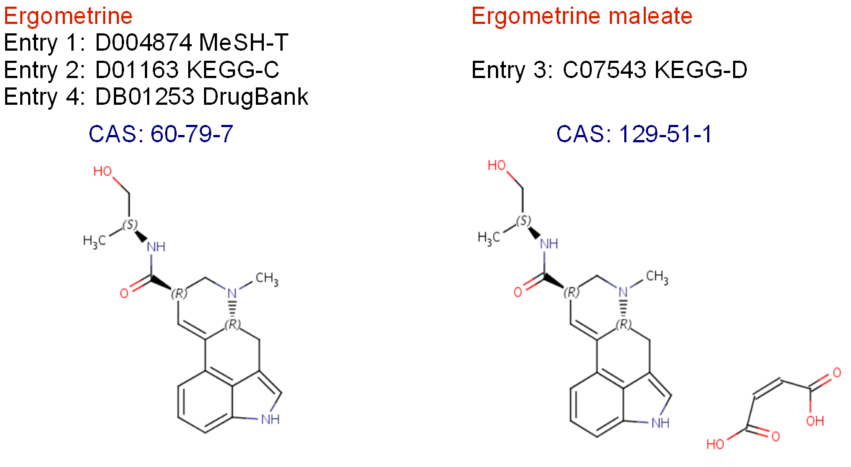


Figure 4.9: Molecular structures of the resource entries provided in Table 4.10.

- **Measure M:**

$$M = \frac{F_O \cdot 100\,\%}{S_{E_n}} \tag{4.1}$$

  M is calculated for every E_n, where S_{E_n} is the total number of synonyms of E_n. It presents which fraction of overlapping synonyms F_O between E_S and E_n in relation to the total number of synonyms S_{E_n} in E_n.

- **Constraint A (C):** The measure M is compared to a user defined threshold T. Merging of two objects E_S and E_n is allowed, if $M \geq T$. This results in the merged object $M_i = E_S \bigcup E_n$.

  If constraint C is not fulfilled, entities are not allowed to be merged.

The example provided in Table 4.11 contains the calculated constraint measures M that are compared to a defined threshold for T as well as the merging decision. During the procedure the merging decision information on each entity pair E_n and E_S is stored. For all E_n that overlap with E_S and fulfill the constraints, merging is done resulting in one new merged entity M_i. After the constraint testing and merging of all entries, the original objects E_n and E_S that were joined by the procedure were replaced by M_i and subsequently removed from the input object list.

Merging by synonyms is an iterative process, because a join of entries could lead to the fact that other entities fulfill the merging constraints in a subsequent iteration step. Hence, object merging has to be repeated until the resulting object number does not change between two succeeding iteration steps. This means that no new merging has been taken place in the last step.

The threshold value T that defines the merging constraint has been determined by experimentation through value variation. Selection criteria were the number of remaining objects and ambiguous synonyms as well as manual inspection of the merging results. Its analysis built the basis for the merging constraint threshold value selection.

Entity	Total No. of object synonyms	F_O	M	Merge decision for E_S and E_n
E_S ($E_S \hat{=} E_1$ of Table 4.10)	8			
E_2	3	3	100 %	**yes**
E_3	2	2	100 %	**yes**
E_4	4	2	50 %	**yes**

Table 4.11: Example of one sorted group of objects overlapping in the synonyms. The entries of Table 4.10 were taken as an example case. The object E_S contains the highest number of synonyms. (It corresponds to entity E_1 in Table 4.10.) The respective overlap in the synonyms F_O between entries E_n and E_S is given as well as the calculated values that correspond to the measure M_{CA}. The example threshold used for constraint C is $T = 30\,\%$. The merging decision is denoted by 'yes' or 'no'.

Merging Workflow Composed of the Single Merging Strategies To exploit the advantages of the three single merging procedures for the generation of a joined dictionary, the single steps were combined to a workflow, which is shown in Figure 4.10. By the fact that only ~20 % of the objects contain InChI and ~48 % CAS identifiers, only those entries were subjected to the corresponding InChI and/or CAS merging procedures. This is depicted by the conditions InChI=yes/no and CAS=yes in Figure 4.10. Since InChI is the potentially more reliable identifier, InChI merging was performed at first. Finally, all objects were passed to the Synonym merging procedure.

Conducting the workflow on the example entities shown in Table 4.10, they would be successively merged in the following order: since Entry 3 and Entry 4 provide the same normalized InChI identifier (data not shown) InChI merging joins both which results in MergeResult 1. It is followed by CAS merging which combines MergeResult 1 with Entity 1 resulting in MergeResult 2. As last step the synonym merging joins MergeResult 2 with Entry 2 finally generating a merged entry.

Analysis of the Merging Results Before the single merging strategy and merging workflow results are analyzed in detail, general assumptions on the expected results are made:

- It is expected that the data sources partially overlap in the provided chemical entities. Thus, the merging of objects from primary utilized resources leads to a reduction of the total object number in the dictionary. If it is assumed that a chemical compound or element is available in all considered resources, the maximal number of primary objects contributing to a joined object should not exceed the number of data sources used. Accordingly, one synonym should maximally occur in as many objects as data

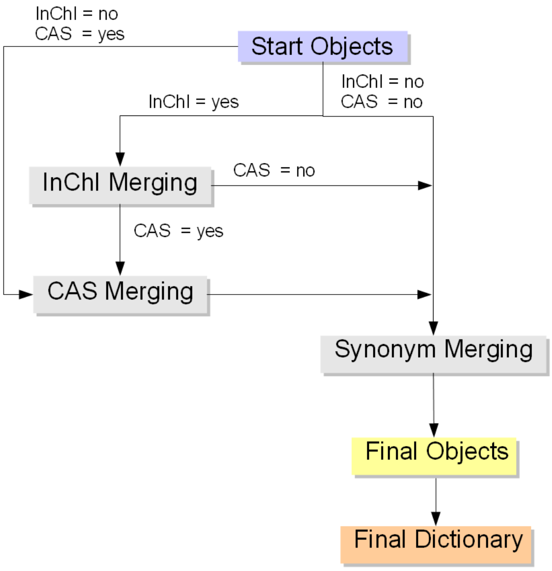


Figure 4.10: Workflow of the merging procedure.

sources were considered. In the present case this number is 7. Merged objects that consist of more than this amount of initial objects point to problems in the data or limitations in the merging procedure.

- The merging of objects should lead to a reduction of ambiguous terms.

The newly generated dictionary and the joined objects were evaluated accordingly. Furthermore, it was expected that the quality of the dictionary reflects the quality of the data. As the names for the closely related chemical compounds *'Ergometrine'* and *'Ergometrine maleate'* exemplarily depicted in Table 4.10 show, the terminological data provided by the resources exhibit some limitations. It was found that not all synonyms that are made available by the data sources are correctly assigned to chemical compounds and associated CAS and InChI identifiers. In the allocated example wrongly assigned names are marked in blue. It reflects that several names in three of the four given data sources are not correctly linked to the two given structures. In consequence of this, a synonym that is found in text for *'Ergometrine'* or *'Ergometrine maleate'* would be related to both structures, irrespective whether the entries were left separate or were merged. As closely related objects, such as organic acids and corresponding salts, are not distinguished through the database suppliers by name space, respective objects can be joined by the merging procedures. Concluding, the name and object space in the final dictionary can only be that distinct as it is provided by the sources.

To investigate the characteristics of the three single merging steps and its results, they were first separately analyzed. Subsequent, the results of the merging workflow, that consecutively connects the three single steps, are presented.

Results of the CAS Merging The CAS merging procedure reduced the object number by 16,102 objects which is a reduction by 9.31 %. Ambiguous terms were decreased by 24,294 (~52 %). Table 4.12 shows the distribution of the number of joined objects. As can be seen, a high amount of objects could not be joined by CAS merging. Which is followed by two, three and four resources which contribute most to joined objects. However, there are also few newly generated ones that were combined from more than 7 initial objects. They mainly consist of objects originating only from the database HMDB. They belong to the class of *'glycosphingolipids'* and *'ceramides'* that are not distinguished via CAS identifiers in HMDB. However, those HMDB entries possess different InChI identifiers, but share most of its synonyms. It shows that this database does not provide correct CAS identifiers for all its entries. This findings suggested to study the number of InChI identifiers that were newly combined in merged objects after CAS merging. The analysis of the results disclosed that 2,288 (~22 %) of the 10,219 newly joined objects are related to more than one InChI identifier. As their analysis revealed, around 2/3 of these InChI identifiers display differences in the Fixed H-layer as was illustrated in Figure 4.8.

Number of resources	Number of objects
1	146,671
2	6,524
3	2,231
4	956
5	384
6	96
7	10
8	3
9	3
10	1
11	5
12	4
13	2

Table 4.12: Distribution of the number of resulting objects joined from resources by CAS merging.

A systematic inspection of the connection between CAS and InChI turned out to be difficult for over 700 identified merged objects that comprise more than one InChI not differing in the Fixed-H layer. Thus, for some merged entities an exemplary analysis of the provided resource information was conducted manually. It revealed that the problems are manifold. Most of them mainly differ in the provided stereochemical information. Some examples of the latter ones are depicted in Figure 4.11.

A list of reasons for multiple InChI identifiers assigned to a newly joined object are given by the following examples:

- Interchange of stereochemistry of chemical compounds; there is confusion with the

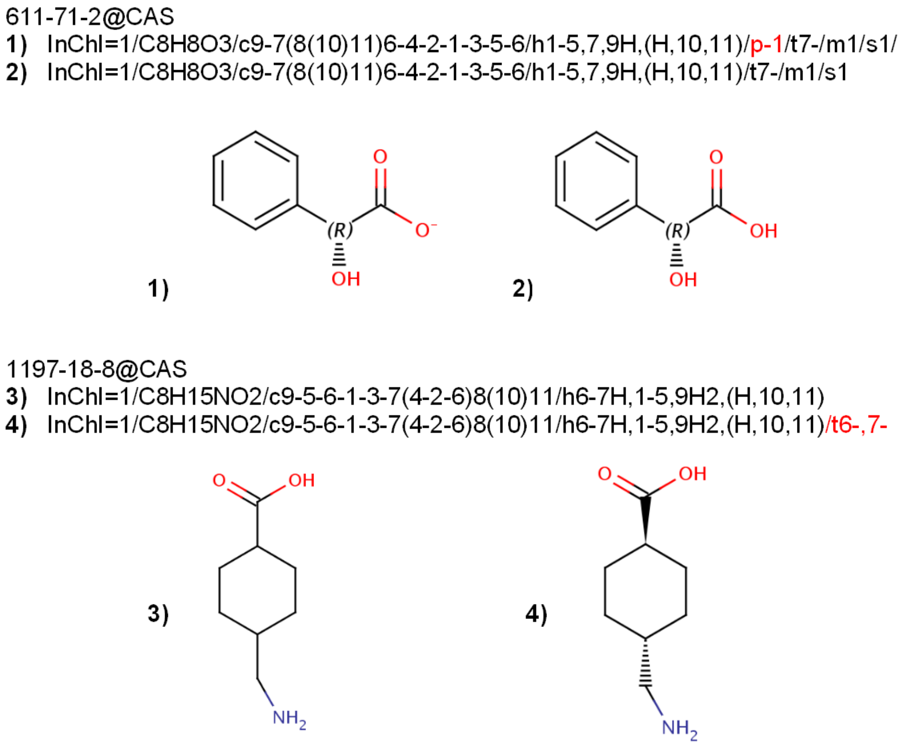

611-71-2@CAS
1) InChI=1/C8H8O3/c9-7(8(10)11)6-4-2-1-3-5-6/h1-5,7,9H,(H,10,11)/p-1/t7-/m1/s1/
2) InChI=1/C8H8O3/c9-7(8(10)11)6-4-2-1-3-5-6/h1-5,7,9H,(H,10,11)/t7-/m1/s1
O
(R)
O⁻
OH
1)
O
(R)
OH
OH
2)
1197-18-8@CAS
3) InChI=1/C8H15NO2/c9-5-6-1-3-7(4-2-6)8(10)11/h6-7H,1-5,9H2,(H,10,11)
4) InChI=1/C8H15NO2/c9-5-6-1-3-7(4-2-6)8(10)11/h6-7H,1-5,9H2,(H,10,11)/t6-,7-
O
OH
3)
NH2
O
OH
4)
NH2

Figure 4.11: Two merged example objects containing more than one InChI identifier are shown. They were obtained as result of the CAS merging process. The InChI identifiers and corresponding structures are numbered.

two chiral forms[15] (S- and R-form) of chemical structures,

- Use of wrong chemical structures in a database entry,
- Error propagation: Wrong assignment of the same CAS number to ionized and unionized structures in the utilized resources because this information has already been adopted from other sources,
- The generation of InChI was based on different computer readable structure representations of the same chemical compound, e.g. one contains stereochemical information and another does not.

This finding documents that the utilized resources inconsistently relate InChI identifiers to CAS numbers. There arises the question, how this can happen. Basically, the linkage between CAS and InChI originates from the utilized resources and results from the collection and assembly of entity information provided by the suppliers. Usually, InChI identifiers are automatically generated from structure information represented in computer readable form, such as mol-files. Thus, if there are more than one InChI identifiers newly combined in merged objects, the structure information provided by the database suppliers differ in the

[15] http://www.cem.msu.edu/~reusch/VirtualText/sterism3.htm

merged entries. Another point could be that some of the CAS identifiers were not correctly assigned. To find out which type of InChI differences occur in objects that are related to more than one InChI identifier, a randomly selected portion of 560 merged objects was manually analyzed. The results are depicted in Table 4.13.

Type of InChI differences	Fraction	Merging correct?
Fixed-H layer	54 %	yes
Stereochemical information	25.4 %	yes
Accompanied moieties	8.4 %	yes
Different structure	7.3 %	no
Charge information	5.2 %	yes
Isotope information	0.18 %	yes

Table 4.13: Results of the InChI identifier comparison of 560 merged objects with more than one InChI. The type of difference in the InChI identifiers and its fraction are given. Furthermore, the correctness of the respective CAS merging was assessed according to the object discrimination by the provided synonyms.

The results in Table 4.13 reflect that CAS merging is correct for most objects that differ in the structure by stereochemical, charge and isotope information as well as accompanied moieties. As these closely related chemical entities are in general not clearly distinguishable by synonyms that are provided by the utilized resources, they could be merged. However, 7.3 % of the analzyed fraction correspond to wrongly assigned CAS numbers or InChI identifiers. In the future a procedure should be included that compares also the compound's structure during CAS merging, identifies contradicting CAS assignments to objects and hinders their merging.

Results of the InChI Merging The results of the InChI merging procedure are shown in Table 4.14. It depicts the number of reduced entities for two cases: One represents the resulting entity number when the Fixed H-layer of all InChI identifiers is not removed and the second one the resulting entity number resulting from the InChI merging with InChI preprocessing to remove the Fixed H-layer. Initially, differences in the Fixed H-layer of assigned InChI identifiers were uncovered through the analysis of the CAS merging results. To remove this specific difference in InChI identifiers, an InChI normalization procedure has been integrated into the InChI merging process. However, other layers of the InChI identifiers cannot be removed without loosing information.

Table 4.14 demonstrates that multiple objects from different resources are in general linked to the same InChI identifier. This is illustrated by the fact that even the simple merging without removing the Fixed H-layer already reduces the total amount of objects in the dictionary by 5,977. Furthermore, the removal of the Fixed H-layer reduces the overall object number by 8,247 entities and the amount of ambiguous synonyms by 11,988 terms (~26 %). According to the results of the CAS merging procedure, it points to the fact that different tautomeric structural InChI representations of chemicals are present in the utilized databases.

	Before merging	Simple InChI merging (Fraction in %)	InChI merging with removing the Fixed H-layer before (Fraction in %)
Number of objects in the dictionary	172,992	166,801 (96.42 %)	164,463 (95.07 %)
Number of objects providing InChI	35,153	28,622 (81.42 %)	26,624 (75.74 %)

Table 4.14: The figure shows the total object number and the number of objects only containing InChI identifiers before and after the InChI merging, when using InChI identifiers as they are or when removing the Fixed H-layer from them.

Similarly to the CAS merging an analysis of joined objects the distribution of the number of joined objects was generated. Table 4.15 shows the results.

Number of resources	Number of objects
1	158,497
2	4,127
3	1,292
4	424
5	96
6	20
7	3
8	2
9	0
10	1
11	0
12	1

Table 4.15: Distribution of the number of resulting objects joined from objects of the utilized resources by InChI merging when Fixed H-layers are removed.

As only a small fraction of all dictionary objects contain InChI identifiers most of them were not joined by this procedure. The major number of joined objects consist of two, three and four initial objects. However, striking are new ones which were combined from over 7 objects. They mainly consist of KEGG-C and KEGG-D entries. They are for instance related to different forms of sugar utilized as pharmaceutic aid, phenol, glycol, and polyacrylic acid.

Results of the Synonym Merging It was expected that the join of objects that overlap in their synonyms combines various terminology of chemical entities from different resources that are not linkable by the discussed identifiers. This is a great advantage for successive approaches which rely on the results of Named Entity Recognition. Furthermore, it reduces redundant data and ambiguous synonyms within the dictionary.

Starting point was the dictionary simply concatenated from the seven single curated resource dictionaries. It implies 172,992 objects, 781,524 chemical compound names, and 46,907 ambiguous synonyms. These are about 6 % of the total number of chemical compound names. Ambiguous synonyms that do not belong the class of chemical formulae and respective overlapping objects were collected and grouped. Following basic results were obtained:

- Ambiguous synonyms omitting chemical formulae: 37,978 (~81 % of the total number of ambiguous synonyms)
- Overlapping objects: 44,551.

In a further primary study the number of ambiguous terms present in more than one object was analyzed. This had following reason: When a synonym is unspecificly provided for many chemical entities within the resources and contributes considerably to the synonym overlap between objects, it would be responsible for the join of objects related to different chemical entities. Thus, synonyms which are related to many objects were analyzed in more detail. Table 4.16 provides the results of this study.

Objects	Number of synonyms	Synonym example
≤ 7	37,772	*chloroethene, histidine, 2-(1-methylimidazol-4-yl)ethanamine, 3,3′,4′,5,7,8-Hexahydroxyflavone*
> 7	164	*d-glucuronic acid, gal-alpha1->4gal-beta1->1′cer, n-acetyl-ganglioside gm1, gm 2, cer, glucosylceramide*

Table 4.16: Analysis of ambiguous synonyms (without chemical formulae) which are provided by more than one object in the curated non-merged dictionary. The respective number of ambiguous synonyms residing in ≤ 7 and > 7 objects as well as synonym examples are provided.

According to the assumption made at the beginning of Section 4.1.2.5 it was considered that every utilized resource theoretically contributes one object that is related to one chemical entity. Thus, ambiguous synonyms should originate from 7 objects maximally. However, the analysis provided in Table 4.16 showed something else. There are over 160 synonyms that reside in more than 7 objects. Some of them reside in up to 19 objects (data not shown). The study of their characteristics revealed that they are not simply unspecific abbreviations, but often complex names of chemical entities. Investigation of correspondingly related resources uncovered that most of these terms reside in objects from databases HMDB. Such highly ambiguous chemical names do not differentiate the different chemical entities. Thus, names that were provided by more than 7 objects were not considered for synonym merging. This finally resulted in 44,520 overlapping objects and 46,763 ambiguous synonyms in total that were considered for synonym merging.

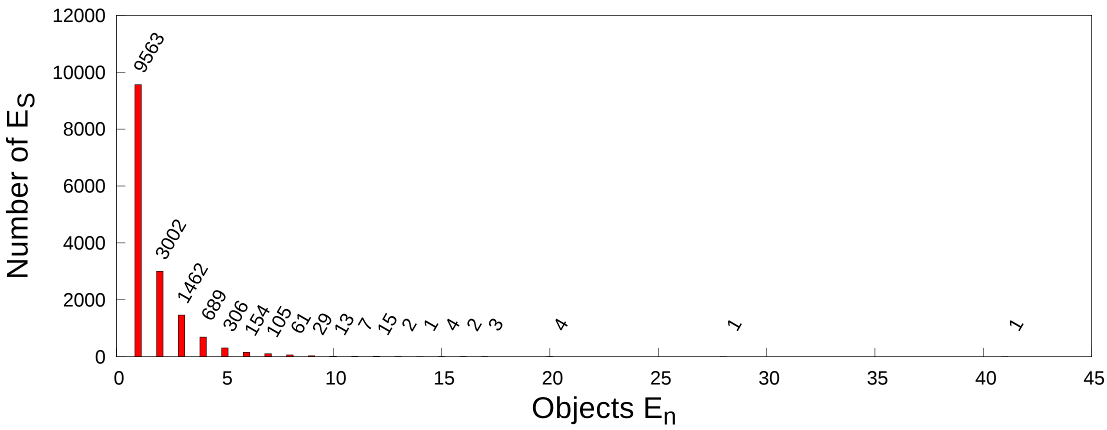


Figure 4.12: Analysis of the object number n of respective objects E_n that overlap with E_S within one group.

In a preprocessing step overlapping objects were collected and combined to 15,467 groups which correspond to a respective set of 15,467 objects defined as standard objects E_S. In a primary study every group was analyzed for the number of E_n objects which overlap with one E_S object per group respectively. Figure 4.12 shows the results. It was observed that standard objects E_S overlap with one, two or three objects E_n in most groups. However, there are 69 groups that are related to E_S objects which share synonyms with 10 to maximal 41 objects E_n. As they could lead to the join of many objects, they were studied in more detail. When all E_n objects of these groups were analyzed altogether, it was observed, that from 219 E_n objects 78 % have a M value of less than 20 % (M was defined in Section 4.1.2.5). Furthermore, the analysis of available InChI and CAS identifiers for respective E_S and E_n of these groups showed that they belong to different chemical entities. Letting them to be merged would join many objects which are related to separate chemical entities. Thus, not all objects partially overlapping in their synonyms should be allowed to be merged without restriction. The constraint C defined in Section 4.1.2.5 is needed to direct the join of object-pairs. It indicates that the synonym overlap of E_n with E_S with respect to the total number of synonyms in E_n has to exceed a given threshold T to allow their join. To determine an appropriate parameter value for T that leads to the best merging result, several experiments were conducted.

For getting an overview on the M values of all overlapping object pairs E_S E_n, the distribution of M values was analyzed. The amount of object pairs E_S E_n possessing the same M values are represented in Figure 4.13.

As it shows, there are many objects E_n which exhibit the same overlap fraction in relation to its total number of synonyms S_{E_n}. This is reflected by an over-representation of objects with distinct M values at 100 %, 50 %, 33 %, 25 %, 20 %, and 16.6 %. The analysis revealed that 27,366 E_n objects have a synonym overlap of less than 100 % and 10,297 objects E_n overlap in 100 % of their synonyms with E_S.

The challenge of the synonym merging was to find an appropriate T threshold that restricts

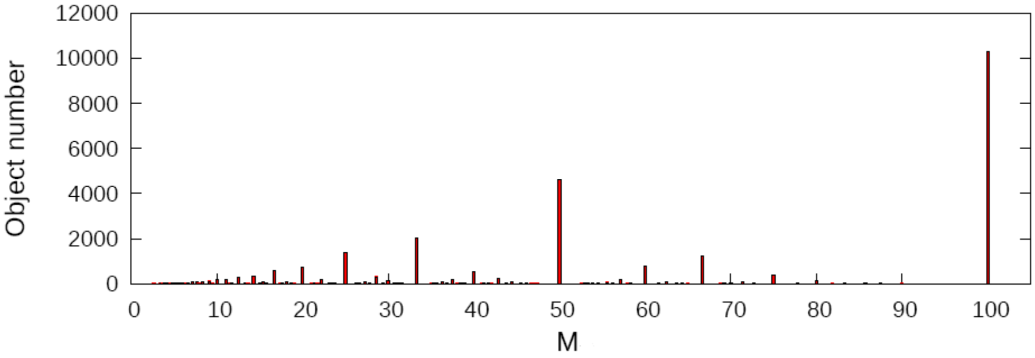


Figure 4.13: Distribution of the number of object pairs E_S E_n in relation to observed M values.

the object merging in such a way that a preferably correct join of objects was performed. As many objects are related to distinct prevalent M values (cf. Figure 4.13), threshold values for T were selected according to the most striking peaks or values close to them. Hence, following T values were chosen: 100, 16, 20, 25, 30, and 40. Most of them are from the low range of found M values. They were selected with the aim to find a low T value that still generates correct synonym merging results.

To get an insight into the quality of the chemical resources the distribution of M values was studied for every single resource on its own with respect to a unique assignment of names to chemical entities or chemical class denominations. Table 4.17 shows the number of synonyms residing within the single curated dictionaries, the ambiguous synonym amount, and the M values for seven ranges.

As can be seen, the MeSH-T dictionary does not provide any ambiguous synonyms. It is followed by the MeSH-C dictionary which comprises very less ambiguous synonyms compared to the remaining ones. In all other resources, except DrugBank, a high number of ambiguous synonyms belong to the class of chemical formulae which were not considered for the merging. However, there are chemical entity names or chemical class denominators which occur in several objects. In the case of KEGG ambiguous synonyms correspond to chemical entities that e.g. belong to the class of sugars. Here, a chemical compound can provide either a ring structure or an open state and are represented by two objects in the database. Other chemicals that contribute to ambiguous synonyms differ slightly in chemical structure (e.g. stereochemical isomers) and are provided as separate objects as well. They possess different InChI and CAS identifiers, so that InChI and CAS merging would not join such entries. Here, often the same chemical names that do not differentiate between specific chemical structure properties are provided by the resource. Hence, when such chemical entities are not specificly named and the synonym overlap between them is high enough, they are expected to be joined through synonym merging. Most ambiguous synonyms were obtained for DrugBank and ChEBI. In DrugBank the same compound can occur in different subsections of the database – in the approved drug part and experimental drug section.

Resource	Ambig. chemical formulae	Ambig. synonyms	Number of E_S E_n pair in defined M ranges						
			0-16	17-20	21-25	26-30	31-40	41-99	100
MeSH-T	0	0	0	0	0	0	0	0	0
MeSH-C	0	23	2	0	0	1	2	4	7
KEGG-C	1,392	28	0	0	0	0	6	8	17
HMDB	454	124	53	6	1	1	12	10	8
KEGG-D	423	139	0	0	0	0	34	93	34
DrugBank	0	315	72	10	3	2	13	65	21
ChEBI	1,665	968	170	99	48	52	152	144	397

Table 4.17: Analysis of ambiguous synonyms residing within the single curated resource dictionaries. The number of ambiguous chemical formulae, ambiguous synonyms without chemical formulae and the number of E_S E_n pairs according to seven M ranges are provided.

Hence, this leads to redundant names within the dictionary. As they provide the same InChI and CAS identifier, they are joined by the InChI merging process, when run before the synonym merging procedure. In ChEBI synonym overlap was often found between objects which describe chemical classes and side groups, elemental, molecular, radical or ionic forms of entities. Furthermore, quite frequently the same synonyms are provided for an entity which is an organic acid, e.g. *'L-Ascorbic acid'* and its ionic form *'L-Ascorbate'*. In both entries *'CHEBI:29073'* and *'CHEBI:38290'* contain the synonym *'L-Ascorbate'*. As already discussed at the beginning of Section 4.1.2.5 and in the CAS merging result part, these closely related chemical compounds are not clearly differentiated by names through the terminology resources. Concluding, such objects should be merged.

Thus, when appropriate merging parameters could be found, object join through synonyms is a promising strategy to combine chemical entities from different resources. Especially in case of those objects that do not provide CAS and/or InChI identifiers. Thus, merging with synonyms was tested for its potential to join entries on chemical entities from different studied resources and to investigate its limitations.

To control the synonym merging of object pairs that are restricted by the C constraint, following test was conducted: For every selected T value 10 object pairs E_S E_n were randomly chosen. The pairs were tested if they belong to the same chemical entity or not. Therefore, it was manually checked whether both chemical entities are related to either the same CAS or normalized InChI identifier. It gives an indication in how far the merging of object pairs connected through the chosen M values is correct or not. Table 4.18 provides the obtained results.

T	Total number of tested pairs	E_S E_n pairs with the same CAS or InChI identifiers (fixed H-layer removed)	E_S E_n pairs with different CAS or InChI identifiers (fixed H-layer removed)
16	10	6	4
20	10	6	4
25	10	6	4
30	10	8	2
40	10	8	2
100	10	9	1

Table 4.18: Comparison of CAS and normalized InChI identifiers related to randomly chosen E_S E_n pairs at selected T values.

The outcome shows that only 60 % of the randomly selected E_S E_n pairs at T of 16, 20, and 25 belong to the same chemical entity. This fraction is 20 % higher at T values of 30 and 40 and reaches 90 % at T values of 100. The latter finding points to a problem in the relation between CAS/InChI and terminology of chemical entities within the studied resources. Thus, it was expected that all objects E_n which share 100 % of their synonyms with E_S would have the same identifiers respectively. Furthermore, the finding leads to the expectation that merging at T values between 16 and 25 would lead to more incorrectly merged objects.

Thus, the simplest strategy would be to merge only those entries that provide a 100 % synonym overlap by setting T to 100. This leads to a reduction of dictionary objects and ambiguous synonym. Furthermore, InChI and CAS identifiers as well as database links are concentrated in fewer newly merged object. However, 95.8 % of these E_n objects possess only one, two or three synonyms; on average they contain 1.4 synonyms. Corresponding detailed data are provided in Table 4.19.

T_{syn}.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Obj. E_n	7596	1870	536	182	53	30	8	11	1	5	0	1	2	1	0	1

Table 4.19: Analysis of the number of synonyms (T_{syn}.) of objects (Obj. E_n) which overlap in 100 % of its synonyms with E_S.

In contrast, objects E_n that only partially overlap with E_S, possess 6.8 synonyms on average. This is 4.9 times more compared to the average of the synonym number of objects E_n which overlap by 100 % with E_S. It shows the tendency that E_n objects with a higher synonym number only partially overlap with E_S objects. This is due to the different amount of synonyms provided by the resources and grades of terminology collection completeness. Thus, a complete overlap does not apply for all E_S E_n pairs. Merging of objects only partially

Iteration step		$T = 16$	$T = 20$	$T = 25$	$T = 30$	$T = 40$	$T = 100$
0	obj.	172,992	172,992	172,992	172,992	172,992	172,992
	a.s.	46,763	46,763	46,763	46,763	46,763	46,763
1	obj.	150,297	150,696	151,354	152,623	154,704	166,299
	a.s.	21,955	20,734	19,875	19,953	21,550	39,839
2	obj.	148,577	149,149	150,000	151,617	154,138	166,283
	a.s.	9,452	9,740	10,964	13,637	18,334	39,817
3	obj.	148,465	148,069	149,943	151,582	154,112	166,279
	a.s.	7,656	8,694	10,390	**13,353**	**18,240**	39,812
4	obj.	148,460	149,066	149,941	stop	stop	166,278
	a.s.	**7,469**	**8,625**	**10,366**			**39,812**
5		stop	stop	stop			stop

Table 4.20: Results of the synonym merging applying different T values. The number of objects (obj.) and ambiguous synonyms (a.s.) obtained in every merging iteration step are given for selected T thresholds. When no merging took place anymore the process stopped. Ambiguous terms remaining after the first iteration step are marked in green, the final object number in blue, and the number of final ambiguous terms in **bold**.

overlapping in their synonyms not only decreases the number of objects and ambiguous synonyms, but it additionally joins terminology that is semantically related to one chemical entity. Therefore, E_S E_n pairs which do not share synonyms by 100 % should also be considered for the synonym merging procedure.

To study the influence of the selected T values onto the join of two objects, six separate synonym merging procedures were performed. These processes generated six differently merged dictionaries respectively. The obtained dictionaries were analyzed with respect to a) the object number and b) the remaining ambiguous synonyms. Since it is an iterative process, the analysis was done after every iteration step. The intermediate and final results related to different T threshold values are provided in Table 4.20.

It was observed that different iteration steps were performed for varying T until the merging procedure terminated. Synonym merging that apply $T = 30$ or $T = 40$ needed three rounds, whereas all others took four iterations until they stopped. In the overall procedure the first iteration round is the most important one. It generates the initial set of merged objects with which further ones are joined in subsequent iteration rounds. When objects are wrongly merged in this first part of the process, errors are propagated until the end of the procedure. An indicator for the merging quality is the number of remaining ambiguous terms that are related to the respective number of generated objects after the first round. It was found that there is a minimum of remaining ambiguous terms at a $T = 25$, although less objects were merged compared to experiments at lower T values. (Ambiguous terms remaining after the first iteration step are indicated in green in Table 4.20.) Compared to $T = 25$ only 78 ambiguous synonyms more remain at $T = 30$ after round one. From this it can be inferred that these two particular merging experiments joined the objects in such a way, that the fewest overlap in synonyms was generated after the first round. At lower T values new merged objects were generated that overlap with additional ones after round one. In contrast, at $T = 40$ and $T = 100$ objects were falsely not merged, leaving ambiguous terms in the dictionary. Finally, most initial objects were joined and most ambiguous terms were reduced when a T threshold of 16 was applied. It is clear that the less synonym overlap fraction M is allowed as merging constraint, the more objects are merged. Thus, the difference in the final number of objects (indicated in blue) and ambiguous synonyms (indicated in **bold**) between low T threshold values and the largest one is very high.

However, the obtained results do not provide information on the correctness of the merged objects. For getting a deeper insight into the merging results, the distribution of the number of original entries newly combined to merged objects was studied. Therefore, the original resource identifiers were counted for every joined object. Furthermore, the fraction of wrongly merged objects that provide both InChI and CAS identifiers was analyzed. It was related to the total number of objects that provide InChI and CAS identifiers for every number of combined objects. This subset of joined objects was chosen, because its analysis could be performed automatically. Since it was decided to join objects according to the list given in Table 4.13, only the first InChI layer with removed accompanied moieties was considered for the evaluation. Therewith, wrongly merged objects consisting of completely different compounds could be found. An object was considered as wrongly merged when it possesses more than one respectively truncated InChI identifier and two CAS identifiers. The latter criterion was chosen because when objects of slightly different structures are allowed to be joined, it is expected that two different CAS identifiers can be found in newly generated objects. The distribution of the final merged objects with regard to the number of combined initial entities (O_R) and the fraction of wrongly merged objects (indicated in grey) are provided in Table 4.21.

As can be seen, most of the final dictionary objects were not merged. They are followed by a high number of joined objects that are a combination of two, three up to 7 single initial objects. The fraction of wrongly merged objects is zero for objects combined from two initial ones for all T values. This seems to be in contrast to the findings in Table 4.18. Manual investigation of the object pairs analyzed in Table 4.18 showed that especially those that were considered as unrelated, were found in merged objects which consist of more than two initial objects. From $O_R > 2$ on the fraction of wrongly merged objects increases with every O_R value. The comparison of the fraction of wrongly merged objects at respective $O_R > 2$ values and different T values revealed that it is low and almost equal for $T \leq 40$ and $O_R \leq 6$. Higher values and larger differences in the wrongly merged object fraction between

O_R	Number of objects (Fraction of wrongly merged objects in %)					
	$T = 16$	$T = 20$	$T = 25$	$T = 30$	$T = 40$	$T = 100$
1	131,720	132,659	133,984	136,534	140,383	160,431
2	9,623 (0)	9,680 (0)	9,713 (0)	9,598 (0)	9,450 (0)	5,074 (0)
3	3,269 (0.37)	3,191 (0.32)	3,106 (0.36)	2,922 (0.28)	2,633 (0.15)	691 (0.57)
4	1,733 (0.87)	1,656 (0.97)	1,543 (1.04)	1,371 (1.03)	1,000 (1.1)	71 (5.63)
5	932 (3.77)	878 (3.77)	782 (4.36)	607 (4.64)	375 (6.13)	7 (14.29)
6	461 (11.09)	414 (9.95)	364 (11.88)	266 (10.90)	158 (11.39)	2 (50.00)
7	269 (20.97)	223 (18.55)	205 (21.67)	142 (27.46)	57 (22.81)	1 (100)
8	156 (16.13)	140 (15.11)	100 (17.00)	68 (25.00)	27 (44.43)	1 (100)
9	92 (38.46)	82 (39.51)	60 (38.33)	32 (34.38)	12 (50.00)	
10	76 (46.05)	51 (43.14)	32 (40.63)	18 (50.00)	9 (77.78)	
11	38 (55.26)	32 (50.00)	18 (55.56)	7 (85.71)	3 (66.67)	
12	21 (66.67)	20 (65.00)	11 (72.73)	5 (80.00)	2 (50.00)	
13	18 (72.22)	12 (75.00)	4 (25.00)	1 (100)	1 (100)	
14	5 (60.00)	3 (100)	1 (100)	1 (100)		
15	6 (100)	3 (100)	3 (66.67)	3 (100)	1 (100)	
16	8 (87.50)	4 (75.00)	4 (50.00)	2 (100)	1 (100)	
17	5 (100)	1 (100)	2 (100)	1 (100)		
18	6 (100)	7 (100)	2 (100)	2 (100)		
19	1 (100)	3 (100)	1 (100)			
20	6 (83.33)					
21	3 (66.67)	1 (100)	3 (100)	2 (100)		
22	1 (100)	3 (100)	1 (100)			
23		1 (100)				
24	1 (100)					
25	1 (100)					
26	1 (100)					
27	2 (100)					
28	2 (100)					
37			1 (100)			
39			1 (100)			
42	2 (100)					
47		1 (100)				
53		1 (100)				
56	1 (100)					
62	1 (100)					
Total	148,460 (2.23)	149,066 (1.79)	149,941 (1.55)	151,582 (1.23)	154,112 (0.76)	166,278 (0.21)

Table 4.21: Distribution of the object number that were combined by the synonym merging procedure in combination with the fraction of wrongly merged objects. The number of objects (O_R) that were included into new objects is given as well as the number of objects with this specific property in relation to different combinations of T values are depicted. The fraction of wrongly merged objects, which possess more than one InChI and two CAS identifiers, is provided in % in grey.

different T values emerge from $O_R \geq 8$. Most newly merged objects that derive from a join of more than 7 initial dictionary objects, exceeding the number of used terminology resources, could be found for T values lower than 100. Its number strongly increases at $T < 30$ and is related to almost 100 % wronlgy merged objects. Corresponding, a high number of different CAS identifiers were combined in these objects that point to unrelated chemical compounds. Thus, T values that lead to a high number of wrongly joined objects should not be considered as merging threshold.

The conducted experiments and the analysis of respective results lead to following conclusion: automated object merging based on synonyms leads to a successful join of objects. The performed studies supported the utilization of threshold $T = 30$. The merging process at this T produced a low number of ambiguous terms after iteration step one and stopped already after the third merging round. Manual investigation of object overlap pairs, provided in Table 4.18, showed a higher object identity compared to lower T values and the same in comparison to $T = 40$. Furthermore, merging at $T = 30$ produced a lower total fraction of wrongly merged objects than experiments with lower T values. These observations lead to the conclusion that $T = 30$ would be an appropriate merging parameter. Synonym merging with this threshold lead to a total reduction by 21,410 objects. This is a descent by 12.38 % in comparison to the initial object number in the curated dictionary. Additionally, ambiguous synonyms were reduced by 33,410. A high advantage of synonym merging is the connection of entries which do not comprise an InChI or CAS identifier with those that provide them. Hence, 7,075 objects from MeSH-C and 1,672 objects from MeSH-T could be connected to InChI identifiers and thus directly to structural information.

There are several challenges and limitations that are related to synonym merging. As it is an iterative procedure the beginning of the merging shapes the whole process because errors made in early steps are propagated. In general, the final results are more laborious to evaluate in comparison to the two other merging strategies. Thus, only a fraction of the merged objects could be analyzed automatically. Similarly to the other two merging strategies, synonym merging is strongly dependent on the quality of the data that is provided by the resource suppliers.

Results of the Merging Workflow The merging workflow combined the three single merging substeps InChI, CAS, and synonym merging with each other. They were performed consecutively according to the workflow shown in Figure 4.10. As synonym merging threshold the determined value $T = 30$ was applied. In compliance with the single merging steps, the distribution of the initial number of objects joined to new ones was analyzed. Furthermore, the fraction of wrongly merged objects was analyzed in the same way as the synonym merging results, described at page 116. The results are given in Table 4.22.

Analog to the separately studied merging approaches most objects origin from one, two, three or four initial objects from the utilized resources. Merged objects that consist of up to 7 initial objects have a low fraction of wrongly merged objects. These values are much lower compared to the single synonym merging results shown in Table 4.21. It demonstrates that the combination of the three single merging procedures yields more correct merging results than the synonym merging alone. Nevertheless, there are 116 objects that consist of more

O_R	Number of objects (Fraction of wrongly merged objects in %)
1	132,253
2	9,871 (0)
3	3,444 (0.03)
4	1,679 (0.18)
5	824 (0.97)
6	348 (3.74)
7	148 (10.14)
8	57 (8.77)
9	29 (20.69)
10	21 (19.05)
11	14 (14.29)
12	8 (12.50)
13	7 (57.14)
14	5 (0)
22	1 (100)
23	1 (100)
26	1 (100)
Total	148,711 (0.41)

Table 4.22: Distribution of merged objects generated from initial objects of the utilized resources joined through the processes InChI merging, CAS merging, and synonym merging combined in the workflow. The synonym merging threshold was $T = 30$. The fraction of wrongly merged objects, which possess more than one InChI and two CAS identifiers, is provided in % in grey.

than 7 initial objects. This are 26 less compared to the synonym merging alone at $T = 30$ and over 100 more in comparison to InChI or CAS merging. Its analysis revealed that 7 % of them consist of objects originating from database HMDB. These objects are related to the same CAS identifiers, but different InChI identifiers respectively. Manual investigation of the other 93 % of the objects showed that most of the combined chemical entities are indeed related. For instance about 96 % of the final merged object that consists of 26 initial ones is related to sugars with 6 carbon atoms. Most of them differ only in charge, stereochemistry, open and closed state in case of sugars or accompanied moieties.

Concluding, the obtained final object number and remaining ambiguous terms were compared between the single merging strategies and the workflow. The intermediate and final results of the complete merging workflow are provided in Table 4.23. The results of the independent single merging processes InChI, CAS and synonym merging are given as well. They are denoted as ‘Single Step’ in Table 4.23.

As the results show, the overall object number was reduced in every step of the merging

	Object number		Number of ambiguous terms	
	Single Step	Workflow	Single Step	Workflow
Curated dictionary	172,992 (100 %)	172,992 (100 %)	46,763 (100 %)	46,763 (100 %)
InChI merging	164,463 (95.07 %)	164,463 (95.07 %)	34,775 (74.36 %)	34,775 (74.36 %)
CAS merging	156,890 (90.69 %)		22,469 (48.05 %)	
Synonym merging	151,582 (87.62 %)		13,353 (28.56 %)	
InChI + CAS merging		154,340 (89.23 %)		18,876 (40.37 %)
InChI + CAS + Synonym merging		148,711 (85.96 %)		10,073 (21.54 %)

Table 4.23: Results of the merging workflow. The number of dictionary objects is provided for the initial concatenated curated dictionary as well as the number of remaining dictionary objects generated by every merging step. The results of every independent merging subprocedure and of single steps of the workflow are listed (indicated in blue). (For synonym merging the threshold $T = 30$ was applied.) Furthermore, the number of ambiguous terms is provided respectively. The fraction of objects and ambiguous synonyms compared to the non-merged dictionary is provided in %.

workflow. Since a low object fraction (~20 %) provides an InChI identifier, only a minor amount of objects could be joined. It lead to a reduction of objects by ~5 % and a decrease in ambiguous synonyms by ~25 %. Since InChI identifiers contain structural information and is the most reliable identifier the workflow has been started with InChI merging. As next step merging was performed through the use of CAS identifiers. It obtained a further object reduction by ~6 % and ambiguous synonyms by further ~34 %. As last step the synonym merging was conducted. It lead to an additional object decrease by factor ~3. Compared to the initial values, the complete merging workflow reduced the overall object number by 24,281 (i.e. 14 %). Only the synonym merging performed as single process could decrease the number of redundant chemical entities in the dictionary in that range alone. Although the difference between the amount of final objects and the object of the unmerged dictionary is relatively low, the overall number of ambiguous synonyms was decreased by a high fraction of ~79 %. This reflects the good performance of the whole merging workflow. The analysis of ambiguous terms that remain after the whole merging procedure revealed that over half of them (56.62 %) belong to the class of chemical formulae. Objects that are linked to the other fraction either do not share InChI or CAS identifiers with other objects, or their synonym overlap with others is too low so that they were not merged. Therefore, the synonym overlap was again analyzed after merging workflow completion. It was found that for a fraction of these objects the chosen synonym merging threshold $T = 30$ was too high. This concerns 38 objects which possess M values between 16 and 29.99. They could be joined by adding a further synonym merging step with a lower T threshold. The other 132 objects which have M values lower than 16 give rise to the fact that their names were wrongly assigned to respective chemical entities by the utilized resources.

In summary, the analysis of the merged objects revealed that object merging by successively applying the three different merging procedures provides a better result than only performing one of them in terms of object number, the number of ambiguous synonyms and correctness of the merging result.

4.1.3 Results of the Chemical Named Entity Recognition

To assess the performance of $\mathrm{ProMiner}_{\mathrm{Chem}}$, the ProMiner version adapted to the chemical domain, it was evaluated on the annotated corpus CHEM-EVAL. Two versions $\mathrm{ProMiner}_{\mathrm{Chem}}$ were generated: $\mathrm{ProMiner}_{\mathrm{Chem2008}}$ comprises the curated and merged dictionary of the terminology from following 2008 resource versions: HMDB, DrugBank, ChEBI, KEGG-D, KEGG-C, MeSH-T, and MeSH-C. This setting was chosen to be able to compare the outcome with the data shown in Section 4.1.2.1 and that were published in [Kolářik et al., 2008]. The second one $\mathrm{ProMiner}_{\mathrm{Chem2009}}$ incorporates the curated merged dictionary from respective chemical data resources version 2009. Their results were compared with the concatenated ProMiner output obtained on CHEM-EVAL with all uncurated raw dictionaries (2008 version) or by omitting the data source PubChem.

Table 4.24 depicts precision, recall and F_1 measure obtained on CHEM-EVAL, whereas two values are given for every measure. The first values were yielded including all chemical name classes that were annotated on CHEM-EVAL, i.e. TRIVIAL, IUPAC, PART, FAMILY, SUM, and ABB. Results shown in brackets were get by leaving out the class PART in the evaluation. These results are presented, because chemical name parts corresponding to class PART, like '3-' are usually not provided by chemical entity databases and will thus not be covered by the dictionary. Only ChEBI contains chemical side group entities which correspond to this class. However, since related names are responsible for false positive partial matchings, these entries were removed by the curation procedure. The resource dictionaries used in the second column correspond to those applied in the generation of $\mathrm{ProMiner}_{\mathrm{Chem2008}}$ and $\mathrm{ProMiner}_{\mathrm{Chem2009}}$. Thus its results were set as standard with which the results of the curated and merged dictionaries were finally compared with. Table 4.24 shows that the curation and merging of the dictionaries lead to a high increase in precision by 55 % and a high improvement of the F_1 measure by 29 % compared to the concatenated result of the raw dictionaries. It demonstrates the success of the performed curation procedure. Nevertheless, the recall was decreased by the curation by 7 %. When entities of class PART were omitted, similar results (shown in brackets) were obtained. However, the recall is only 6 % lower compared to the raw dictionaries. Considering only the comparison of $\mathrm{ProMiner}_{\mathrm{Chem2008}}$ and $\mathrm{ProMiner}_{\mathrm{Chem2009}}$, it can be observed that there is no difference in all evaluation results

	Concatenated ProMiner results of raw dictionaries including PubChem (from Table 4.5)	Concatenated ProMiner results of raw dictionaries omitting PubChem	$\mathrm{ProMiner}_{\mathrm{Chem2008}}$	$\mathrm{ProMiner}_{\mathrm{Chem2009}}$
Precision	0.13 (0.12)	0.15 (0.15)	0.70 (0.69)	0.70 (0.69)
Recall	0.49 (0.50)	0.47 (0.50)	0.40 (0.44)	0.40 (0.44)
F_1 measure	0.21 (0.19)	0.22 (0.23)	0.51 (0.54)	0.51 (0.54)

Table 4.24: Evaluation of the concatenated ProMiner results obtained with all uncurated raw dictionaries or by leaving out PubChem, $\mathrm{ProMiner}_{\mathrm{Chem2008}}$ and $\mathrm{ProMiner}_{\mathrm{Chem2009}}$ on CHEM-EVAL. The result values in brackets correspond to the evaluation that leaves out terms of the chemical annotation class PART.

between the two versions. It leads to the conclusion that only entries were added to the resource versions 2009 which are not relevant for the evaluation of $\mathrm{ProMiner}_{\mathrm{Chem}}$ on the corpus CHEM-EVAL.

For identifying the strength and weakness of $\mathrm{ProMiner}_{\mathrm{Chem2009}}$ according to the six single annotation classes TRIVIAL, IUPAC, PART, FAMILY, SUM, and ABB the recall has been studied for every class separately. Figure 4.14 provides the number of class entities that were manually annotated on CHEM-EVAL. Furthermore, it displays the number of entities recognized by the ProMiner incorporating the uncurated dictionary and $\mathrm{ProMiner}_{\mathrm{Chem2009}}$. Although the dictionary-based approach performs moderate in the recognition of entities from all classes, it works well in the case of entities that belong to the entity class TRIVIAL. $\mathrm{ProMiner}_{\mathrm{Chem}}$ recognizes such entities with a recall of 79 % on CHEM-EVAL. On the contrary, the graphic depicts that $\mathrm{ProMiner}_{\mathrm{Chem}}$ has problems with the recognition of the other classes, especially with names corresponding to the classes IUPAC, ABB, and SUM. The low performance according to class ABB and SUM partially results from the removal of one- and two-letter names by the curation procedure, because they lead to many unspecific matches. Furthermore, 6 abbreviations were filtered through disambiguation after the recognition, because they are highly ambiguous and should only be found when a further synonym occurs within the abstract. The remaining not recognized named entities were not covered by the dictionary.

Comparison of the $\mathrm{ProMiner}_{\mathrm{Chem}}$ Performance with Available Results of Other chemNER Approaches To assess the performance of $\mathrm{ProMiner}_{\mathrm{Chem}}$ it was compared to the output of the approach OSCAR3 [Corbett and Murray-Rust, 2006], the only freely available software for the recognition of chemical named entities, and the IUPAC-tagger [Klinger et al., 2008] developed complementary to ProMiner in-house by R. Klinger. Hence, both could directly be evaluated on the corpus CHEM-EVAL. OSCAR3 finds names of entities which

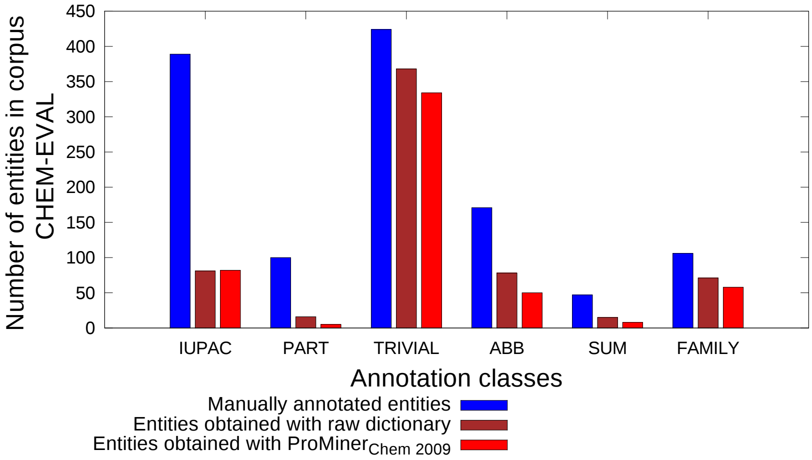


Figure 4.14: Evaluation of the recall according to defined chemical entity annotation classes TRIVIAL, IUPAC, PART, FAMILY, SUM, and ABB. The number of manually annotated entities on CHEM-EVAL is compared with the recognized entities applying ProMiner either with the raw dictionaries or ProMiner$_{\text{Chem}}$.

correspond to six entity classes. For evaluation only entities of the class CM, that marks up recognized chemical entities, were utilized. Furthermore, Hettne et al. [2009] used the corpus CHEM-EVAL for testing and evaluating Peregrine, a dictionary-based approach, which was adapted to the chemical domain as well. The best result obtained by Peregrine with regard to a comparable dictionary composition was included in the approach comparison. Two results are given that are related to the curation of the dictionary and the disambiguation as postprocessing step of recognized named entities. Table 4.25 depicts precision, recall and F_1 measure of all evaluable approaches and the best obtained ProMiner$_{\text{Chem}}$ results on the corpus CHEM-EVAL. The annotation class PART was not considered in the evaluation.

	ProMiner$_{\text{Chem}}$	Peregrine		OSCAR3	IUPAC-tagger
		Curated	Disambiguated		
Precision	0.69	0.55	0.67	0.53	0.71
Recall	0.44	0.46	0.40	0.75	0.35
F_1 measure	0.54	0.50	0.50	0.62	0.47

Table 4.25: Evaluation of ProMiner$_{\text{Chem}}$ comprising the curated and merged chemical dictionary of resource versions from 2009, Peregrine with a curated or curated and disambiguated dictionary omitting PubChem, OSCAR3 and the IUPAC-tagger on the corpus CHEM-EVAL not using annotation class PART for the evaluation.

Comparing firstly the two dictionary-based approaches $\mathrm{ProMiner}_{\mathrm{Chem}}$ and Peregrine separately from the other two, it is obvious that $\mathrm{ProMiner}_{\mathrm{Chem}}$ yielded a higher precision and F_1 measure than Peregrine. Compared to Peregrine using a curated dictionary, $\mathrm{ProMiner}_{\mathrm{Chem}}$ obtained a 14 % higher precision and a 4 % higher F_1 measure. In contrast, the recall achieved by $\mathrm{ProMiner}_{\mathrm{Chem}}$ was 2 % lower. It is remarking, that although Peregrine includes chemical terminology of two more resources, i.e. UMLS[16] and ChemIDplus[17], it could not obtain a higher recall compared to the concatenated results of the raw dictionaries (cf. second column in Table 4.24 for comparison). The approach $\mathrm{ProMiner}_{\mathrm{Chem}}$ corresponds best to Peregrine that applies additionally a disambiguation of the recognized entities. In this case, the precision obtained by $\mathrm{ProMiner}_{\mathrm{Chem}}$ is only 2 % better. However, $\mathrm{ProMiner}_{\mathrm{Chem}}$ provides a 4 % higher recall and a 4 % better F_1 measure than Peregrine. It leads to the conclusion that $\mathrm{ProMiner}_{\mathrm{Chem}}$ includes a better disambiguation strategy than Peregrine. Summarizing the results of both dictionary approaches, they show a similar achievable recall and precision.

In contrast, OSCAR3 has a 16 % lower precision, but the highest recall compared to all other approaches. It shows, that an approach which is not only restricted to a given terminology resource can find a high amount of chemical named entities in text. However, it is on the cost of a much higher number of false positive names recognized and classified as chemical term. Their investigation revealed that OSCAR3 found entities which are no chemical entities at all, like *'vitro'*, *'beta-cell'* or *'Aspergillus flavus'*. The other machine learning-based approach IUPAC-tagger provides a 2 % better precision than $\mathrm{ProMiner}_{\mathrm{Chem}}$, but a lower recall than the two dictionary-based approaches and OSCAR3. As Figure 4.15 on page 125 shows, both the IUPAC-tagger and OSCAR3 are superior in recognizing entities of the class IUPAC. It displays the recall of entities recognized by $\mathrm{ProMiner}_{\mathrm{Chem2009}}$, OSCAR3, and the IUPAC-tagger according to the defined annotation classes. The graph shows that in general dictionary-based approaches provide its best results for entities of class TRIVIAL. In contrast to Peregrine, $\mathrm{ProMiner}_{\mathrm{Chem2009}}$ additionally finds a high number of terms which belong to the FAMILY class.

Complementary, the IUPAC-tagger obtains high numbers of IUPAC-names and entities of class PART. A high recall was found for entities from all classes that were recognized with OSCAR3. However, the drawback of the two mainly machine learning-based methods is that they cannot normalize chemical entities in a straight forward way. Thus they cannot directly map synonyms of one chemical entity to a representative identifier or a chemical structure. Since normalization is of great importance for applications that are built on Named Entity Recognition results, such as information retrieval or relation extraction, it has to be conducted as a separate step. However, until now there is no procedure freely available which can conduct this post-processing in an appropriate way. Klinger et al. [2008] reviewed existing approaches for the transformation of chemical names to structures. They tested OPSIN[18] which is included in the OSCAR3 package and the only available term to structure transformation program for academic use. They came to the conclusion that only 16.24 % of chemical named entities recognized with the IUPAC-tagger and 30 % of 100,000 sampled IUPAC-names from PubChem could be transformed into structures. In contrast, the mapping

[16] `http://www.nlm.nih.gov/research/umls/`

[17] `http://www.nlm.nih.gov/pubs/factsheets/chemidplusfs.html`

[18] `http://oscar3-chem.sourceforge.net`

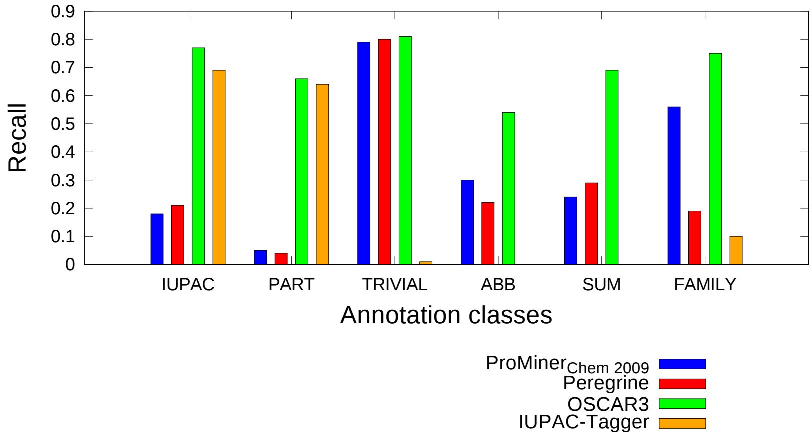


Figure 4.15: Recall of the recognized entities on CHEM-EVAL according to defined chemical entity annotation classes TRIVIAL, IUPAC, PART, FAMILY, SUM, and ABB. The recall is given for ProMiner$_{\text{Chem}}$, Peregrine, OSCAR3, and the IUPAC-tagger respectively.

of names is easy realizable with dictionary-based methods when data resources provide a maximal number of synonyms that is combined with structural information. Hence, dictionary-based approaches have a strong advantage in this concern.

4.1.4 Application of Recognized Chemical Named Entities for Information Retrieval

For demonstrating the benefit of the Named Entity Recognition of chemical compounds in an application, ProMiner$_{\text{Chem}}$ was operated on complete MEDLINE. Subsequently, the identified chemical entities in MEDLINE articles have been added to a text index generated from complete MEDLINE by Lucene. This index has been included into SCAIView for visualizing the recognized entities within the documents. It supports text search and document retrieval as well as knowledge discovery. Thus, the query term entered in the search field of SCAIView defines a specified subcorpus for which the frequency of entities of a specified entity class is calculated and compared to its frequency in whole MEDLINE. From this the relative entropy introduced in Section 3.3 is calculated and provides a ranking of the entities, starting with those that are at most distinctive for the selected subcorpus compared to complete MEDLINE.

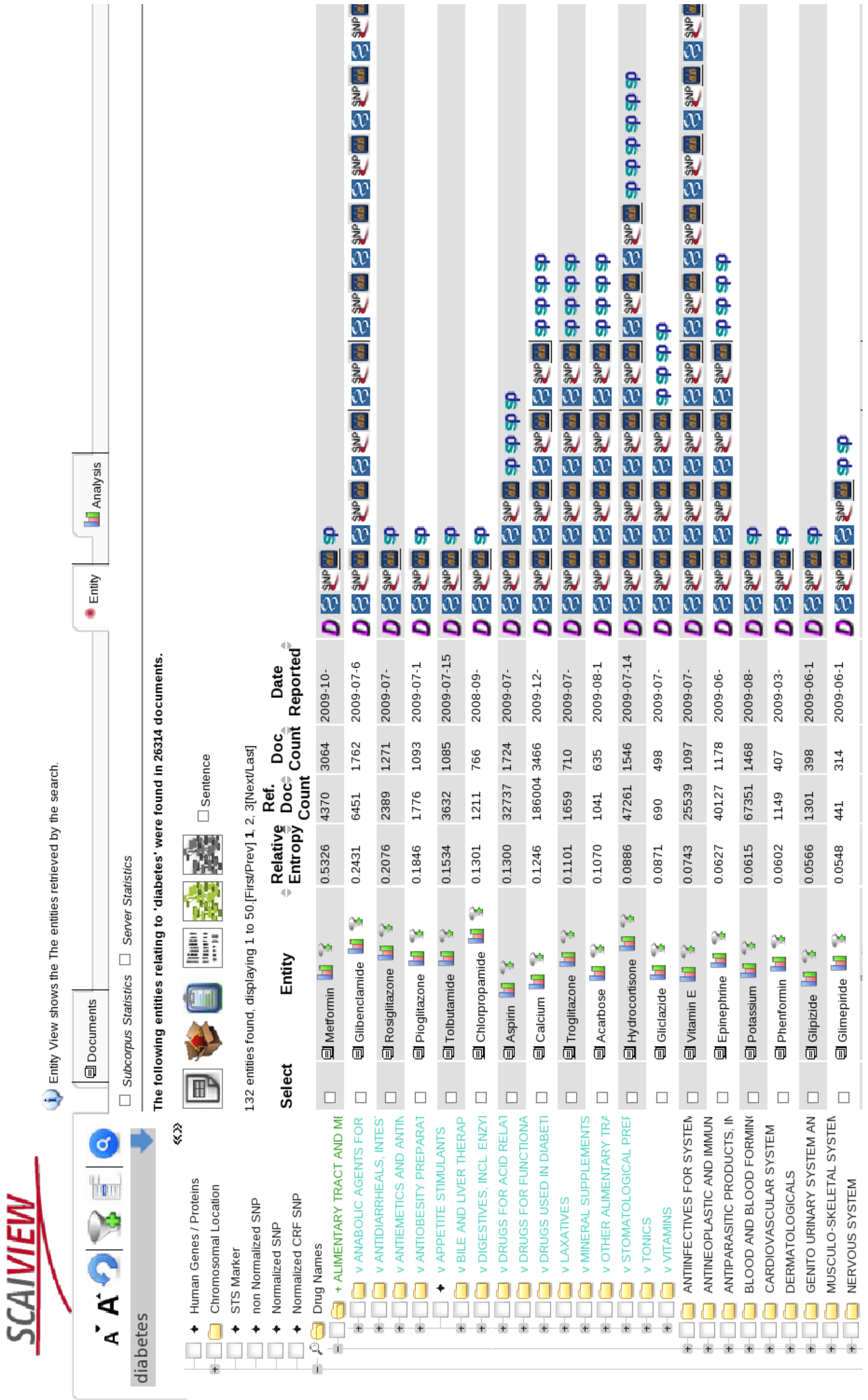
The integration of hierarchical entity classification systems into approaches that enable text search provide a good opportunity for structuring text. Thus, text retrieval at different levels of granularity is supported. Classifications to which chemical compounds, elements and pharmaceuticals are assigned to by functional or structural properties are for instance the pharmacological drug classification system ATC, the ontology ChEBI and the MeSH-T hierarchy. As the resources DrugBank and KEGG relate their chemical entities to ATC, it has been included into SCAIView. Figure 4.16 on page 127 shows the ATC hierarchy and the result of the performed text search. Here, only chemical entities were considered for the query which belong to the branch of *'A Alimentary Tract and Metabolism'*. As they are potentially involved in the development or in the treatment of metabolic diseases they provide a more focused text search than using the complete set of chemical entities.

The top 18 highest ranked chemical entities of the obtained list have been chosen for a deeper analysis. It was figured out that they are strongly associated with the disease *'diabetes'*. Therefore, retrieved articles that are related to every single chemical entity were studied. Table 4.26 shows the selected chemical compounds and its relation to *'diabetes'*.

Chemical compound	Function in relation to *'diabetes'*
Metformin	Antidiabetics, inhibition of gluconeogenesis
Rosiglitazone, Pioglitazone	Antidiabetics, increases sensitivity of liver, muscle and lipidic tissue for insulin
Glimepiride	stimulating the release of insulin by pancreas and increases activity of intracellular insulin receptors
Phenformin	improves insulin sensitivity, withdrawn
Epinephrine	hormone and neurotransmitter, inhibits insulin secretion by the pancreas
Aspirin	Antipyretikum, prevention of angiopathy and coronary heart disease which are secondary malfunctions of diabetes
Calcium	Calcium homeostasis plays role in development of diabetes
Troglitazone	Antidiabetic and antiinflammatory, withdrawn drug
Hydrocortisone	Induces high blood sugar as side effect and can cause complications for diabetics
Acarbose	inhibits enzymes needed to digest carbohydrates, prevents the degradation of complex carbohydrates into glucose
Gliclazide, Glibenclamide, Tolbutamide, Gliclazide, Chlorpropamide	Antidiabetics, stimulate insulin secretion by the pancreas
Potassium	Diabetes is accompanied by potassium deficiency
Vitamin E	Supplement, to prevent angiopathy which is increased in diabetics because of disturbed sugar metabolism and increased oxidative stress

Table 4.26: Top 18 chemical compounds related to *'diabetes'* obtained from SCAIView and their function description.

As was found out, all chemicals highly ranked are related to diabetic conditions. At first sight, most of them are pharmaceuticals directly used for the treatment of diabetes. However, in the list are also the elements *'Calcium'* and *'Potassium'* which are related to diabetes induction or malfunctions. The substance *'Vitamin E'* and the drug *'Aspirin'* are



Entity View shows the The entities retrieved by the search.

□ *Subcorpus Statistics* □ *Server Statistics*

The following entities relating to 'diabetes' were found in 26314 documents.

□ Sentence

- □ ✦ Human Genes / Proteins
- □ Chromosomal Location
- □ ✦ STS Marker
- □ ✦ non Normalized SNP
- □ ✦ Normalized SNP
- □ ✦ Normalized CRF SNP
- Drug Names
  - □ + ALIMENTARY TRACT AND ME
    - □ v ANABOLIC AGENTS FOR
    - □ v ANTIDIARRHEALS, INTES
    - □ v ANTIEMETICS AND ANTIN
    - □ v ANTIOBESITY PREPARAT
    - □ ✦ v APPETITE STIMULANTS
    - □ v BILE AND LIVER THERAP
    - □ v DIGESTIVES, INCL. ENZYI
    - □ v DRUGS FOR ACID RELAT
    - □ v DRUGS FOR FUNCTIONA
    - □ v DRUGS USED IN DIABETI
    - □ v LAXATIVES
    - □ v MINERAL SUPPLEMENTS
    - □ v OTHER ALIMENTARY TRA
    - □ v STOMATOLOGICAL PREF
    - □ v TONICS
    - □ v VITAMINS
  - □ ANTIINFECTIVES FOR SYSTEM
  - □ ANTINEOPLASTIC AND IMMUN
  - □ ANTIPARASITIC PRODUCTS, IN
  - □ BLOOD AND BLOOD FORMING
  - □ CARDIOVASCULAR SYSTEM
  - □ DERMATOLOGICALS
  - □ GENITO URINARY SYSTEM AN
  - □ MUSCULO-SKELETAL SYSTEM
  - □ NERVOUS SYSTEM

132 entities found, displaying 1 to 50.[First/Prev] **1**, 2, 3[Next/Last]

Select	Entity	Relative Entropy	Ref. Doc Count	Doc Count	Date Reported
□	Metformin	0.5326	4370	3064	2009-10-
□	Glibenclamide	0.2431	6451	1762	2009-07-6
□	Rosiglitazone	0.2076	2389	1271	2009-07-
□	Pioglitazone	0.1846	1776	1093	2009-07-1
□	Tolbutamide	0.1534	3632	1085	2009-07-15
□	Chlorpropamide	0.1301	1211	766	2008-09-
□	Aspirin	0.1300	32737	1724	2009-07-
□	Calcium	0.1246	186004	3466	2009-12-
□	Troglitazone	0.1101	1659	710	2009-07-
□	Acarbose	0.1070	1041	635	2009-08-1
□	Hydrocortisone	0.0886	47261	1546	2009-07-14
□	Gliclazide	0.0871	690	498	2009-07-
□	Vitamin E	0.0743	25539	1097	2009-07-
□	Epinephrine	0.0627	40127	1178	2009-06-
□	Potassium	0.0615	67351	1468	2009-08-
□	Phenformin	0.0602	1149	407	2009-03-
□	Glipizide	0.0566	1301	398	2009-06-1
□	Glimepiride	0.0548	441	314	2009-06-1

Figure 4.16: Screenshot of SCAIView showing marked articles that were obtained by a query with the search term *'diabetes'* and chemical entities related to the branch *'A Alimentary Tract and Metabolism'* of the included ATC classification hierarchy.

both utilized as supplements. Both are taken to reduce secondary conditions like oxidative stress which is involved in kidney and liver injuries or to prevent stroke. On the other side, *'Hydrocortisone'* rises the blood sugar level as side effect which is problematic for diabetics.

The brief example shows the advantages of the application of Named Entity Recognition in combination with text retrieval and knowledge discovery. It allows to obtain articles on topics of interest, a filtering and ranking of entities of a certain type which is relevant in the given field. Furthermore, highlighted entities in the text allow to get a fast overview on its content.

4.1.5 Summary of the Chemical Named Entity Recognition and Discussion

The recognition of chemical named entities in text addressed the basic requirement of the information aggregation framework. In the course of conducted experiments a dictionary-based approach ProMiner was modified for its utilization in the context of the chemical domain. The fundamental challenge was the generation of a dictionary comprising denominations of chemical entities, i.e. elements, chemical compounds and structure families. Therefore, eight structured resources, i.e. five databases collecting physicochemical as well as pharmaceutical properties and chemical names, the ontology ChEBI as well as the MeSH thesaurus as well as supplementary material from MeSH have been considered as terminology sources. To get a general overview on the resources, the number of chemical entities as well as the coverage of chemical synonyms have been analyzed with respect to their potential to build up a dictionary of chemical entities. Thus, primary raw dictionaries generated from the eight single resources have been searched with ProMiner only using basic recognition settings. The study of the found chemical entities and its comparison to the annotated entities from the newly generated corpus CHEM-EVAL revealed a potential recall boundary that is achievable with this approach. In addition, basic problems of chemical entity name recognition have been identified, whereas nine error classes were defined. It shows the impact of terminology resources and challenges with which the text mining community is confronted with. Usually, data resources that are allowed to be used for named entity recognition approaches were not specifically generated for it. Their quality is of great importance since dictionary-based approaches rely on terminology that is provided by data resources.

For solving the recognition problems caused by false positive matches, automated dictionary curation – the automated synonym and object processing as well as synonym classification were introduced. Additionally, the modification of the approximate string matching to the chemical domain and its influence onto the search was described. The potential of the precision improvement has been demonstrated. The automated dictionary curation has a high advantage in reducing efforts when the dictionary has to be updated from time to time. It is important since resources exploited for their terminology are extended in their entities and synonyms with the publication of a new version. The curation procedure provides several options for increasing the performance: The removal of synonyms and object as well as a classification of synonyms. The latter one defines the search behavior of ProMiner for synonyms assigned to the diverse classes. In short the tokenization of text and its influence onto the performance was discussed. The final curation setting and utilized modified search capabilities reduced the number of false positive matchings. This resulted in a high increase in precision by 36.38 % for the single dictionaries on average which was related to an acceptable low decrease in recall. It was demonstrated that false positive terms from all error classes could be removed automatedly, except those from class 4 which are responsible for partial term matches.

It was furthermore shown that no single data resource serves for the generation of a final dictionary, because they comprise entities from different subdomains, like metabolomics or pharmacology. Hence, recognition results obtained with single raw dictionaries were analyzed giving priority to seven of them for further use: DrugBank and KEGG-D representing pharmaceuticals, HMDB, KEGG-C, the supplementary material of MeSH (MeSH-C) and ChEBI containing metabolites, nutraceuticals, and small compounds found in living beings as well as MeSH-T and ChEBI mainly covering structure family information. It was expected that their combination would assure the enclosure of many chemical entities from different subfields studied in biomedical research. As resources utilized for dictionary generation overlap in their entities they lead to redundant entries in a simply concatenated dictionary. Hence, they likely contain the same or complementary information in form of chemical objects and names several times. To ensure an easy use of the NER results for information retrieval and other successive approaches relying on textual data, it was required to map the same chemical entities from different resources onto each other. Hence chemical entities from considered resources had to be combined which goes along with a combination of synonyms as well as a reduction of redundant objects within the dictionary and ambiguous terms. Thus, it reduces the size of the final dictionary in comparison to the initial curated one. For object merging three processes have been developed. They are based on the mapping by InChI and CAS identifiers as well as the consideration of synonyms that are shared by chemical entities from the utilized resources. For the first time chemical entities from the seven utilized repositories were joined through the overlap of synonyms. The merging was restricted through the determined threshold T whose best value was 30. To exploit the advantages of all three merging procedures they were integrated into an overall merging workflow. As result a dictionary with a reduced redundancy was obtained exhibiting a decrease in the object number by 24,281 compared to the simply concatenated dictionary that contains 172,992 entries. This resulted in a reduction of ambiguous synonyms by ~79 %. However, the analysis of the single merging results also disclosed problems which come along with the information provided by the resources. Although InChI was introduced as standard for describing chemical structures by a string representation, its use for data mapping provides several pitfalls. Repositories represent the same chemical compound by different InChI representations. Hence, to normalize InChI representations of different tautomeric variants provided for the one chemical compound, the Fixed H-layer has been removed. In the case of CAS identifiers also wrong assignments to chemical entities were found. Furthermore, not for all chemical entities terminology information is strictly used by the considered chemical information resources. This resulted in remaining ambiguous synonyms. Possibly, commercial providers of chemical information offer more correct name-to-structure assignments, however, this type of resource is not allowed to be used for dictionary generation and hence was not analyzed in this work.

Finally, the developed dictionary generation procedure was applied on data resource versions of 2008 and 2009 from which two dictionary versions have been generated. They were separately included into ProMiner, that resulted in two versions $\mathrm{ProMiner}_{\mathrm{Chem2008}}$ and $\mathrm{ProMiner}_{\mathrm{Chem2009}}$ adjusted for the chemical domain. Their performance was evaluated on the newly annotated text corpus CHEM-EVAL. Furthermore, it was compared with the results obtained by concatenation the outputs achieved with the separate raw dictionaries individually integrated into ProMiner. The best performance obtained with $\mathrm{ProMiner}_{\mathrm{Chem}}$ was additionally compared to approaches specialized in the recognition of chemical named entities that were either available or that evaluated their approach on CHEM-EVAL. Thus, OSCAR3, IUPAC-tagger, and results of Peregrine were analyzed. The performance comparison revealed that OSCAR3 is the best approach with an F_1 measure of 0.62, recall of 0.75, and precision of 0.67. It is based on a Naïve Bayes approach utilizing overlapping 4-Grams, a lexicon and rules for the composition of single tokens to a chemical term. In terms of F_1 measure OSCAR3 is followed by $\mathrm{ProMiner}_{\mathrm{Chem}}$, which is 8 % lower. The IUPAC-tagger, relying on the machine learning approach Conditional Random Fields, obtained the highest precision of 0.71. The second best value was achieved by $\mathrm{ProMiner}_{\mathrm{Chem}}$ which is only 2 % lower. As turned out, the two dictionary-based approaches $\mathrm{ProMiner}_{\mathrm{Chem}}$ and Peregrine yielded similar overall results, whereas $\mathrm{ProMiner}_{\mathrm{Chem}}$ obtained a better precision and thus a higher F_1 measure. Entities assigned to the class TRIVIAL were recognized very well by both approaches. It demonstrates the applicability of $\mathrm{ProMiner}_{\mathrm{Chem}}$ for this specific class of chemical named entities.

The results of the NER approach comparison also show that it is difficult to recognize the entire chemical name space with $\mathrm{ProMiner}_{\mathrm{Chem}}$ or Peregrine. Basically, the performance of a dictionary-based approach is dependent on the synonyms that are provided by the utilized terminology resources. Hence, especially novel coined synonyms that were not entered into resources cannot be detected.

For improving the overall recall of $\mathrm{ProMiner}_{\mathrm{Chem}}$ in the future, a further inclusion of publicly available chemical information resources, e.g. ChemIDplus[19], which is a database provided by the National Institutes of Health (NIH), and maybe a part of PubChem might be profitable. Furthermore, a combination of the dictionary-based approach with a technique potent in IUPAC-name recognition, like the IUPAC-tagger published by Klinger et al. [2008], could additionally improve the overall recall and complement the weakness of the chemical version of ProMiner. Taking into account only the longest match of systematic chemical names could also reduce the partial matches of terms from error class 4. However, entities recognized with a Machine Learning approach have to be normalized by an additional procedure. In contrast, entity normalization, the linkage of a named entity to an unique representation like an identifier of a database or ontology or a chemical structure representation, is straight forward with $\mathrm{ProMiner}_{\mathrm{Chem}}$. This is enabled through the storage and linkage of resource, InChI, and CAS identifiers in conjunction with the synonyms of the entities. The inherent normalization is a clear advantage of $\mathrm{ProMiner}_{\mathrm{Chem}}$. On the contrary, OSCAR3 and the IUPAC-tagger, can only recognize chemical names in text, but are not able to normalize it concomitantly. They have to be combined with a follow up method that maps the found chemical names to an unique representation – either an identifier or a chemical structure. However, as was shown in Section 3.1.3.2 normalization of chemical names through its transformation to a chemical structure is a challenging task, which is not yet completely

[19] http://chem.sis.nlm.nih.gov/chemidplus/

solved.

The benefit of the chemical named entity recognition approach ProMiner$_{\text{Chem}}$ was demonstrated in an application scenario. Therefore, ProMiner$_{\text{Chem}}$ was operated on complete MEDLINE. The recognized chemical entities were used to extend a text index generated by Lucene which builds the data basis for SCAIView. In a defined scenario for disease *'diabetes'* the application of recognized chemical entities for knowledge discovery was shown. Thus, the query for this disease was combined with a subset of chemical entities defined by selection of a single class of the drug classification hierarchy ATC included in SCAIView. On one hand side the highlighting of chemical entities besides disease terms, proteins, etc. in text eases the sifting through large amounts of text. On the other side, the potential of Named Entity Recognition results was demonstrated when applied in a knowledge discovery approach which is based on scientific articles.

4.2 Extraction of Function Annotation Information on Chemical Compounds from Text

The identification of principles underlying the pharmacological action, information on side effects, etc. of small molecules provide a basis for the development of new therapeutic agents. Chemical and biological properties as well as pharmaceutical effects are expressed and communicated by natural language in form of terms which reflect the underlying concepts respectively. Some example terms available for aspirin are depicted for clarification in Table 4.27. They comprise for instance information on effects onto protein targets, like the inhibition of an enzyme, or processes which are affected by them.

Chemical compound	Classification terms
Aspirin	*'Anti-Inflammatory Agents, Non-Steroidal'*, *'Anticoagulants'*, *'Cyclooxygenase Inhibitors'*, *'Fibrinolytic Agents'*, *'Platelet Aggregation Inhibitors'*, *'Salicylates'* ATC Codes: A01AD05, B01AC06, N02BA01

Table 4.27: Example classification terms and ATC codes for *'Aspirin'* provided by DrugBank.

Such concept descriptions are provided by various drug classification schemes, like ATC or the Therapeutic Category of Drugs (TCD)[20]. They allow for ordering chemical compounds according to the affected organ systems or effects on specific targets for instance. Furthermore, they enable the establishment of relationships between classes of chemicals. The classification of pharmacological effects requires a substantial understanding of the concepts and their relationships used in the domain of pharmacology.

Likewise Gene Ontology terms are applied for proteins and genes, chemical entities in databases, like DrugBank, are annotated with terms of such classification schemes. However, the annotation of drugs in public databases is far from being complete. Nacher and Schwartz [2008] showed in their work that not all entities of DrugBank are linked to ATC identifiers. Furthermore, classification schemes do not contain all available drug classification concepts, because they have been generated with certain constraints.

In fact, most of the knowledge on pharmaceutical effects is communicated through natural language text. Hence, a high amount of research results concerning drug properties, especially the most recent findings, are available only as unstructured text in scientific publications, patents and drug safety reports. The exploitation of this information resource is therefore of high value. It can be achieved by extracting pharmaceutical properties of compounds from textual resources for annotating chemical compounds with new information and for extending or developing new classification schemes.

To meet this aim, approaches and workflows have been developed for extracting terms from text that describe pharmacological, systemic, chemical or biological properties of chemical compounds. Finally, they were utilized for extending the annotation of chemical entities

[20] `http://www.genome.jp/kegg-bin/get_htext?br08301.keg`

and expanding a classification schemes with further chemical compounds.

4.2.1 Developed Methodology for the Extraction of Function Annotation Information on Chemical Entities from Text

Terms which describe properties of chemical compounds in text often occur in phrases which correspond to hypernymic constitutions, like $NounPhrase_1$ is a $NounPhrase_0$ which corresponds to the example *'Adinazolam is a benzodiazepine derivative'*. Therefore, the extraction of phrases, introduced in Section 3.2, is fundamental for gaining property information from text which is directly related to a chemical compound. This assures the correct assignment of property terms to their corresponding named chemical entities. The perpetuation of the assignment is basically important, because only correct relations allows for the utilization of property terms for annotation purposes.

Hence, the developed term extraction approach is based on work published by Hearst [1992] and Fiszman et al. [2003b,a]. Both made use of the hypernymic proposition as linguistic concept; Hearst [1992] for finding terms to extend machine readable dictionaries like WordNet and the latter ones for providing a method to find and semantically interpret these phrases in biomedical research articles.

4.2.1.1 Description of A New Method for the Extraction of Property Information on Chemical Compounds from Text

The basic aim was to find new property terms related to chemical compounds in text and thus new information applicable for chemical compound annotation, drug classification extension or ontology generation. Hence, similarly to Hearst's technique the developed approach is independent from pre-encoded knowledge on the chemical entities. It means that the related information type on the chemical entities was not specified. Therefore all noun phrases associated with a chemical named entity in a phrase structure following hypernymic phrase constitution have been extracted from text. Figure 4.17 illustrates the complete workflow of the developed drug property term extraction approach.

Phrases representing hypernymic proposition were identified and extracted from substance-specific texts at first (cf. step (1) in Figure 4.17). This is equivalent to Fizman's initial step. Subsequently in step (2), all phrases were filtered to remove those not containing any chemically relevant information. Therefore, the chemical name recognition approach developed in this work and described in Chapter 4.1 was applied. The remaining phrases were split into its fragments and assigned to the classes hypernym and hyponym. Additionally, the information about the conjunction of a hypernym with a certain chemical substance is stored (cf. step (3)). In the last step (4) spelling variants of hypernym terms are removed by a term canonicalization and mapping process.

In the following paragraphs the outlined single steps are discussed in more detail.

Step 1: Extraction of Hypernymic Phrases Biological and pharmacological descriptions of drug effects are usually represented by nested multi-word terms of complex structure. Such terms usually consist of base noun phrases often containing protein names that are

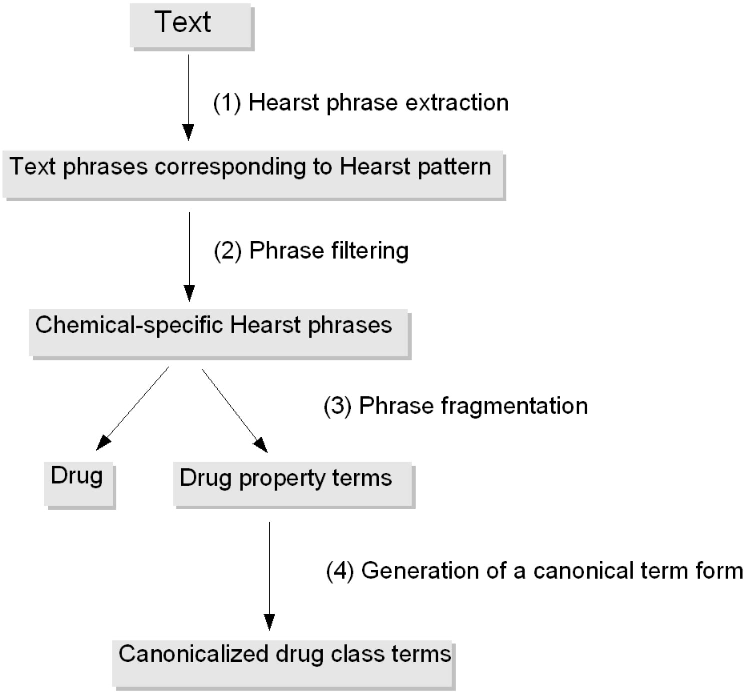

Text
(1) Hearst phrase extraction
Text phrases corresponding to Hearst pattern
(2) Phrase filtering
Chemical-specific Hearst phrases
(3) Phrase fragmentation
Drug
Drug property terms
(4) Generation of a canonical term form
Canonicalized drug class terms

Figure 4.17: Workflow of the term extraction process.

complex by themselves. Furthermore, they comprise inserted numbers, name abbreviations or adjectives written inside or outside of parentheses as can be seen in the following examples:

a) *'competitive beta (1)-selective adrenergic antagonist'*

b) *'angiotensin-converting enzyme (ACE) inhibitor'*

c) *'inhibitor of the cyclooxygenase pathway of arachidonic acid metabolism'*.

d) *'selective high-affinity antagonist of human substance P/neurokinin 1 (NK1) receptor'*

Classical chunkers developed for base noun phrase extraction, introduced in Section 3.2 identify only a part of these complex noun phrases implying pharmacological interesting information. Hence, they miss phrase segments which are fundamental for the meaning of extracted drug related terms. Lets take the example term d). A base noun phrase chunker, like **Analytics** (commercially provided as Skill Cartridge by Temis[21]), identifies only the phrase fragment *'selective high-affinity antagonist of human substance'* and thus loosing information about the protein. Therefore, such incomplete terms would be of no avail for drug annotation.

Since there was no corpus annotated with complex noun phrases available to train a machine learning system and avoid the necessity starting from scratch, an existing noun

[21] www.temis-group.com

phrase chunker was taken as a basis. Furthermore, annotation of such a training corpus requires linguistic experts that were not at hand. For this reason the noun phrase chunker **Analytics** was chosen. Table 4.28 provides examples for nouns, proper names, and noun phrases extractable with **Analytics**.

The software from Temis, that includes **Analytics**, allows a user specific creation of Skill Cartridges with new grammar rules which can easily be incorporated into the already existing system. Hence, noun phrase patterns were extended with additional rules in the new defined Skill Cartridge **ExtAnalytics**. Some example patterns and corresponding complex noun phrases extracted with **ExtAnalytics** are shown in Table 4.28 as well.

Chunker	Pattern	Examples
Analytics	Noun phrase (NP)	*'selective high-affinity antagonist of human substance'*
	NP	*'P/neurokinin'*
	Noun (N)	*'receptor'*
	Proper Name (PN)	*'NK1'*
ExtAnalytics	NP NP Bracket PN Bracket N	*'selective high-affinity antagonist of human substance P/neurokinin 1 (NK1) receptor'*
	N NP Bracket PN Bracket	*'azapeptide HIV-1 protease inhibitor (PI)'*

Table 4.28: Example patterns incorporated into the noun phrase chunkers **Analytics** and **ExtAnalytics** as well as corresponding noun phrase examples.

Similarly to the Skill Cartridge recognizing complex noun phrases, rule sets have been manually established to generate a Skill Cartridge that identifies hypernymic propositions.

Applied Hypernymic Proposition Patterns: Eight lexico-syntactic structures applied in the developed approach of this work were collected from [Cimiano et al., 2005, Hearst, 1992, Rindflesch and Fiszman, 2003] and are described by following patterns:

Pattern 1: NP_1 is (a | an) NP_0

Pattern 2: NP_1 is one of (the | a | an) NP_0

Pattern 3: $NP_1, NP_2, \ldots,$ and NP_n are NP_0

Pattern 4a: NP_0 such as $NP_1, NP_2, \ldots, NP_{n1}$ (and | or) NP_n

Pattern 4b: such NP_0 as $NP_1, NP_2, \ldots, NP_{n1}$ (and | or) NP_n

Pattern 5: NP_0 (including | especially | like) NP_1

Pattern 6: NP_0 for example $NP_1, NP_2, \ldots, NP_{n1}$ (and | or) NP_n

Pattern 7: NP_1, $NP_2, \ldots, NP_n$ (and | or) other NP_0

Pattern 8: NP_1, (a | an) NP_0.

Nominal modifications (introduced in Section 3.2.2.2) have not been applied in this work, since nouns of the last proposition could not be easily assigned either to hypernym or hyponym without semantic analysis. Given that the intention was to find new terms and not searching for predefined ones, they were not used in the developed approach.

To allow for the comparison of the performance of both noun phrase chunkers, they were integrated into two separate Hearst phrase chunkers; one incorporating the original noun phrase chunker, referred to **Analytics-HP chunker**, and the second one contains the extended NP chunker, named as **ExtAnalytics-HP chunker**.

Step 2: Phrase Filtering The obtained phrases are primarily semantically not specified, which means that all in text occurring Hearst phrases were extracted. Since there was an interest only in chemical information, phrases were automatedly filtered so that phrases without drug-specific information were omitted. For this $ProMiner_{Chem}$, was applied to obtain chemical substance related phrases. All phrases containing chemical entity names covered by the Drug Name Dictionary were further processed. Thus, the approach is held generic so that it can easily be adapted to other domains by only exchanging the named entity recognition to focus on various entities of interest.

Step 3: Phrase Fragmentation Following the filtering step, terms describing drug properties were extracted from remaining phrases. The phrases were automatedly split and assigned to their meaning parts, i.e. drug names (the hyponyms) – NP_1, $NP_2, \ldots, NP_n$ and terms describing drug properties or effects (hypernym) – NP_0. Partitioning of the phrase *'Adinazolam is a benzodiazepine derivative'* given as an example would result in: *'Adinazolam'* – a drug and *'benzodiazepine derivative'* – a drug property term. The latter one – the NP_0 of the Hearst phrase – is the term of interest and that is used for further processing and analyses.

Step 4: Generating Canonical Term Forms As described in Section 3.1.3 different variants of terms representing one concept are extensively used in texts, terminologies as well as in databases. To ascertain whether the extracted terms were novel compared to annotation terms available in the database, it was necessary to deal with this difficulty. Examples of occurring term variations are provided in Table 4.29 as well as derivated canonical terms.

Therefore, the available tool Lexical Variant Generator (lvg2006)[22] developed by National Library of Medicine (NLM) was integrated into the workflow to obtain canonical term forms. The following processing steps were applied to each term: First it was tokenized with a Genia tagger-based tokenizer and POS-tagged. Nouns and adjectives lemmatized by the Genia tagger [Tsuruoka et al., 2005] were transformed into a canonical representative form with the UMLS lexical tool Lexical Variant Generator. Syntactic variants of a term, like *'inhibitor of protein synthesis'*, were automatically normalized by a developed heuristic, resulting in *'protein synthesis inhibitor'*. Furthermore, a dictionary of synonymous expressions

[22] `http://www.nlm.nih.gov/research/umls/meta4.html`

and synonyms not covered by lvg2006 was generated for the automated mapping to a canonical term form. They encompass synonymous head nouns typical for the chemical and pharmacological domain, like *'agent'*, *'drug'* or *'compound'*. Examples for semantic equivalent term parts that can differ in the number of words are *'blocker'* vs. *'blocking agent'*. Terms varying in these single words were automatically mapped to each other because they share an equivalent meaning.

Term variation type	Term variations	Canonical term form
Orthographical	*'antiinflammatory agents'*, *'anti-inflammatory agent'*	*'anti inflammatory agent'*
Morphological	*'inhibitor's'*	*'inhibitor'*
Syntactic	*'inhibitor of protein synthesis'*	*'protein synthesis inhibitor'*
Lexico-semantic	*'blocking agent'* *'antihypertensive agent'*	*'blocker'* *'antihypertensive'*

Table 4.29: Term variation types and term variation examples normalized with UMLS lexical tool Lexical Variant Generator (lvg2006) and a developed heuristic.

4.2.1.2 Generation of Evaluation Corpora

Two corpora DRUGBANK-HP and MEDLINE-HP have been generated to assess the extraction quality of the developed Hearst phrase chunkers. Primarily, the two following criteria were applied for the annotation:

1.) Basically, a true positive phrase has to fit syntactically to the given Hearst patterns.

2.) Semantically, the phrase content needs to make sense in a pharmacological way. This means it should be a part of an explicit description providing some biological, chemical or pharmacological properties of a drug or an enumeration of drugs. It does not need to be just a subordinate clause that has no drug property or effect term referring to a drug.

DRUGBANK-HP encompasses phrases from free text fields of DrugBank that contain many sentences with linguistic hypernymic propositions. They describe the mode of action and the pharmacological effect of a substance. The text of these fields has been utilized to generate a text corpus. Half of it was taken for manual annotation of text phrases corresponding to Hearst patterns. This resulted in 572 selected phrases serving as DRUGBANK-HP gold standard.

The second corpus MEDLINE-HP is based on MEDLINE abstracts. Primarily, 1089 abstracts dealing with the pharmaceutical *'Ibuprofen'* were chosen arbitrarily. From these texts

101 Hearst phrases containing pharmacological information on *'Ibuprofen'* were manually extracted.

4.2.1.3 Evaluation of the Hypernymic Phrase Extraction

The analysis of the extracted phrases was done semi-automatically. Phrases not exactly matching the examples in the standard corpus were inspected by hand and were classified into three classes of false positives (FP) defined in Table 4.30. The remaining phrases were considered as true positives (TP). Tables 4.30 and 4.31 show the results of the evaluation on the DrugBank-HP and Medline-HP corpus respectively.

Description	Analytics-HP	ExtAnalytics-HP	Example phrases
Number of automatically extracted phrases	417	500	
True positives	345	451	
(1) FP partial: Phrase is too short compared to standard	65	26	'**clarithromycin, a macrolide** antibiotic'
(2) FP too long: Phrase is too long, at the beginning or at the end	4	19	'**availability of dopamine, a brain chemical**'
(3) FP wrong content: Phrase matches Hearst-pattern, content makes no sense	3	4	'**ibuprofen, a 61-year-old woman**'
False negatives	155	72	
Recall	0.69	0.86	
Precision	0.83	0.90	
F_1 measure	0.75	0.89	

Table 4.30: Results of the Hearst phrase extraction with 'Analytics-HP' and 'ExtAnalytics-HP' on corpus DrugBank-HP (FP = false positives). The extracted example phrases are depicted in bold, whereas the correct phrase section is underlined.

As Table 4.30 illustrates, the application of the ExtAnalytics-Hearst Phrase chunker increased precision, recall, and F_1 measure on DrugBank-HP in comparison to the Analytics-Hearst Phrase chunker. It clearly shows, the extension of the noun phrase chunker leads to an improvement in the identification correctness of the complete Hearst phrase. The extraction of phrases that are too short is based on errors in the Part-of-speech tagging of syntactic ambiguous words. In almost all of these cases words were detected as adjectives which are in the given pharmaceutical context nouns, e.g. the word *'antibiotic'* in *'macrolide antibiotic'* or *'analgesic'* in *'synthetic opioid analgesic'*. According to the given patterns, the last word in a phrase has to be a noun and hence the extension of the phrase is stopped too early. This results in *'macrolide'* and *'synthetic opioid'*

Explanation	Results
Number of manually annotated phrases	101
Total number of automatically extracted phrases	108
True Positives	79
(1) FP partial	2
(2) FP too long	18
(3) FP wrong content	9
False Negatives	2
Recall	0.73
Precision	0.97
F_1 measure	0.83

Table 4.31: Evaluation of the ExtAnalytics Hearst Phrase chunker on MEDLINE-HP (for descriptions of false positives (FP) variants cf. Table 4.30.)

Table 4.31 provides the phrase recognition results on the MEDLINE-HP corpus. Compared to the DRUGBANK-HP corpus a higher recall, but a lower precision was achieved. The proportion of true positives recognized in DRUGBANK-HP is similar (~78 %). It turned out that a lower fraction of FP partial was extracted, but a higher amount of FP too long and FP with wrong content. It shows that the two corpora differ in their term characteristics and content.

4.2.1.4 Results of the Function Annotation Term Extraction Procedure

In the following, terms extracted for 11 drugs, that are mentioned in the context of various therapeutic areas, were evaluated. Table 4.32 provides the drug list. They have been arbitrarily selected with the solely constraint that for all of them a considerable number of MEDLINE abstracts (>4200) was retrieved. It was set to assure the analysis of an adequate number of excerpted Hearst phrases.

For each of the 11 drugs the term extraction procedure was done separately. Table 4.32 lists the number of obtained Hearst phrases, unique terms (all redundant terms were removed), and unique normalized terms. It shows, that a high amount of Hearst phrases and potential drug annotation terms could be extracted from MEDLINE abstracts. Canonicalization of terms reduced the number by 16 % on average. The ratio between the extracted Hearst phrases and canonicalized terms is similar for the selected drugs and lies between 1.58 and 2.08.

Drugs	Hearst phrases	Number of extracted unique terms	Number of unique terms after canonicalization
Ibuprofen	597	404	329
Diclofenac	309	203	156
Atenolol	235	175	136
Mitomycin	320	223	203
Metoprolol	207	152	125
Tamoxifen	1003	635	527
Ciprofloxacin	529	363	316
Nifedipine	975	601	507
Chlorpromazine	427	320	271
Phentolamine	321	186	154
Midazolam	403	284	252

Table 4.32: Number of Hearst phrases extracted from MEDLINE, unique terms, and normalized terms for 11 drugs.

4.2.1.5 Summary of the Developed Workflow for the Extraction of Function Annotation Terms and Discussion of the Results

As could be shown in this work, Hearst phrase extraction in combination with the recognition of chemical names is a valuable approach to find new annotation terms that are not yet applied for chemical compounds in databases like DrugBank or drug classification schemes like ATC.

The developed Hearst phrase extraction approach was driven to obtain a high performance. It was achieved by an extension of patterns used for the recognition of noun phrases going beyond basic noun phrases and its incorporation into the Hearst phrase chunker. This resulted in a high F_1 measure of 0.89 on the DrugBank test set and 0.83 on an arbitrarily chosen test set from MEDLINE. It lead to an improvement of 14 % in F_1 measure compared to a Hearst phrase chunker that is based on patterns ready to identify basic noun phrases.

Compared to SemSpec developed by Fiszman et al. [2003b] the results obtained by the presented approach on the DrugBank standard corpus (90 % precision and 86 % recall) and on the MEDLINE standard corpus (97 % precision and recall 73 %) were higher, both in recall and precision. Fiszman et al. [2003b] reported a precision of 83 % and recall of 69 %. It can be concluded that the newly introduced system has a better recognition performance for Hearst Phrases on MEDLINE abstracts than SemSpec. However, a direct one-to-one comparison is difficult, because the patterns for hypernymic proposition recognition used by SemSpec was not completely explained and thus may differ. Furthermore, the two systems were evaluated on two different text corpora.

For 11 selected drugs the presented term extraction procedure has been applied on MEDLINE. Many terms could be obtained from the Hearst phrases for every single drug. Through the canonicalization of term variations the overall term number could be reduced by 16 % on average. This is an important aspect when dealing with textual data.

4.3 Developed Framework for Information Aggregation and Annotation of Chemical Compounds

Function annotation of chemical entities in databases is done by assigning concept denominations from predefined classification schemes or ontologies in general. Providing annotations to chemical is related to the parsing of these structured resources and assigning the obtained identifiers to chemicals. However, it only allows for incorporating well established property information. In addition, Nacher and Schwartz [2008] illustrated that even classification schemes, like ATC are not complete. Hence, providing chemical entities with new annotations, other resources like recent publications or reports have to be read and respective information has to be extracted. To support the process of extracting potentially new annotation information from text and its comparison to annotation information available in structured resources, a new approach has been developed.

4.3.1 Description of the Workflow for Finding New Function Annotations of Chemical Entities in Text

The complete process basically relies on a given annotation terminology which is compared to potentially new annotation terms extracted from text, thus building on the workflow previously described in Section 4.2.1.1. Although the conception of the process is generally applicable, two main resources have been considered for its design in this work; the database DrugBank and the bibliographic database MEDLINE which contains over 17 million scientific articles. DrugBank was chosen because it provides a database field comprising function annotations of chemical compounds as well as effect and mechanism's descriptions in form of complete sentence descriptions embodied in free text fields.

In the course of the developed workflow annotation terminology which is related to chemical entities from DrugBank was extracted and compared to property terms specifically extracted for these entities from free text fields of DrugBank and MEDLINE by the methodology described in Section 4.2.1.1. Figure 4.18 provides an overview on the developed workflow.

In the following the three single steps depicted in Figure 4.18 are explained in more detail:

- **Step 1:** Property terms were extracted from MEDLINE titles and abstracts and canonicalized corresponding to the new method introduced in Section 4.2.1.4. They were named as MEDLINE-Text Terms. Furthermore, the developed term extraction methodology was applied on DrugBank free text fields, whereas the canonicalized term set is named DrugBank-Text Terms.

- **Step 2:** Annotation terms and ATC identifiers have been extracted from DrugBank from respective database fields. Additionally, ATC identifiers were mapped to their corresponding terms provided by the WHO. Finally, all terms were transformed to canonical term forms building the term set called DrugBank-Annotation Terms.

- **Step 3:** MEDLINE-Text Terms and Drug-Bank-Text Terms were both compared to DrugBank-Annotation Terms using ProMiner. Therefore, term sets extracted either from MEDLINE or DrugBank text were separately used as a dictionary incorporated

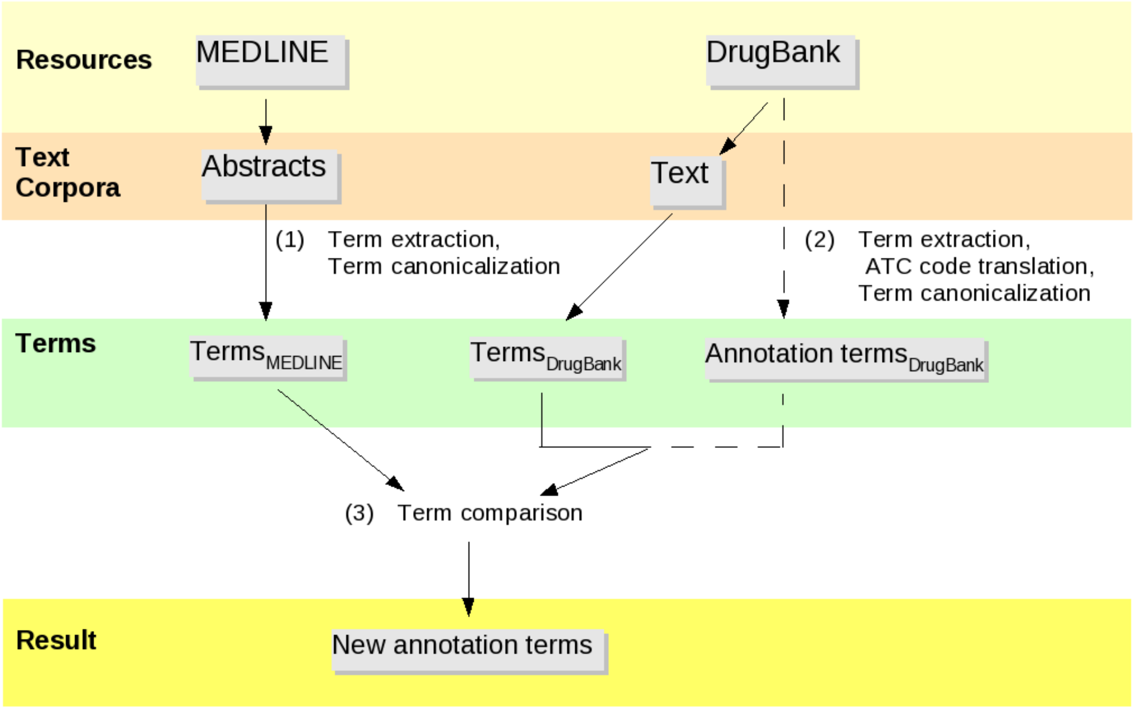
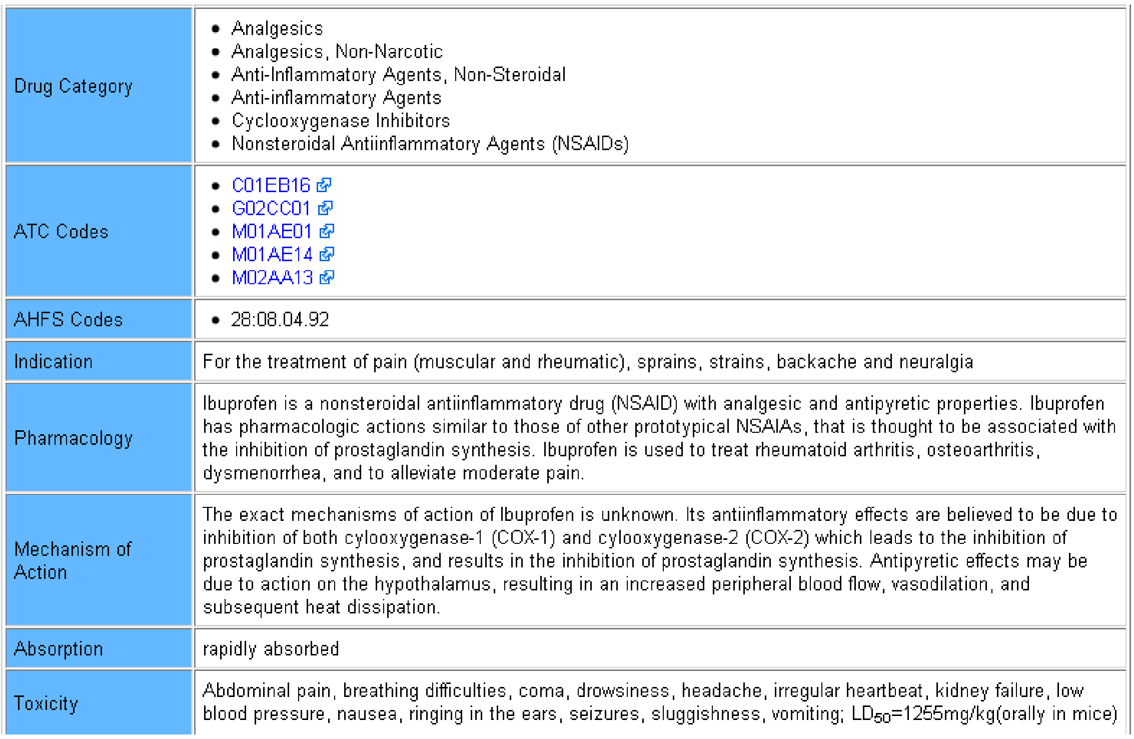

Resources
MEDLINE
DrugBank
Text Corpora
Abstracts
Text
(1) Term extraction, Term canonicalization
(2) Term extraction, ATC code translation, Term canonicalization
Terms
Terms$_{MEDLINE}$
Terms$_{DrugBank}$
Annotation terms$_{DrugBank}$
(3) Term comparison
Result
New annotation terms

Figure 4.18: Developed workflow of the term extraction and comparison process. The three process steps are enumerated and are described in more detail in Section 4.3.1.

into ProMiner. It was used to search for corresponding terms in the list of canonicalized DrugBank-Annotation Terms. Those terms of the sets MEDLINE-Text Terms DrugBank-Text Terms that were not found in the DrugBank-Annotation Term list were considered as potentially new annotation terms. They have been manually evaluated to analyze their applicability as new drug classification/annotation terms. A constraint that needed to be fulfilled for defining a term as novel drug classification/annotation is the following: A term should contain relevant pharmacological, biological effect or chemical property information on a drug.

In the previous Section 4.2.1.4 it was shown that a high number of terms were extracted from MEDLINE for selected drugs, for which an at least moderate number of articles have been published. It demonstrates that text is a valuable concept term source. However, it does not give evidence that these terms are relevant for different scenarios i.e. chemical entity annotation or other applications, like classification system expansion. Hence, experiments had to be conducted to test the information quality and applicability of the extracted and canonicalized terms. At first place the annotation of chemical entities in DrugBank has been studied.

In Section 4.2.1.4 it was shown that the number of the extracted terms from MEDLINE was quite high and the average number of annotation terms in DrugBank is four to five. For finding out whether the remaining terms comprise new information, they had to be



Drug Category	• Analgesics • Analgesics, Non-Narcotic • Anti-Inflammatory Agents, Non-Steroidal • Anti-inflammatory Agents • Cyclooxygenase Inhibitors • Nonsteroidal Antiinflammatory Agents (NSAIDs)
ATC Codes	• C01EB16 • G02CC01 • M01AE01 • M01AE14 • M02AA13
AHFS Codes	• 28:08.04.92
Indication	For the treatment of pain (muscular and rheumatic), sprains, strains, backache and neuralgia
Pharmacology	Ibuprofen is a nonsteroidal antiinflammatory drug (NSAID) with analgesic and antipyretic properties. Ibuprofen has pharmacologic actions similar to those of other prototypical NSAIAs, that is thought to be associated with the inhibition of prostaglandin synthesis. Ibuprofen is used to treat rheumatoid arthritis, osteoarthritis, dysmenorrhea, and to alleviate moderate pain.
Mechanism of Action	The exact mechanisms of action of Ibuprofen is unknown. Its antiinflammatory effects are believed to be due to inhibition of both cylooxygenase-1 (COX-1) and cylooxygenase-2 (COX-2) which leads to the inhibition of prostaglandin synthesis, and results in the inhibition of prostaglandin synthesis. Antipyretic effects may be due to action on the hypothalamus, resulting in an increased peripheral blood flow, vasodilation, and subsequent heat dissipation.
Absorption	rapidly absorbed
Toxicity	Abdominal pain, breathing difficulties, coma, drowsiness, headache, irregular heartbeat, kidney failure, low blood pressure, nausea, ringing in the ears, seizures, sluggishness, vomiting; LD_{50}=1255mg/kg(orally in mice)

Figure 4.19: Screenshot of a part of a DrugBank entry for the drug *'Ibuprofen'*. The extracted database fields *'Drug Category'*, *'ATC Codes'*, *'Indication'*, *'Pharmacology'*, *'Mechanism of Action'* and *'Toxicity'* are depicted.

manually inspected. Hence, the term comparison task was exemplarily accomplished only for the 11 selected drugs from DrugBank already introduced in Section 4.2.1.4.

In the following the results at separate steps of the workflow described in Section 4.3.1 are explained. At first the three term sets generated during steps 1 and 2 and their characteristics are explained in more detail.

- **DrugBank-Annotation Terms:** Chemical entities provided by databases like DrugBank are related to pharmacological annotations in form of terms and/or identifiers of classification systems. DrugBank contains identifiers of the ATC classification scheme and additional non-systematic pharmacological class terms. An example section from DrugBank is shown in Figure 4.19. The analysis of the available annotations is depicted in Figure 4.20 on page 145. It provides the distribution of the number of annotations per drug. As can be seen, a large quantity of drugs are annotated with four or five drug category terms.

  1073 annotation terms and ATC identifiers were extracted from the database fields *'Drug Category'* and *'ATC Codes'* of all approved drugs of DrugBank. Identifiers of the ATC classification system used for chemical entity annotation, were automatedly mapped to their corresponding terms provided by the WHO. To be sure to work with representative invariant terms all annotation terms were canonicalized by the process

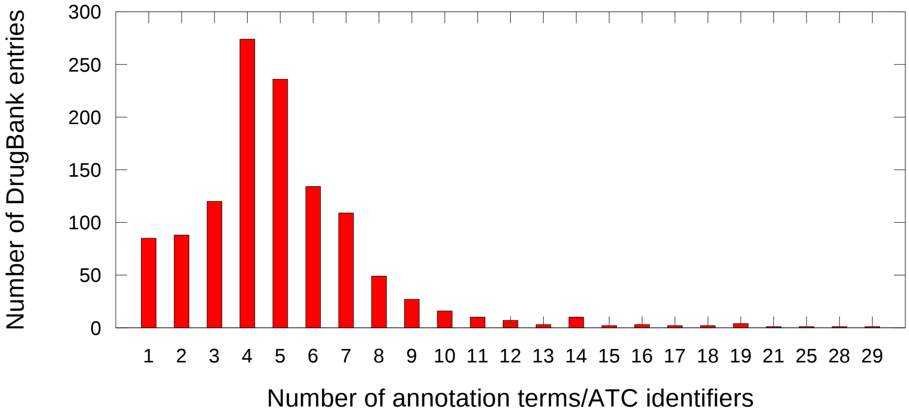


Figure 4.20: Distribution of the number of annotation identifiers or terms assigned to drugs in DrugBank.

developed for canonicalizing terms from text. This procedure reduced the overall number of annotation terms to 966, which is a decrease by about 10 %. It was detected that different term variants of one drug category were assigned to several drugs in DrugBank. They are of morphological, orthographical and lexico-syntactic type. It accounts that even terms used for annotation vary in databases.

- **MEDLINE-Text Terms:** The term set obtained from MEDLINE corresponds to the term extraction results described in Section 4.2.1.4.
- **DrugBank-Text Terms:** To assess the information content of terms used in free text fields of DrugBank, the complete free text of following fields has been extracted from DrugBank: *'Indication'*, *'Pharmacology'*, *'Mechanism of Action'* and *'Toxicity'*. This corpus was subjected to the term extraction pipeline, whereas the assignment to the database entities was stored. A total of 1164 Hearst phrases containing drug names were automatically obtained from the entire DrugBank text corpus. They comprise 860 terms, which were reduced to 829 after canonicalization. This shows that, even in database text, different term variants are used for the description of the same concept. As result a second drug specific term set obtained from DrugBank text was generated. It turned out that for most approved drugs one or two terms were extracted from DrugBank text.

4.3.1.1 Results of the Term Comparison Procedure

The term comparison was performed with three experimental settings.

(1) In a first experiment the overlap between the DrugBank-Annotation Term set, the MEDLINE-Text Term set and DrugBank-Text Term set has been studied. Therefore, DrugBank-Annotation Terms and terms derived from the DrugBank and MEDLINE text corpora were compared. Three classes were defined to which the DrugBank-Annotation Terms can be assigned to: 'Terms only used as DrugBank-Annotation Terms', 'DrugBank-Annotation Terms

also found in DrugBank text', and 'DrugBank-Annotation Terms also found in MEDLINE'.

(2) In a second experiment the overlap between the DrugBank-Text Term set, the DrugBank-Annotation Term set, and the MEDLINE-Text Term set has been studied in a similar manner as above. Again three classes were defined like above. The analysis of the result shows that 84 % (694) of the total number of DrugBank-Text Terms have not been used as drug annotation terms so far within DrugBank. They contain pharmaceutically relevant information on drugs. Most of the terms give more detailed information than the provided annotation terms, e.g. about the protein that the drug influences or the mechanism of the drug effect, e.g. *'irreversible proton pump inhibitor'*. Other new terms describe the natural resource of the drug or even a new reaction mechanism of the drug onto a protein target (*'histamine h2 agonist'*).

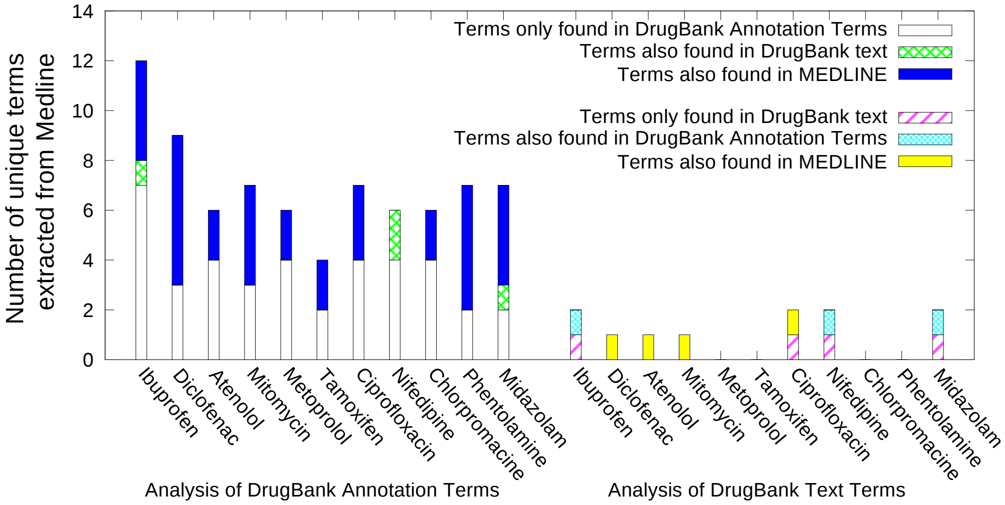


Figure 4.21: Analysis of drug property-describing terms from DrugBank-Annotation Terms and DrugBank-Text Terms compared to MEDLINE Text Terms. The left side shows the comparison of DrugBank-Annotation Terms with the other two term sets. The right side depicts the analysis of DrugBank-Text Terms compared the remaining two.

The diagram of Figure 4.22 shows the detailed result of both experiments for the defined 11 drugs, whereas the terms are represented by the three described classes. The left side presents the first experiment, whereas the right side the second one. This experiment shows that even database text contains terms that have the potential to be applied as annotation terms. Furthermore, they can add additional information to a terminology or classification system. Figure 4.22 also demonstrates that most of the DrugBank-Text Terms are either available in MEDLINE or in the DrugBank-Annotation Term set.

(3) In a third experiment MEDLINE-Text Terms have been compared with DrugBank Annotation Terms for each of the 11 drugs separately. The results depicted in Figure 4.22 illustrate that only a limited number of them overlap. Only 1.3-6.4 % are already in use in the DrugBank annotation field. The remaining terms were checked manually to ascertain

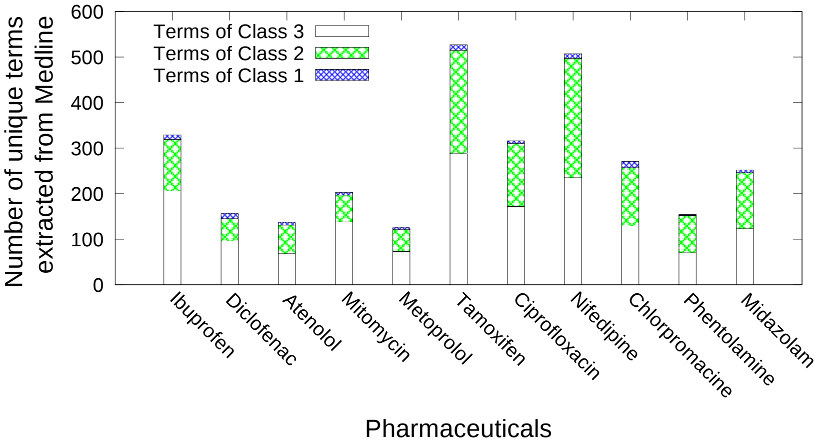


Figure 4.22: Terms extracted from MEDLINE and their assignment to three classes (a more detailed description is provided in the text):
Class 1: Terms available in DrugBank Annotation Terms and DrugBank text,
Class 2: New drug annotation terms and
Class 3: Terms not usable for drug annotation.

its novelty. It turned out, a high portion – 29-53 % – of terms extracted from MEDLINE abstracts could serve as annotation terms and have not been used in DrugBank so far. This means that they would add new information to the database if they were used for drug annotation. A deeper analysis of the valid new terms shows that they can be assigned to various drug property classes. A list of classification types and term examples from DrugBank and MEDLINE abstracts is given in Table 4.33. Analyzed terms that were not considered as new, either originate from false positive Hearst phrases with wrong content, from too long phrases that incorporate a term already existing in the DrugBank Annotation Terminology, or they contain non-relevant additional information about a drug.

It became apparent, the ATC drug classification schema as well as the internal annotation types of DrugBank is restricted to some drug property classes, like pharmacological property or chemical structure classes. As can be seen in Table 4.33, some of the new terms found in MEDLINE can be assigned to new annotation categories not contained in the DrugBank annotation terminology. With that not only additional drug property terms were found in text, but also new classes of information. Furthermore, new pharmacological concepts not contained in ATC have been extracted from the MEDLINE text corpus. Some examples are listed in Table 4.34.

These new terms usually contain descriptions of a compound's effects on a certain protein like the inhibition of an enzyme or agonistic action on a receptor. With regard to the term content, they correspond to concepts of the ATC scheme at level 4. It demonstrates that the drug classification by this system is limited and not comprehensive which constricts its use for diverse research applications. Hence, the extracted drug classification concepts could be used to augment the ATC scheme.

Classification	Terms in DrugBank	New terms from MEDLINE
Pharmacological property	*'antipyretic agent'*	*'radiosensitizer'*
Chemical structure class	*'methylhydrazine'*	*'flurbiprofen derivative'*
Effect on biological processes	*'antiperistaltic'*	*'nf kappaB activation inhibitor'*
Effect on protein	*'neuraminidase inhibitor'*	*'cytochrome P-450-monooxygenase inhibitor'*
Chemical property	—	*'racemic drug'*
General molecular effect	—	*'free radical scavenger'*
Biotransformation	—	*'short half life drug'*
Biological resource	—	*'rheum palmatun anthraquinone component'*
Combination	—	*'aspirin like anti inflammatory drug'*

Table 4.33: Drug classification categories with example terms from DrugBank and MEDLINE.

4.3.2 Application of the Information Aggregation Framework: Drug Classification Schema Extension

Classification systems of medical compounds are generated on the basis of efficacy studies, mechanism of action, clinical outcomes and market strategies and are maintained by pharmacopeias of different countries. Usually they assign drugs to classes of several hierarchical levels reflecting its pharmaceutical, therapeutic or other properties. They are applied in drug utilization studies, as annotations of chemical compounds in databases as well as for the clear organization of pharmaceuticals on web interfaces to databases, etc. The classification system most prevalently applied is the Anatomical Therapeutic Chemical (ATC) classification system maintained by the World Health Organization (WHO). This scheme hierarchically classifies drugs providing five different levels of granularity, i.e. the organ system they act on as well as therapeutic, pharmacological, chemical properties of drugs, and the drugs. Although the classification system is an international standard for drug utilization studies, it does not contain the entirety of drugs, because it considers new entries only upon requests made by manufacturers, regulatory agencies or researchers[23]. Therefore, the system does not include substances for which no requests have been made or withdrawn

[23] http://www.frma.org.au/atc/maintainence.htm

Pharmacological concepts not contained in ATC
'Angiotensinogen inhibitor'
'Organic Anion Transport Protein inhibitor'
'Alpha-1-acid glycoprotein antagonist'
'voltage-gated potassium channel inhibitor'
'voltage-gated sodium channel antagonist'
'DNA (Cytosine-5-)-Methyltransferase inhibitor'
'Membrane stabilizer'
'Cytochrome P-450 CYP2D6 substrate'
'Alpha 1A adrenergic receptor agonist'
'Dopamine beta hydroxylase agonist'
'Alpha 1B adrenergic receptor agonist'

Table 4.34: Denominations of concepts extracted from the MEDLINE text corpus which are not covered by the drug classification scheme ATC.

drugs.

When ATC is used in network studies, this problem becomes more evident. For instance there is no possibility to learn from drugs that are not on the market anymore. Nacher and Schwartz [2008] realized that 138 drugs of DrugBank do not contain ATC identifiers and hence could not be integrated into their drug-therapy network study. Their result was an incomplete network making clear that such studies highly depend on information contained in resources used, which might limit their value. This case points to a clear drawback of ATC.

Hence, generally the incorporation of drug instances into classification systems like ATC is of high potential, because it helps to close the information gap for allocating more information to applications that are built on such schemes. A way to accomplish this task is to predict respective scheme classes of pharmaceuticals or chemicals not represented by the systems. Thus, for internal research purposes, they can be included in an automated way independently from suppliers like the WHO.

4.3.2.1 Description of a New Method for the Extension of Classification Schemes by New Drug Instances

For the prediction of ATC classes of drugs the classification technique was utilized. There are several established algorithms and implementations available that can be applied and compared with each other. Classifiers are dependent on features that represent the data of the different classes. The basically new idea was to apply property concepts on chemicals which were extracted from text by the methodology described above as class features. They were allocated to ATC classes through chemical compounds that are already assigned to them.

This work was done in cooperation with H. Gurulingappa whose master thesis [Gurulingappa, 2008] was supervised by me. The underlying conception was developed by me, whereas Gurulingappa was responsible for the application of selected methods as well as the development of conception details, i.e. additional term normalization, classifier testing and evaluation was conducted by Gurulingappa. The description of the applied methods can be found in [Gurulingappa et al., 2009]. Here just a brief outline of the methodology is given, whereas the focus lies on the obtained results to demonstrate the potential of the concept denominations extracted from text.

For experimentation the classification was restricted to a subset of ATC classes to diseases of the cardiovascular system. The super class *'Cardiovascular Agents'* of ATC contains 400 drugs which constitute 13 % of the entire drugs of this scheme. It implies 38 therapeutic/pharmacological classes at level three and 118 pharmacological/chemical classes at level four. From the third level 21 ATC classes related to 390 drugs were chosen as training set. On average they contain 10 drugs per class. To analyze the approach on drugs therapeutically applied for diseases of the cardiovascular system that are not covered by the ATC classification system, a test set was generated. It consists of 114 drugs with an indication on diseases of the cardiovascular system selected from the United States Pharmacopeia (USP)[24] and Therapeutic Category of Drugs (TCD)[25] maintained by the Japanese Pharmacopeia. To evaluate the classification results, all drug instances of the test set were manually annotated with ATC classes by H. Gurulingappa.

For all drugs of the training and test set property terms were extracted from MEDLINE titles and abstracts and were canonicalized with the developed approach described in Section 4.2.1. For drugs of the training set 3051 unique property terms have been obtained. Gurulingappa introduced an additional step for the normalization of obtained terms through a mapping to standard biomedical concepts contained in the UMLS metathesaurus. It has the additional advantage that synonyms and term variants are represented by a single concept identifier of UMLS. Synonyms, like *'5-HT antagonists'* and *'Antiserotonergic agents'* are mapped to the identifier *'C0037753'*. This step, performed with MMTx (Version 2.4.C), reduced furthermore the redundancy in the data set. Even though UMLS contains over two million concepts, not all extracted property terms are present. Hence, 32.6 % could not be mapped to UMLS identifiers. In order to overcome the problem of partial and unmapped concepts, a list of property concepts not present in UMLS was generated, which contain manually normalized terms that are linked to self created concept identifiers. Furthermore non-informative terms like *'drug'* have been removed. Finally, a total set of 368 unique concept identifiers were obtained for all 390 drugs of the training set. The overall number of the terms was reduced by 87.94 %.

The 368 unique concept identifiers were used to generate a feature vector for every considered chemical compound, whereas every concept represents one feature within that vector, which is depicted in Table 4.35.

For the purpose of experimentation, two types of feature vectors were generated for every drug. The first one is a binary feature vector with 368 numerical positions. The second one is a weighted feature vector. Its positional values equal to the number of the frequency with which the corresponding feature occurs within the corpus, normalized on a logarithmic

[24] http://www.usp.org/

[25] http://www.genome.jp/kegg-bin/get_htext?br08301.keg

Vector position	Concept ID	Term
1	C0001219	*'Acrylates'*
2	C0001413	*'Adenine Nucleotides'*
3	C0001640	*'alpha adrenergic receptor agonist'*
4	C0001641	*'Adrenergic alpha-Antagonists'*
⋮	⋮	⋮
368	S10000121	*'Antihyperlipoproteinemic'*

Table 4.35: Depiction of the feature vector conception.

scale. Examples of both types are illustrated in Table 4.36.

For finding a suitable classifier to predict ATC classes for not contained chemical compounds, Gurulingappa investigated the performance of four classifiers, i.e. Naïve Bayes, Decision Tree, k-Nearest Neighbor and Support Vector Machines. The training was done with different subsets of the features ranked by Chi-square [Simon, 2006] and 100-fold bootstrapping. Every classifier was tested with different parameter sets respectively (data not shown here), whereas the best one was chosen for the final comparison. The curves in Figure 4.23 reflect the course of the prediction accuracy obtained with the four classifiers and 10 % - 100 % of the features that were increased in steps of 10 %.

Figure 4.23 illustrates that the Naïve Bayes classifier with the weighted feature vector and SVMs with the binary feature vector outperformed the other classifiers. Furthermore it shows that the classifiers performed different when applying binary feature vectors instead of weighted feature vector. The analysis of all classifiers disclosed that the Naïve Bayes classifier generated a global maximum with a classification accuracy of 89.47 ±2.13 % when the top 70 % weighted features were applied.

Binary vector	1	0	0	0	0	1	…	0
Weighted vector	0.9	0	0	0	0	0.69	…	0
Position	1	2	3	4	5	6	…	368

Table 4.36: Schematic depiction of a section of the binary and weighted feature vectors. Every position of the vector corresponds to a concept and hence to an identifier either from UMLS or from a separately defined concept list for those concepts not contained within UMLS.

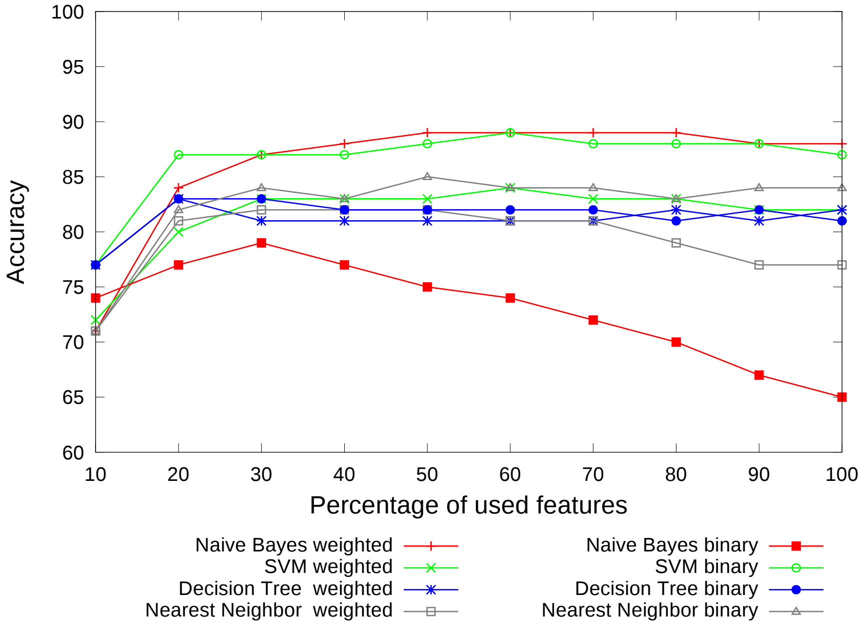


Figure 4.23: Performance of k-Nearest Neighbor, Naïve Bayes, Decision Tree, and SVM using weighted and binary feature vectors, validated by 100-fold bootstrapping. The classifier parameters were k=1 for k-Nearest Neighbor, application of the Kernel Estimator for Naïve Bayes, no pruning for Decision Tree and polynomial kernel for SVM. (The figure was adapted from [Gurulingappa et al., 2009].)

4.3.2.2 Results of the Classification Schema Extension

After selecting an adequate feature subset and Naïve Bayes as optimal classifier from the four tested classification methods it was evaluated on the test set of 114 drugs.

For comparison, the structure-based ATC classification method SuperPred was tested, which is introduced in Section 3.2.3.1. Therefore, SMILES string representations of all drugs in the test set have been provided as system input. The output of SuperPred provides structurally most similar compounds of the system's basic data to the query molecule that are ranked by decreasing Tanimoto coefficient values as well as their ATC classes. The ones of the top ranked chemical compound were compared to the manually annotated ATC classes of drugs in the testing set.

The number of correctly, wrongly and non-classified drugs obtained with both approaches is shown in Figure 4.24. Additionally, Figure 4.25 provides the results for recall, precision and F_1 measure. The evaluation of the classification results illustrates that the concept-based approach was able to outperform the structure-based approach SuperPred by classifying the drugs with a precision of 94 %, a recall of 82 %, and an overall classification F_1 measure of 86 % on the limited test set of cardiovascular drugs. In contrast, SuperPred provided a much lower precision of 60 %, the same recall of 82 %, and an overall classification F_1 measure of 69 %.

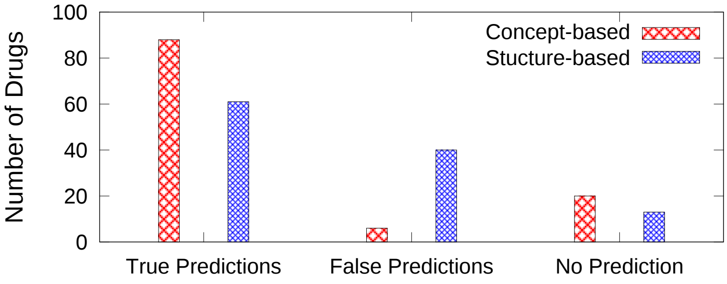


Figure 4.24: Comparison of concept-based with structure-based ATC class prediction for drugs in the test set comprising 114 drugs. The number of correctly, wrongly and non-classified drugs is provided for both approaches. (Figure was adapted from [Gurulingappa et al., 2009]).

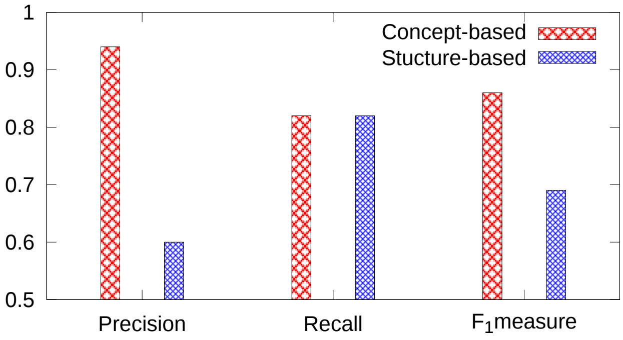


Figure 4.25: Evaluation of concept-based and structure-based ATC class prediction for drugs in the test set. Recall, precision and F_1 measure are given.

4.3.3 Summary of the Developed Information Aggregation Framework and Discussion of the Results

The developed information aggregation framework combines the recognition of chemical named entities described in Section 4.1 with the extraction of related function information explained in Section 4.2. To demonstrate the application potential of the framework, 11 drugs have been selected to obtain function annotation terms from MEDLINE articles which were exemplarily compared with annotation concepts provided by DrugBank. The analysis of property terms subsequently extracted from MEDLINE illustrated that 29-53 % of them could be identified as valid new drug classifications. They contain new valuable annotation information and hence could serve as novel annotation terms. Additionally, the extracted and canonicalized terms could enlarge the drug classification spectrum and speed up the annotation process, even if not all found terms might be useful. In general this approach can be applied for annotating chemical compounds or pharmaceuticals from other databases. Furthermore, it is generic and thus can be applied to other areas as well only by exchanging named entity recognition to focus on Hearst phrases that are content-specific for a certain domain. The developed approach can be utilized to assist database curators to scan free text data resources which represent the most up-to-date sources of information. Therewith, the work of information gathering and manual entity annotation is supported.

The extracted property information was employed in the developed drug classification scheme extension. Therefore, property concepts extracted from text were utilized as features of property descriptors that were applied by an automated process to extend the drug classification scheme ATC with new drug instances by class prediction methods. It was demonstrated that ATC class prediction, using a defined biological, pharmaceutical weighted feature vector and a Naïve Bayes classifier, was successful for the selected subset of ATC proved by a high F_1 measure of 86 %. The method outperformed SuperPred – a recently published structure-based ATC class prediction methodology, which achieved an F_1 measure of 69 %.

The value of utilizing textual data for pharmaceutical research, especially for drug repurposing, is also supported by a study published by Campillos et al. [2008] at the same time the study on the extension of the ATC classification was done. They explored side-effect information from drug package inserts that have been generated from the use of marketed drugs. They extracted terms as well and classified side effects using the Unified Medical Language System ontology for medical symptoms. Their idea was to infer molecular activities of drugs that are not implicitly encoded by their chemical structure and hence not solely inferable from structure and sequence similarity of their known targets. They predicted new targets for well-known drugs by applying side-effect similarity measures and successfully verified their hypotheses with binding assays in vitro. Like the previously described approach, the introduced method also incorporated biomedical concepts and clearly demonstrated that it is worth to use them for the prediction of new targets for already marketed drugs.

A drawback of the concept-based approaches, like the presented classification, in general is its dependency on available textual data provided in a specific linguistic form. Thus, the extracted concept features are strongly dependent on the text corpus underlying the complete concept extraction process that could lead to an inherent feature sparseness and data bias because of missing data. Additionally, the approach does not take organism-specific pharmacological actions of chemical substances into account, so that the ATC class prediction is organism-independent. This could be solved by filtering the articles for certain organisms before the Hearst phrase extraction is performed. However, it might decrease the amount of property concepts available on chemical compounds in text. To partially overcome this problem, other text corpora or full text articles could be added to the phrase extraction pipeline. Furthermore, an extraction of other relations than the discussed ones could enhance the extraction and collection of more information on chemical compounds to enrich the feature space. However, the approach does not allow its use for completely novel chemical compounds. Therefore, additional concepts from further resources, like other databases or classification schemes that contain pharmaceutical and biological property concepts could be incorporated.

The application of biological and pharmaceutical property concepts for extending classification schemes by further chemical entities has proved to be applicable and provided good results on a subset of ATC. In the future, efforts have to be made for including a complete classification scheme. Therefore, the above discussed extension of the feature space is an important step. However, with further improvement of the method it could be a valuable technique applied as one additional approach for decision making in pharmaceutical research in the future.

Chapter 5

Histone Modification Recognition in Text

Histone modifications play a key role in epigenetic mechanisms. Growing research in this field resulted in a steep increase of literature data in the last years. However, no automated approach has been published so far that support the recognition of histone modifications descriptions in text and allow for its identification.

In July 2008, MEDLINE contained over 24,600 abstracts dealing with epigenomics, wherefore PubMed was queried using the term *'epigenetics'*. About half of them contain information about histone modifications as was depicted in Figure 1.3 of Section 1.1. On average, over 1000 articles have been published every month in the last two years, which is a high publication rate.

Querying PubMed for histone modifications reveals several problems. The same example terms of H3K9me3 provided in the introduction have been used to search for articles containing information of that histone modification type. Table 5.1 provides the abstract's quantity obtained with a PubMed query for every given term. The results illustrate the high dependence of the retrieved number of articles from the search term. This is a definite disadvantage when searching for documents on a specific histone modification type and indicator for the absence of defined histone modification terms from the controlled vocabulary of MeSH.

5.1 Development of a CRF-based NER Approach for Recognizing Histone Modifications

Analyzing the selected terminology examples, it was expected that a simple search strategy is not able to find all description variants related to a certain histone and modification type. Further on, the generation of a dictionary from scratch would be difficult, because not all description variants that authors create, could be foreseen in advance, especially enumeration variations of histone modifications.

Therefore, a Machine Learning-based system, like Conditional Random Fields (CRFs) introduced in Section 3.1.2.3, was trained to support the identification of histone modifications in text. To develop such a system an annotated training corpus was required from which the histone modification description characteristics could be learned and a model to be applicable on new unseen examples could be induced.

Histone modification term variants	Number of obtained articles from PubMed
H3K9me3	41
Me3-K9 H3	1
Me(3)-K9 H3	78
H3K9 tri-methylation	7
H3-K9 trimethylation	28
H3 Lys9 trimethylation	11
H3 tri-methylated at lysine 9	14
histone H3 trimethylated at lysine (K) 9	3
K9 trimethylation at histone H3	36
K9-trimethylated histone H3	15
tri-methylation of H3 at lysine residues K9	0
trimethylated H3K9	18
di- and trimethylated H3K9	8

Table 5.1: Term variant examples of one histone modification type. The number of obtained articles from PubMed are given (data from July 2008).

5.1.1 Corpus Selection and Annotation

5.1.1.1 Corpus Generation

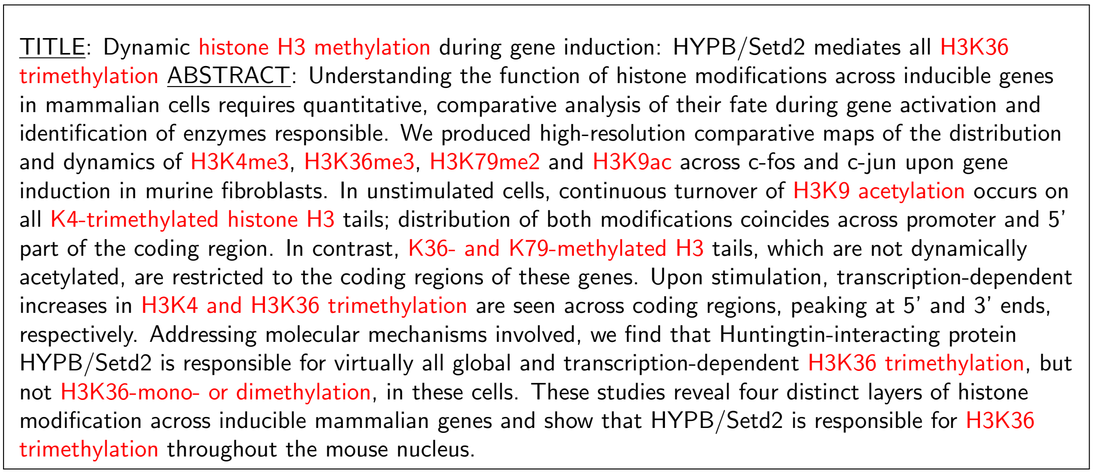
For training a CRF model an initial corpus (referred to as EPI-TRAIN) of 187 MEDLINE titles and abstracts has been selected manually from a corpus in which both histones and modification terms occur together. This was obtained by a co-occurrence MEDLINE search with ProMiner using two generated separate dictionaries. One contains the terms *'histone'* and *'histones'*. Contrarily, the second dictionary comprises 75 modification terms and some spelling variants that represent different modification types in general. These are for instance *'di-methylation'*, *'dimethylation'*, *'ubiquitation'* and *'acetylation'*. With this approach 10,576 articles have been obtained. From that corpus 187 titles and abstracts have been selected manually. It was ensured that every histone modification type is covered by the corpus.

For testing the trained model, a corpus called EPI-TEST has been generated on the basis of a PubMed search using the MeSH term *'epigenetics'*. From the 24,653 obtained articles 1,000 titles and abstracts have randomly been chosen. They are distinct from the articles contained in the EPI-TRAIN corpus. For further evaluation of the system a general third corpus EPI-TEST-R has been randomly sampled from complete MEDLINE. Table 5.2 shows the properties of all three corpora; the number of documents, sentences, and tokens.

5.1.1.2 Corpus Annotation

As for the annotation of the CHEM-EVAL corpus WordFreak has been used for this task. All three corpora have been annotated with the entity class **Hmod**. A term to be annotated as entity type **Hmod** had to fulfill following constraints: The term had to contain at least one histone type and one modification term, e.g. '*histone acetylation*' or '*histone 3 dimethylation*'.

The removal of a modification, like '*H3K9 demethylation*', has also been annotated, because an existing modification is changed. Instead, if a histone modification term fraction is part of an enzyme, e.g. in '*H3K9 methyltransferase*', the term is not annotated. Enumerations are handled as follows: If modification terms, similar to the official nomenclature, occur in an enumeration, like '*H3K36me3, H3K79me3 and H3K9ac*', they have been annotated as single terms. By contrast, long forms, like '*H3K36-mono- or dimethylation*', have been annotated as a whole phrase. An annotated text containing different examples terms is provided in Figure 5.1. The number of annotated entities for every corpus is provided by Table 5.2 as well.



TITLE: Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation ABSTRACT: Understanding the function of histone modifications across inducible genes in mammalian cells requires quantitative, comparative analysis of their fate during gene activation and identification of enzymes responsible. We produced high-resolution comparative maps of the distribution and dynamics of H3K4me3, H3K36me3, H3K79me2 and H3K9ac across c-fos and c-jun upon gene induction in murine fibroblasts. In unstimulated cells, continuous turnover of H3K9 acetylation occurs on all K4-trimethylated histone H3 tails; distribution of both modifications coincides across promoter and 5' part of the coding region. In contrast, K36- and K79-methylated H3 tails, which are not dynamically acetylated, are restricted to the coding regions of these genes. Upon stimulation, transcription-dependent increases in H3K4 and H3K36 trimethylation are seen across coding regions, peaking at 5' and 3' ends, respectively. Addressing molecular mechanisms involved, we find that Huntingtin-interacting protein HYPB/Setd2 is responsible for virtually all global and transcription-dependent H3K36 trimethylation, but not H3K36-mono- or dimethylation, in these cells. These studies reveal four distinct layers of histone modification across inducible mammalian genes and show that HYPB/Setd2 is responsible for H3K36 trimethylation throughout the mouse nucleus.

Figure 5.1: Example title and abstract with histone modifications annotated as entity type **Hmod** (PMID: 18157086, Edmunds et al. [2008]).

Corpus	Documents	Sentences	Tokens	Annotated entities
EPI-TRAIN	187	1,605	44,876	601
EPI-TEST	1,000	8,880	236,160	221
EPI-TEST-R	1,000	5,313	123,920	0

Table 5.2: Number of documents, sentences, tokens and annotated entities for the three selected training and test corpora.

5.1.2 CRF Training and Feature Selection

For obtaining a CRF model to be used for identifying histone modifications in documents MALLET [McCallum, 2002] was used as basic implementation. The package provides default features for representing text tokens, like morphological features (cf. Table 3.4). Nevertheless, further features have been added, like spaces which were described in Klinger et al. [2008].

	EPI-TRAIN	EPI-TEST
Recall	0.81 (± 0.05)	0.76
Precision	0.87 (± 0.05)	0.87
F_1 measure	0.84 (± 0.05)	0.81

Table 5.3: Recall, precision, and F_1 measure are provided if a CRF model with the optimal feature set is used to label the selected training corpus EPI-TRAIN and the test corpus EPI-TEST. The numbers in brackets for the EPI-TRAIN corpus provide the standard deviation of the 10-fold cross-validation.

A granular tokenization of the text was performed, splitting it at white spaces, before and after diverse bracket forms and special symbols, like '-' or comma, etc. In the following a tokenized example text snippet is depicted ('|' displays the text string separator):

'... | Indeed | , | Rtf1 | is | required | for | H2B | ubiquitination | , | suggesting | that | its | effects | on | H3 | - | Lys4 | and | H3 | - | Lys79 | methylation | are | an | ...'.

To validate the trained CRF model a 10-fold cross-validation was performed on EPI-TRAIN. Following initial results were obtained for precision, recall and F_1 measure: 0.87 ± 0.05, 0.81 ± 0.07, 0.84 ± 0.05.

5.1.2.1 Feature Selection

As Klinger et al. [2008] showed in the application of CRFs for the recognition of IUPAC and IUPAC-like names, single features have a different impact on precision and recall. They demonstrated that some features are very important for the performance of the system, whereas others do not change the F_1 measure. They studied the influence of features by leaving single ones or combinations of some feature types out. Omitting non-informative features reduces its number that has to be considered for the CRF model. This could be advantageous in processing time when larger text corpora are tagged by the system.

Feature sets proven to lead to a good recognition performance on one entity type might result in a poor outcome for another. Hence the optimal feature set used for the recognition of IUPAC names [Klinger et al., 2008] might not be an optimal one for identifying the new entity type **Hmod**. That is why an analysis of the features to be selected for obtaining good recognition results was performed.

To prove the impact of the different features used for the recognition of histone modifications, single features belonging to various classes or combinations of them have systematically been left out. For every modified feature set a single model has been trained on EPI-TRAIN and was validated by 10-fold cross-validation. The obtained precision, recall, and F_1 measure are displayed in Figure 5.2.

The best feature set has a high performance in recall (0.81), precision (0.87), and F_1 measure (0.84) on EPI-TRAIN. Comparing the initial results with the new ones, an improvement of 1 % could be obtained with this step (cf. Table 5.3). The optimal feature set includes

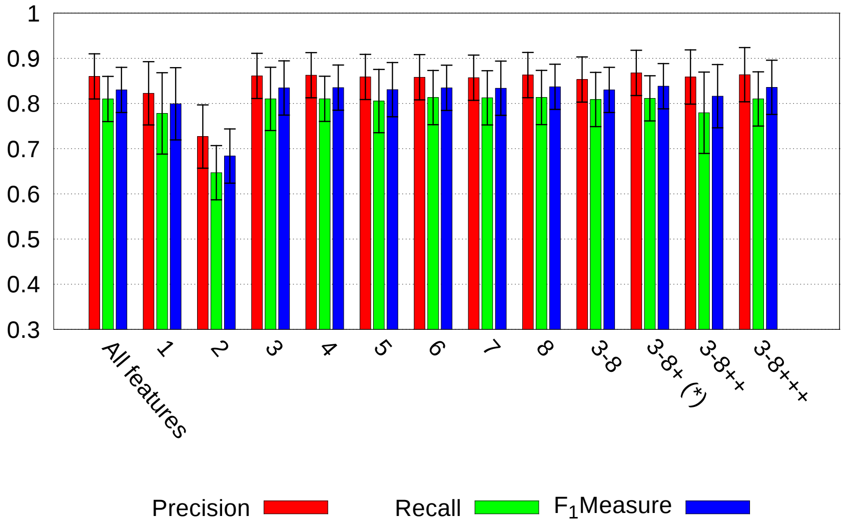


Figure 5.2: Recall, precision, and F_1 measure are given for every single feature analysis experiment on EPI-TRAIN. Non-used features or combinations of them belong to two classes *Automatic generated morphological features (AM)* and *Context (C)* and are depicted in the plot. The final used feature combination is marked with '(*)'.
1: No Word as class (C)
2: No Spaces (C)
3: No natural number (AM)
4: No Capletter (AM)
5: No Capsmix (AM)
6: No InBracket (C)
7: No Init caps (AM)
8: No Single char (AM)
3-8: Combination of features from 3 to 8 not used
3-8+: Combination of features including feature 3 to 8 and 24 additional morphological features not used
3-8++: Combination of features including feature 3 to 8 and 31 additional morphological features not used
3-8+++: Combination of features including feature 3 to 8 and 57 additional morphological features not used

following features and feature generating methods: *Prefix*, *Suffix*, *InBrackets*, *Words as Class*, *Spaces*, *wordClass*, and *doBriefWordClass*. The features from class *Static morphological* have no impact on the result, hence they have been omitted altogether. On the contrary, leaving out *Spaces* and *Words as Class* affect the histone modification term recognition and lead to a considerable decrease in precision, recall, and F_1 measure. It points out that it is relevant whether the token is preceded or succeeded by white space and if words occurring in histone modification descriptions are learned by the system. The first one is important especially in enumerations or abbreviations of terms to separate them from each other. This feature already indicated a high impact on the identification of IUPAC and IUPAC-like names [Klinger et al., 2008].

The apparent optimal feature set identified by feature selection was applied to tag both test corpora EPI-TEST and EPI-TEST-R. On EPI-TEST-R a precision of 1 was obtained (data not shown). It illustrates that no false positive term was found in the general sample test corpus, containing no histone modification term. It shows that the system in general is able to distinguish between non-histone modification terms from positive terms. Recall, precision, and F_1 measure that were yielded on EPI-TEST are provided in Table 5.3. Compared to the results on the training corpus, tagging of EPI-TEST lead to the same precision (0.87), but to lower recall (0.76). It gives some indication that the model was not general enough for finding all histone modification terms. This might be due to the relative small training corpus EPI-TRAIN containing too few annotated histone modification term examples. However, the model is already sufficient to identify terms in EPI-TEST not previously seen as training example. This is supported by the fact that 68.6 % of the true positive terms did not occur within the training corpus.

5.1.2.2 Recognition Problems

The tagged entities have been examined to identify false positive and false negative terms. False positives have been divided into four classes. The following list provides a brief description of every class and shows some examples:

- **Error class 1:** Modification descriptions without histone mentions: *'acetylation and methylation'*.
- **Error class 2:** Enzymes introducing or removing histone modifications: *'H3K9 methyltransferase'*.
- **Error class 3:** Incorrect recognition of term boundaries: *'H3 - K9) with no sign of histone H2AX phosphorylation'*, *'H3K9me3 at pericentric heterochromatin. H3K27me3 and H4K20me'*. The red marked entities should have been found as separate entities here.
- **Error class 4:** Terms with other meaning: *'phosphorylation of IRS'*, *'eradication of H7N1, H7N3 and H5'*.

It turned out that most of the false positives belong to classes 3 and 4. More training data could reduce the number of false positives. Furthermore, a post-processing step decreases their number which is later described in Section 5.2.

The analysis of the unified false negative modification terms not found in EPI-TEST showed that 6 histone modification terms (20 %) occurred in at least one document within the training corpus EPI-TRAIN. Histone modification terms which were not recognized are for instance: *'histone acetylation'*, *'histone methylation'*, and *'ubiquitylation of histone H2B'*. This might be due to the fact that the CRF learns the properties of the context tokens which surround an entity within a text. If new adjacent tokens with other feature distributions as those represented by the trained model occur in a new example, the modification term is not recognized, which is a clear disadvantage of this method. Similar to solving the reduction of false positives, an increase of annotated documents could help to enhance the performance of the CRF for overcoming this problem and to improve the recognition of false negative terms as well.

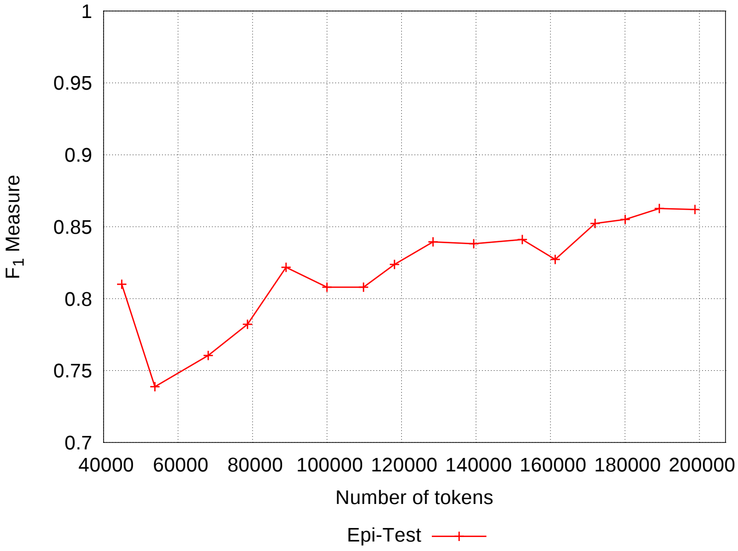


Figure 5.3: F_1 measures obtained on EPI-TEST after every active learning step in relation to the total number of tokens in the iteratively extended training corpus.

5.1.3 Improvement of a primary CRF Model by the Extension of the Training Corpus through Active Learning

The necessity to enhance the histone modification recognition emerged from the seen problems. This could be achieved by increasing the training corpus used for building up the CRF model. The intention was to select example texts to be annotated and added to the training corpus EPI-TRAIN. This was done via active learning based on uncertainty sampling introduced in Section 3.1.4.1.

The sampling procedure could keep the annotation effort of new training samples in a passable limit. Since active learning is an iterative process, several rounds of annotation, CRF training and testing on the test corpus EPI-TEST have been performed. In every iteration step new sample documents were chosen from complete MEDLINE for annotation on which the system was least confident.

The enhancement of the system's F_1 measure, evaluated on EPI-TEST, was studied. The convergence of F_1 measures was utilized as stopping criterion for the active learning process. 15 iteration steps have been done until convergence in the F_1 measure was observed on EPI-TEST. Table 5.4 shows the number of iterations, the annotated entities as well as the rising number of tokens of the enlarging training corpus in every iteration round. The process resulted in an extended training corpus EPI-TRAIN-AL containing 434 documents more compared to the initial training corpus. The F_1 measure depending on the number of tokens are depicted in Figure 5.3.

It turned out, the extension of the training corpus improved the system's precision by 6 %, and both recall as well as F_1 measure by 5 %. The evaluation results obtained for the CRFs either trained on the original training corpus or on EPI-TRAIN-AL are provided in Table 5.5. All evaluation measures stayed the same on EPI-TEST-R (data not shown).

Run	Annotated entities	Tokens
AR 1	502	53,785
AR 2	522	68,083
AR 3	550	78,648
AR 4	567	88,994
AR 5	624	99,987
AR 6	647	109,804
AR 7	672	118,100
AR 8	709	128,456
AR 9	733	139,382
AR 10	769	152,425
AR 11	814	161,271
AR 12	879	171,974
AR 13	896	180,052
AR 14	917	189,237
AR 15	936	198,803

Table 5.4: Result of the active learning procedure. The number of iteration (AR) No. 1–15 is provided at the left followed by the total number of annotated entities and tokens for every iteration step.

	EPI-TEST	
	Initial model	Extended model after 15 active learning rounds
Precision	0.87	0.93
Recall	0.76	0.81
F_1 measure	0.81	0.86

Table 5.5: Comparison of the identification results on EPI-TEST before and after active learning.

The new model (referred to as extended model in Table 5.5) trained on the extended training corpus improved the ability to differentiate between positive and non-histone modification terms, which is reflected by a high precision of 0.93. Furthermore, it reduces the number of false positive terms by 22 and increases true positives by 27. False positives belonging to all error classes, introduced within the previous section, were decreased.

5.2 Canonicalization of Histone Modification Terms

Machine learning techniques utilized for named entity recognition like the CRFs are only able to provide a classification of tokens with which an assignment to a certain predefined entity class is possible. Hence, the recognition of single histone modification descriptions or enumerations of them in text provides only positional term information useful for highlighting histone modifications in text. However, it does not support semantic search, because different term variations of one modification description are still considered as different entities. An inevitable next step is to map different description variants of histone modifications onto a canonical term form respectively, which shown in Figure 5.4, and to filter out false positive terms. This supports semantic retrieval of all documents related to one histone modification type, which has not been possible up till now and solves the document retrieval problem depicted in Table 5.1.

Canonical terms that come into consideration for histone modification are those which follow the 'Brno nomenclature'. It defines a specific abbreviation type of histone modification terms. These terms provide several advantageous properties; they are short, their generation is defined and they are unique. According to the nomenclature, histones are defined by a single '*H*', followed by a number or a combination of a number and letters illustrating a certain histone type. The amino acids are described by single characters defined by the one-letter code[1] in combination with their position information within the protein chain. The modification types are depicted by two or three letters. It is followed by a number which defines the amount of modifications of one type on this specific amino acid. An example is '*H3K9me2*'.

5.2.1 Canonicalization Workflow

To map recognized histone modification terms onto canonical standard terms corresponding to the 'Brno nomenclature' they are transformed by an automated canonicalization procedure which was developed and implemented in a workflow. In general, a term is passed through checking units which translate the information implied in the term to produce a canonicalized standard term form. Figure 5.5 depicts a general overview on the canonicalization workflow.

The workflow units, i.e. the filters and transformation units include rule sets which are applied to every found term. They have been established by analyzing entities from the two annotated corpora EPI-TRAIN (before active learning was performed) and EPI-TEST. Primarily, rules have been developed using all entities from EPI-TRAIN. Subsequently, they have been tested on manually transformed entities from EPI-TEST. For reducing the number

[1] http://www.chem.qmul.ac.uk/iupac/AminoAcid/A2021.html

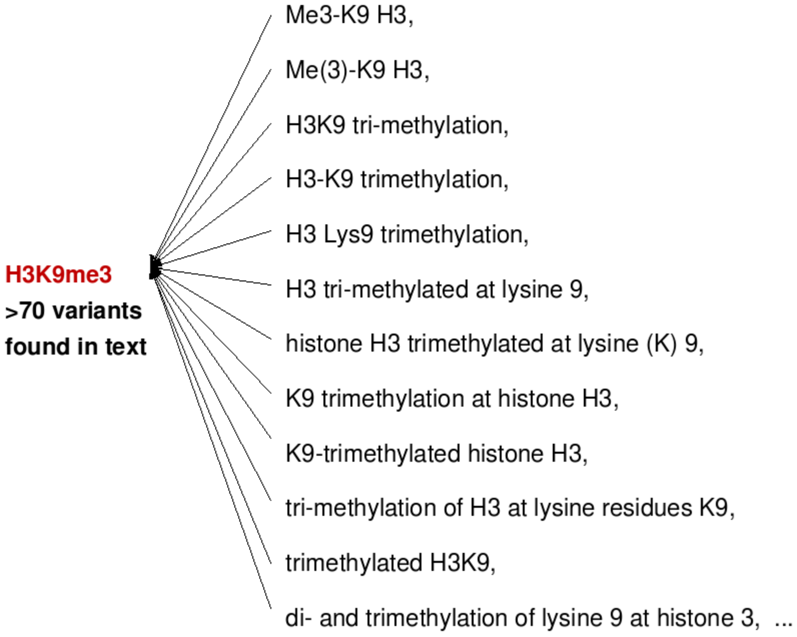

H3K9me3
>70 variants
found in text
Me3-K9 H3,
Me(3)-K9 H3,
H3K9 tri-methylation,
H3-K9 trimethylation,
H3 Lys9 trimethylation,
H3 tri-methylated at lysine 9,
histone H3 trimethylated at lysine (K) 9,
K9 trimethylation at histone H3,
K9-trimethylated histone H3,
tri-methylation of H3 at lysine residues K9,
trimethylated H3K9,
di- and trimethylation of lysine 9 at histone 3, ...

Figure 5.4: Mapping of different apparent term variants of the entity *'H3K9me3'* recognized in text to one standard term corresponding to the 'Brno nomenclature'.

of false positives, further rules have been incorporated into the system after testing on EPI-TEST entities. In the following every step of the workflow is described in more detail.

- **Validity check:** Is performed to filter false positive terms recognized by the learned CRF model. One basic rule is the absence of a histone type, for instance. Furthermore, several regular expression patterns are applied which describe false positive terms, e.g. C3-H(1|2) which matches C3-H1.

- **Term length check:** If a term passed the previous filter, it is checked for a general property, the term length. Long terms and short terms are treated differently in the remaining part of the workflow because different canonicalization processes are used. Short terms already consist of abbreviations and either correspond to the nomenclature or are similar to it, like '*Me3-K9 H3*'. Long forms include complete modification type, amino acid or histone descriptions, like '*dimethylated lysine 20 of histone H4*'.

  - **Term transformation of long terms:** Rules check the presence of basic histone modification information and transform it to the standard representation, example term: '*dimethylated lysine 20 of histone H4*':
    - Histone type '*H4*',
    - Amino acid '*H4 K*',
    - Position of the amino acid in the protein sequence '*H4 K 20*', and
    - Quality of the modification '*H4 K 20 me 2*'.

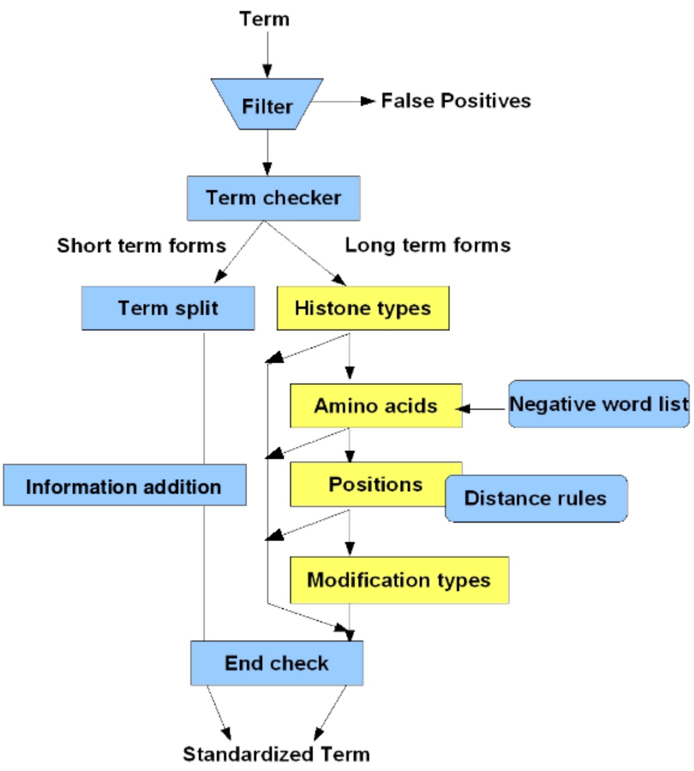


Figure 5.5: Canonicalization workflow which transforms recognized histone modification terms into canonical term forms.

The first property needs to be there, others are optional to create a standard output term.

- **Term transformation of short terms:** If the term consists of several modification types, the term is split, two terms are generated, and the missing information are added to the second one, an example term is *'H3K9me2S10p'*. Otherwise, the term is transformed to result in a canonicalized form:
  - Splitting of the term *'H3K9me2'*, *'S10p'*,
  - Addition of remaining information '**H3***S10p*' (*'H3K9me2'* stays the same).

- **End check:** Every resulting term is allowed to have only one representative of all four information types. In case the algorithm produces output terms not following this constraint, they are considered as false canonicalization results and are filtered, i.e. if:
  - A term has more than one amino acid like *'H3 K 4 S 10 me 3'* resulting from: *'histone H3 lysine 4 trimethylation (Me(3)-K4 H3) and histone H3 serine 10 phosphorylation'*.
  - More than one modification type is contained, like *'H4 K 18 ac ac'* from *'acetylation on H3K14 and H4K5, and hypoacetylation on H3K18'*

	EPI-TRAIN	EPI-TEST
Manually canonicalized terms	414	123
Terms correctly canonicalized	397 (96 %)	121 (98 %)

Table 5.6: Evaluation results of the term canonicalization process obtained on entities of the corpora EPI-TRAIN and EPI-TEST. Given are the number of manually annotated histone modification terms and the number and fraction of terms correctly automatedly canonicalized.

- Wrong modifications have been assigned to the wrong amino acid, like *'H3 K 4 ph'* from *'phosphorylated histone H3 displayed mostly Lys-4 dimethylation'*. In this case lysine cannot be phosphoylated.

Terms containing enumerations of histone types, modified amino acids, modifications or positions, like *'di- and trimethylation of lysine 4 at histone 3'*, can be resolved by the procedure as well. The canonicalization of such terms would lead to more than one standard term. The transformation of the given example term results in the two terms *'H3K4me2'* and *'H3K4me3'*.

5.2.2 Evaluation of the Canonicalization

To enable the evaluation of the canonicalization process, every annotated and unique histone modification of EPI-TRAIN and EPI-TEST was manually assigned to canonical terms leading to two standard test sets. They have been used for the automated evaluation of the canonicalization results. The number of manually transformed histone modification terms and results of the canonicalization procedure are given in Table 5.6. It shows a very good performance of the system. Over 96 % of the entities from the EPI-TRAIN corpus and over 98 % of the entities from the EPI-TEST corpus have been transformed correctly.

To evaluate the system on a large text corpus, complete MEDLINE was tagged by the CRF-based system. This resulted in 82,981 terms that have been recognized. They were subsequently passed to the canonicalization procedure. The study of the filtering result revealed that 63,314 (76.30 %) of all recognized terms are classified as false positives. Most of these are abbreviations occurring very often in MEDLINE and histone types which do not describe a modification. Table 5.7 shows some example terms and its frequency in the large text corpus. The large amount of wrongly found abbreviations is a surprising result, as the evaluation of the CRF approach on the randomly chosen test corpora EPI-TEST and EPI-TEST-R, both consisting of 1000 abstracts and titles, yielded a high precision (cf. results in Section 5.1). It demonstrates that training and test corpora are restricted representatives of a large text corpus. Even though the training corpus EPI-TRAIN-AL consists of a high number of selected articles and good results were obtained through evaluation on the test sets, the finding of that many false positives could not be prevented at a real test case, i.e. the tagging of complete MEDLINE.

However, from these filtered false positives only 0.08 % were wrongly classified as such,

Example term	Frequency in MEDLINE
AChE	21,700
ECoG	2,012
BChE	1,678
histone	1,511
BUdR	1,224
UNaV	1,158
SIgA	1,045
FUdR	890
BLyS	871
IUdR	822

Table 5.7: Examples of false positive terms recognized by the CRF approach in complete MEDLINE.

thus actually being true positives. It demonstrates a high performance of the filtering step. Through a more detailed analysis it turned out that most of the false positive terms comprise a different meaning and hence belong to the error class 4 (introduced in Section 5.1.2).

The remaining true positive terms have been subjected to the subsequent term transformation process. This lead to 16,250 canonicalized entities obtained from complete MEDLINE. From those 4,086 terms were evaluated. They were chosen in such a way so that for every concept (364) every unique synonym was analyzed. A high fraction of 3,628 (~89 %) were transformed correctly, whereas 458 (~11 %) were wrongly canonicalized. Examples of them are shown in Table 5.8. As can be seen, histone types, amino acids and modification types were wrongly combined.

Many of the terms that caused problems in the canonicalization correspond to recognition error class 3 (cf. Section 5.1.2.2), which indicate the identification of wrong term boundaries. Furthermore, long enumerations with several different histone modification descriptions

Example term	Wrongly automated canonicalization	Correct canonicalization
histone H2B ubiquitination affects H3K79 trimethylation	H2BK79ub3, H2BK79me3	H2Aub, H3K79me3
H3K27 trimethylation , H2A ubiquitination	H2AK27me3, H2AK27ub3	H3K27me3, H2Aub

Table 5.8: Example of wrongly canonicalized histone modification terms and its correct canonicalization.

provide a challenge for the canonicalization. Hence, for improving canonicalization, new rule sets have to be included into the term transformation process in the future.

5.3 Development of a Dictionary-based Approach for Recognizing Histone Modification Terms in Text

5.3.1 Generation of a Histone Modification Term Dictionary

A basic problem of the CRF approach is the influence of direct surrounding context of a certain token on its assignment to a certain state from the IOB-label alphabet. As discussed in Section 5.1.2, this leads to the result that the system does not always find histone modification terms in new unseen documents which do exist in the training corpus. It is a clear drawback, because it lowers the reliance of the approach. As NER builds the basis for document retrieval and information extraction, non-recognition of present entities would lead to a loss of available information.

To overcome this problem, the idea was to use histone modification terms recognized by the CRF system in a large text corpus like MEDLINE to generate a dictionary. This is then utilized by the dictionary-based NER approach ProMiner. Since the canonicalization procedure maps all terms that belong to one histone modification type to a canonicalized term representation and filters false positives, i.e. non-histone modification terms, the dictionary generation is straightforward to accomplish. Through this, the histone modification dictionary HmodDict was assembled and included into ProMiner resulting in ProMiner$_{\text{Hmod}}$. Figure 5.6 shows the workflow of the dictionary generation procedure.

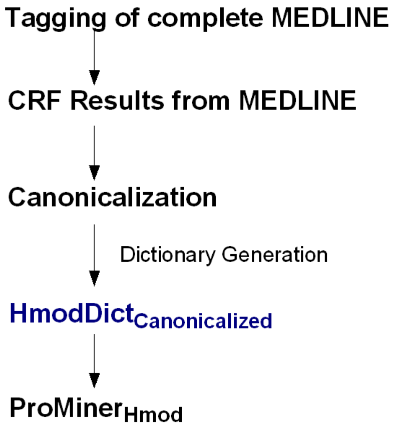


Figure 5.6: Depiction of the generation of the histone modification dictionary HmodDict which was included into ProMiner. This leads to a ProMiner$_{\text{Hmod}}$ version.

As described in the workflow, the canonicalized entities are the resource for the dictionary HmodDict. It comprises 364 unique histone modification objects which are related to 4,086 synonyms that was included into ProMiner resulting in ProMiner$_{\text{Hmod}}$.

5.3.2 Comparison of the Results from ProMiner$_{\text{Hmod}}$ and the CRF Approach

To be able to analyze the impact of the generated dictionary in comparison to the CRF approach, ProMiner$_{\text{Hmod}}$ and the CRF were operated on a large text corpus – the complete MEDLINE. Subsequently, the obtained histone modification terms were subjected to the canonicalization procedure. The respective number of extracted terms from complete MEDLINE achieved with ProMiner$_{\text{Hmod}}$ and the CRF approach are given in Table 5.9. Additionally, it shows the number of true positives and canonicalized entities.

		CRF	ProMiner$_{\text{Hmod}}$
Recognized terms on complete MEDLINE		82,981	21,273
False positives	Total	67,455	0
	Fraction of histone type terms	3,488	0
True positives		15,526	21,273
After filtering through canonicalization procedure		16,250	21,913

Table 5.9: Comparison of the CRF system with the dictionary approach for recognizing histone modifications on MEDLINE. The total number of histone modification terms recognized on complete MEDLINE, and the fraction of false positives and true positives are provided for both approaches. The entity number obtained after term canonicalization is given in the last row.

Compared to ProMiner$_{\text{Hmod}}$ the CRF approach provides a 3.9 times higher number of recognized terms on complete MEDLINE. However, the terms recognized with the CRF include a large amount of non-histone modification descriptions that are false positive findings. Finally, only 18.71 % of the recognized terms are true positive histone modification descriptions. The canonical entities are the result of the transformation of histone modification information comprised by the terms. Its number is provided in the last row of Table 5.9. Enumerations of histone modifications embodied in one term lead to its mapping to more than one entity. This explains the higher number of resulting entities in comparison to the true positive terms. Hence, the canonicalization of terms recognized by the CRF reduces the number of entities by the factor of 5.1.

As the results of ProMiner$_{\text{Hmod}}$ show, the dictionary-based approach finds 5,747 true positives (1.37 times) more compared to the CRF method. Here, canonicalization reduces the term number by a factor of 1.05 only. If canonicalized entities are compared between both approaches, ProMiner$_{\text{Hmod}}$ yields 5,663 entities (1.35 times) more compared to the CRF approach.

Entity type	Additional histone modification concepts recognized by $\mathrm{ProMiner}_{\mathrm{Hmod}}$
Hac	1644
H3ac	539
Hme	342
H3ph	284
H1ph	232
Hph	229
H3K9me	192
H3K4me	139
H4ac	131
H3me	124

Table 5.10: Difference in the number of canonicalized entities between the CRF approach and $\mathrm{ProMiner}_{\mathrm{Hmod}}$. The 10 most differing entity types in relation to the difference in the number of canonicalized entities is shown.

To analyze the cause of the difference in the entity number, the unified true positive histone modification terms obtained by the two approaches were compared with each other. It revealed that 1,672 new term variations were not contained in the primary generated dictionary HmodDict and hence were newly found in MEDLINE by $\mathrm{ProMiner}_{\mathrm{Hmod}}$. This is caused by the approximate string search introduced in Section 3.1.2.2. It allows for permutations of single tokens of a term, when the token number exceeds a given threshold. In the experiment the default setting was used that permits permutation of terms which comprise a minimum amount of four tokens. In a second study the occurrence of canonicalized histone modification entity types were compared between the two approaches respectively. The result for the top 10 most differing entities is provided in Table 5.10. As the result shows, $\mathrm{ProMiner}_{\mathrm{Hmod}}$ recognizes more terms with respect to specific histone modification types in MEDLINE than the CRF approach. When summed up, this portion already presents 3,856 histone modification entities more compared to the CRF results. This finding could be related to two causes: On the one hand side, this finding fits to the observation described in Section 5.1.2.2. There it was found that not all histone modification terms were recognized in the testing corpus EPI-TEST although they were available in the training corpus. It also shows that even though the training corpus has been extended by active learning, the generated model is not able to find all histone modifications available in the training corpus. Better recognition results could only be obtained with more training data which includes a high additional annotation effort. On the other side, the dictionary method $\mathrm{ProMiner}_{\mathrm{Hmod}}$ provides more reliable results than the Machine Learning-based NER approach when it is endowed with many term forms and synonyms. It finds a term that is provided by the dictionary every time it occurs in the processed text corpus. This is not the case for the CRF approach. As was shown, the approximate search algorithm of ProMiner has also the ability to find novel term variations of given synonyms. However, ProMiner$_{\text{Hmod}}$ it is not able to find new histone modification concepts which are not contained in the dictionary. Thus, Machine Learning-based methods are more flexible in recognizing novel coined terms in text that correspond to new entity types. Hence, they can be used as a tool to utilize text as resource for the generation of dictionaries when no other terminology source is available.

5.4 Generation of a Histone Modification Concept Hierarchy

Scientists working in epigenomic research have different information needs concerning histone modifications. They would like to obtain scientific articles with different focuses for getting an overview on the research in their own or related fields. Some would possibly ask a text retrieval system general questions, like:

‘Search for all documents that contain modifications of histone 3’.

Others might like to perform a more specific text search, like:

‘Search for all documents dealing with trimethylated lysine at position 9 of histone 3’.

The first question implicitly includes the second one in this case. It describes a demand that semantic text retrieval systems, like Textpresso[2] [Müller et al., 2004] and SCAIView [Hofmann-Apitius et al., 2008], can cope with. In such a system the recognized named entities are mapped to concepts of a hierarchy which is used for the organization of texts and allows for semantic search.

Available hierarchical structured terminologies and ontologies potentially applicable for a semantic search system on histone modifications have been analyzed. MeSH-T, Gene Ontology, PSI-Mod[3], and HistOn [Post et al., 2007] were examined for their usability as histone modification concept hierarchy. It turned out, there is no resource exhaustively covering histone modifications.

Therefore, an organism-independent hierarchy of histone modification concepts was established. In general, the hierarchy could be generated from two different points of view: Modification-centric or histone-centric. The decision was taken for a histone centric organization, for which the canonicalized terms were used as basic concept denominations. Herewith, getting a fast overview on all modification types of a certain histone type is enabled. Furthermore, applied in a semantic text retrieval system, it allows for organizing scientific texts related to one histone type at different granularity levels of modifications. A section of the complete generated hierarchy is given below for histone 3 as an example, whereas five possible methylation states are provided (mono-methylation: me1, di-methylation: me2, asymmetric di-methylation: me2a, symmetric di-methylation: me2s, tri-methylation: me3, and unspecified modification type: me) at two amino acids (K: lysine and R: arginine) and two positions (2 and 4):

```
0.3.0   H3
0.3.0.1   H3me
0.3.0.1.0.1   H3R2me
0.3.0.1.0.2   H3K4me
```

[2] www.textpresso.org

[3] http://psidev.sourceforge.net/mod/data/PSI-MOD.obo

```
0.3.0.1.1   H3me1
0.3.0.1.1.1   H3R2me1
0.3.0.1.1.2   H3K4me1
0.3.0.1.2   H3me2
0.3.0.1.2.1   H3R2me2
0.3.0.1.2.2   H3K4me2
0.3.0.1.2.a   H3me2a (asymmetric)
0.3.0.1.2.a.1   H3R2me2a
0.3.0.1.2.s    H3me2s (symmetric)
0.3.0.1.2.s.1   H3R2me2s
0.3.0.1.3   H3me3
0.3.0.1.3.1   H3R2me3
0.3.0.1.3.2   H3K4me3
```

To every term in the hierarchy a unique number has been assigned. It has at most 7 levels. A basic term set consisting of general histone modification concepts has been assigned to every included histone type. Subsequently, the hierarchy has been populated by canonicalized terms from Gene Ontology (GO), MeSH-T, HistOn, manually collected specific histone modification terms from the antibody online catalog of Abcam[4], and MEDLINE articles. The terms of the developed hierarchy have been automatically compared with the canonicalized ones from these resources. Those which have not been used so far within the hierarchy have been proposed by the system for its extension. An analysis of the impact of the single term resources is given below.

Since there was no existing comprehensive hierarchy ready to use, we developed our own, including 462 concepts. It is a manually created text file which was transformed into an xml-format. The used term resources contribute to the hierarchy concepts as follows:

- **Histone types (13)**
  - 13 histone types connected to GO obtained with Gene product search using AmiGO,
  - 13 in MeSH-T,
  - 7 in the online catalog of Abcam,
  - 8 in HistOn,
  - 10 in MEDLINE articles
- **General histone modification types (262)**
  - 16 in GO,
  - 47 in MEDLINE articles,
  - 1 in MeSH-T
- **Specific modification types from different resources (156)**

[4] http://www.abcam.com/

- 148 from online catalog of Abcam,
- 52 in MEDLINE articles,
- 1 in GO and HistOn.

The terms from the different resources overlap in content. GO and MeSH-T are the best of the considered resources for histone types, whereas Abcam and MEDLINE articles are the most useful resources for general and specific histone modification types. The current version of the created histone modification hierarchy covers the most important histone types and was integrated into the knowledge discovery system SCAIView, introduced in Section 3.3. Figure 5.7 depicts a section of the hierarchy as it is provided to users.

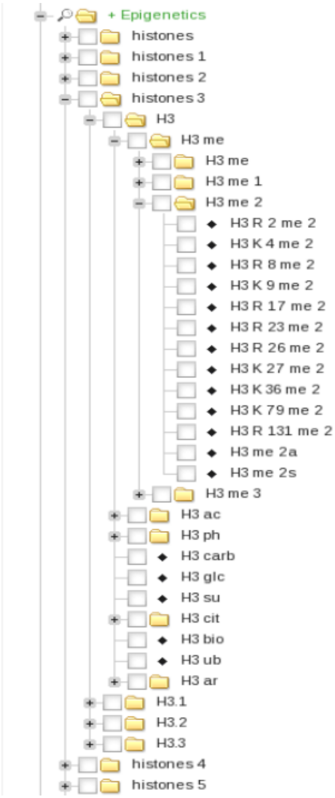


Figure 5.7: Section of the histone modification hierarchy included into SCAIView.

5.4.1 Automated Support of the Hierarchy Extension

New histone modifications will be described in the literature in the future, not yet contained in the hierarchy. Since manual search for new histone modification concepts would require to scan the literature regularly, it would take time until a new modification term could be included into the hierarchy. That is why a strategy was developed to automatedly support the regular hierarchy extension by proposing new concepts. Therefore, the canonicalized histone modification terms obtained from a ProMiner$_{\mathrm{Hmod}}$ run on complete MEDLINE are taken together with their occurrence frequency. They are compared to the existing concepts of the hierarchy and ranked by their occurrence frequency in MEDLINE. This procedure results in a list of novel histone modification concepts not yet contained in the hierarchy. It can be utilized to manually extend the histone modification concept hierarchy, whereas the concept frequency supports the inclusion decision. Table 5.11 provides the new histone modification concepts obtained with this procedure and the results of its semi-manual analysis. It shows the total amount of newly found concepts, the number of potentially new concepts to be included into the hierarchy and disregarded concepts. The latter ones were excluded because of a) a $occurrencefrequency < 2$ in MEDLINEand b) wrongly canonicalized concepts.

	Concept frequency in MEDLINE
Total number of newly found histone modification concepts	168
Number of potentially new concepts for hierarchy extension	90
Number of disregarded concepts	78

Table 5.11: Results of the extraction of new histone modification concepts from MEDLINE.

As was found, 57% of the valid new concepts describe a specific histone modification, such as *'H3S139ph'*. 53% contain the information on a general modification type to which *'H4Kac'* belongs as an example. The results let expect that with new findings published in the literature, the histone modification concept hierarchy is expected to grow with time until all occurring modification types are experimentally discovered and extensively mentioned in the literature.

5.5 Applications of the Histone Modification Identification Approach and the Concept Hierarchy

5.5.1 Analysis of the Information Content of Histone Modification Descriptions Extracted from MEDLINE

To see which histone modification concepts were found most often in MEDLINE an analysis of its frequency was conducted. Therefore, the canonicalized terms extracted with $\text{ProMiner}_{\text{Hmod}}$ were counted and sorted by its frequency. Table 5.12 provides the results of the top 20 most often occurring histone modification concepts in MEDLINE.

As the analysis of the results reveals, the most often recognized terms are histone types. Furthermore, often only the modification type in relation to a certain histone protein is given. More detailed modification descriptions including the amino acid and side group number are less often mentioned in the articles of MEDLINE. Hence, the examination of recognized histone modification terms from MEDLINE revealed that not every term contains

Histone modification concept	Concept occurrence in MEDLINE
Hac	6,149
H2A.Xph	2,445
H4ac	1,362
Hme	1,089
H3ac	871
Hph	725
H3K9me	613
H1ph	601
H3ph	539
H2A.X	489
H3K4me	385
H3K4me3	380
H2Aub	331
H3S10ph	275
H3K27me3	263
H3K9me2	255
H2Bub	226
H3me	212
H3K9me3	200
H3K9ac	199

Table 5.12: Frequency of the 20 most occurring histone modification concepts in MEDLINE.

detailed modification information. To get a general overview on the information content of the terms they were analyzed and automatedly translated into patterns reflecting the order and particular basic information units of the term; i.e. histone (H), amino acid (As), sequence position (Pos), and modification (Modi). Therefore, histone modification terms recognized with $\text{ProMiner}_{\text{Hmod}}$ were analyzed and translated. The generated patterns were statistically investigated, whereas the most often occurring patterns, corresponding term numbers, the relative amount and example terms are depicted in Table 5.13.

The depicted statistics of the 10 most often occurring patterns reflects the results shown in Table 5.13 above. A difference of the pattern creation to the canonicalization is the order of the information which was left as it appears in the terms in connection with prepositions or coordinating conjunctions. It was maintained in the patterns to be able to determine how histone modifications are described.

5.5.2 Document Retrieval

As became apparent by the PubMed query results presented in Table 5.1, MeSH does not index articles in MEDLINE with defined histone modification terms. Hence, PubMed can only be queried by typing the modification term in the search window. It has the consequence

Pattern type	No of occurrence (relative amount in %)	Term examples
H Modi	7,347 (38.24 %)	*'histone 3 acetylation'*, *'H3 acetylation'*
Modi H	3,473 (18.08 %)	*'methylated histone 3'*
Modi of H	2,261 (11.77 %)	*'methylation of histone 3'*
H As Pos Modi	2,150 (11.19 %)	*'H3 lysine 9 trimethylation'*, *'H3K9me3'*
Modi of H As Pos	352 (1.83 %)	*'methylation of H3K9'*
Modi H As Pos	214 (1.11 %)	*'methylated H3 lysine 9'*
H Modi and Modi	175 (0.91 %)	*'H3 methylation and acetylation'*
H Modi As Pos	158 (0.82 %)	*'histone 3 methylated K9'*
H As Modi	120 (0.63 %)	*'H3 lysine methylation'*
H As Pos	108 (0.56 %)	*'H3 lysine 9'*

Table 5.13: Patterns of the 10 most often occurring histone modifications encode how they are described in text. Given are pattern type, frequency, relative amount in %, and example terms. Abbreviation definitions:
H: Histone type,
As: Amino acid type,
Pos: Position of the amino acid,
Modi: Type of modification.

that different article numbers are retrieved when using varying query terms, because they are split into its word fragments and searched by co-occurrence. This could furthermore lead to false positive documents provided to the user.

To demonstrate the advantage of both developed histone modification recognition approaches combined with the term canonicalization process, the document retrieval was tested on complete MEDLINE. From the obtained recognition results the article identifiers from MEDLINE, which are PMIDs, were collected, unified and counted for every canonicalized histone modification concept. Table 5.14 presents the number of obtained articles for some frequently described histone modifications concepts when employing the listed terms in a PubMed query or using the canonicalization results of the CRF-based system and the dictionary-based system ProMiner incorporating HmodDict.

As the number of retrieved articles for the selected histone modification concepts clearly shows, all developed approaches considerably outperform the document search from PubMed. For instance, when searching PubMed for the term *'H3K9me'*, the number of documents provided is only a small fraction – 4.9 % – compared to number of articles obtained with the best performing term recognition approach ProMiner$_{\text{Hmod}}$ combined with the term canonicalization procedure. It demonstrates the high potential of the developed system.

Now it is also possible to retrieve and analyze published articles from MEDLINE that imply information on histone modifications in respect of publication year. For the first

Modification type	PubMed search	CRF		ProMiner$_{\mathrm{Hmod}}$
		I. model	E. model	
H3K9me	18	231	285	368
H3K4me	10	173	208	241
H3K4me3	92	104	171	190
H3K9me3	55	90	120	124
H3K9me2	61	80	113	145

Table 5.14: Number of articles from MEDLINE (version from December 2008) obtained for some selected histone modifications retrieved by a PubMed search, with CRFs using the initial (I.) and extended (E.) model (before and after active learning) for term recognition or ProMiner$_{\mathrm{Hmod}}$. The recognized terms were canonicalized by the workflow described in Section 5.2 above.

time this provides a global picture of the histone modification research history based on literature, which was not possible to analyze before. Figure 5.8 shows the result of the investigation depicting the distribution of the accumulated publication number for single selected modification types that have been released during a time span of 1962 and 2008. Concepts that do not describe modifications were omitted.

The array of curves shows quite different publication rate distributions for single histone modification types in the last 46 years. In general, the publication rate on unspecific histone modification descriptions is much higher compared to specific ones in this period. The highest number of publications and a similar curve progression was observed for descriptions on the acetylation and phosphorylation of histones in the time span from 1965 to 1995. Subsequent, the publication rate for *Hac* has risen much stronger since 1995. In this year enzymes of the class histone acetyltransferases (HATs) have been isolated that are responsible for the acetylation of histone's lysine residues [Kimura et al., 2005].

In comparison, the fraction of publications on specific histone modifications that specify the amino acid, its position and the type of modifications is much smaller. However, an increase in publication number can also be observed for these modification types within the last 12 years. This could be due to progress in experimental methods that made it easier to analyze histone modifications by high throughput methods. On the other side, since abstracts have been analyzed for this study they might not represent the information of the complete publications. Hence more specific histone modification types could have been identified if the publications would be available as full text.

5.5.3 Application of the Histone Modification Finding Results

As discussed in Section 3.3 visualization is still an open area of research in text mining for making the contents of large document collections easier to navigate, which hence supports an easier finding of interesting information.

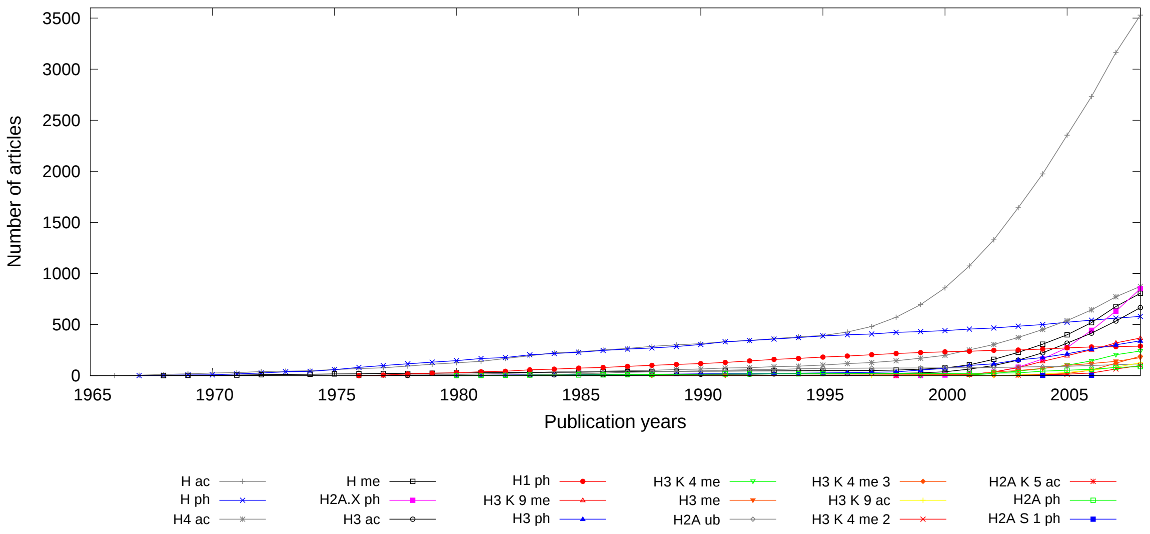
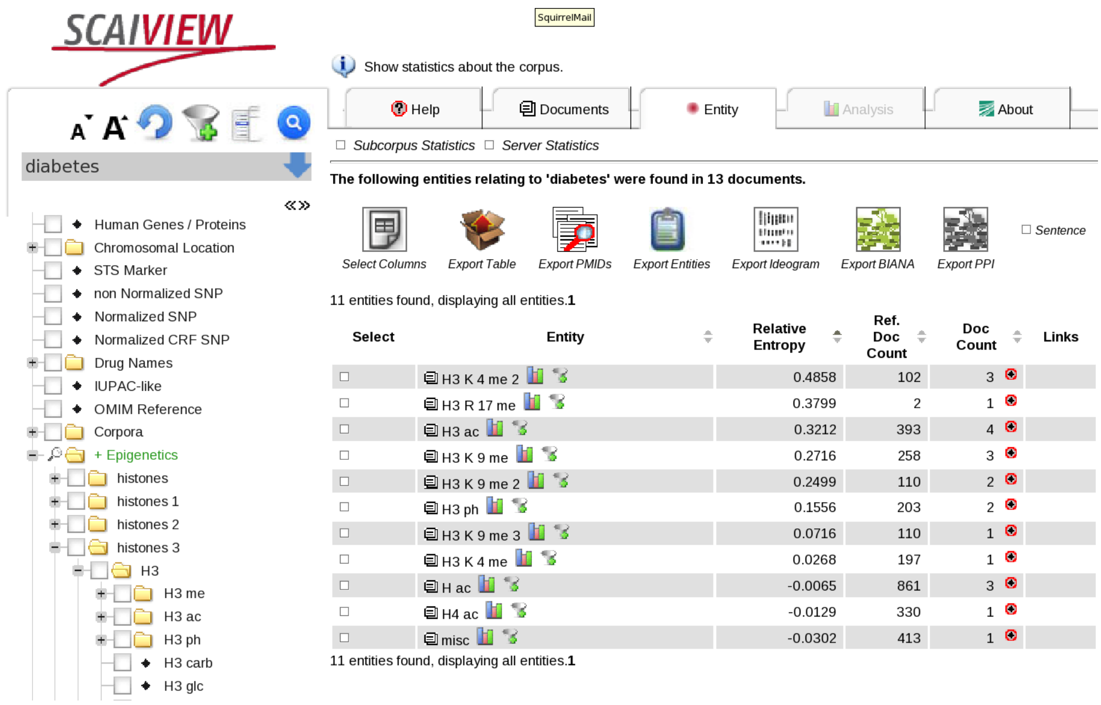


Figure 5.8: Distribution of the accumulated publication number for selected histone modification concepts from 1962 – 2008.

Information Retrieval Improved by Named Entity Identification The hierarchical organization of histone modification concepts which are linked to their denominations in text and its graphical presentation provides a good possibility to accomplish this task. Every concept node in the hierarchy is related to corresponding entities identified in the utilized text corpus. Furthermore, it provides identified entities of its subnodes by default. This allows for semantic histone modification search at different levels of granularity and for the confinement of articles to be considered for document retrieval.

Adequate to the integration of NER recognition results of chemical entities into SCAIView, histone modification concepts identified in articles of MEDLINE have been included into SCAIView. Hence, the filtering of large text corpora for articles that contain histone modifications is now possible at a semantic level. This can be realized by selecting all or specific concepts of the provided histone modification hierarchy and a query term. Furthermore, the text corpus can be filtered for further entity types, like diseases, chemicals, genes or proteins. This additionally specifies the search and might reduce the number of articles to be investigated.

An exemplary application scenario is the investigation of histone modifications related to *diabetes* – an onset disease appearing in adulthood and already used for the application of chemical entities in Section 4.1.4. SCAIView provides those modification concepts which are probably highly connected to the selected disease. This is shown by a list of histone modification concepts ranked according to their relative entropy. Thus one obtains a good overview on histone modification concepts relevant in a certain defined subfield. Since entities are related to respective MEDLINE articles the huge amount of articles are filtered at the same time so that only relevant articles related to the query are provided to the user.



Show statistics about the corpus.

The following entities relating to 'diabetes' were found in 13 documents.

11 entities found, displaying all entities.1

Select	Entity	Relative Entropy	Ref. Doc Count	Doc Count	Links
□	H3 K 4 me 2	0.4858	102	3	
□	H3 R 17 me	0.3799	2	1	
□	H3 ac	0.3212	393	4	
□	H3 K 9 me	0.2716	258	3	
□	H3 K 9 me 2	0.2499	110	2	
□	H3 ph	0.1556	203	2	
□	H3 K 9 me 3	0.0716	110	1	
□	H3 K 4 me	0.0268	197	1	
□	H ac	-0.0065	861	3	
□	H4 ac	-0.0129	330	1	
□	misc	-0.0302	413	1	

11 entities found, displaying all entities.1

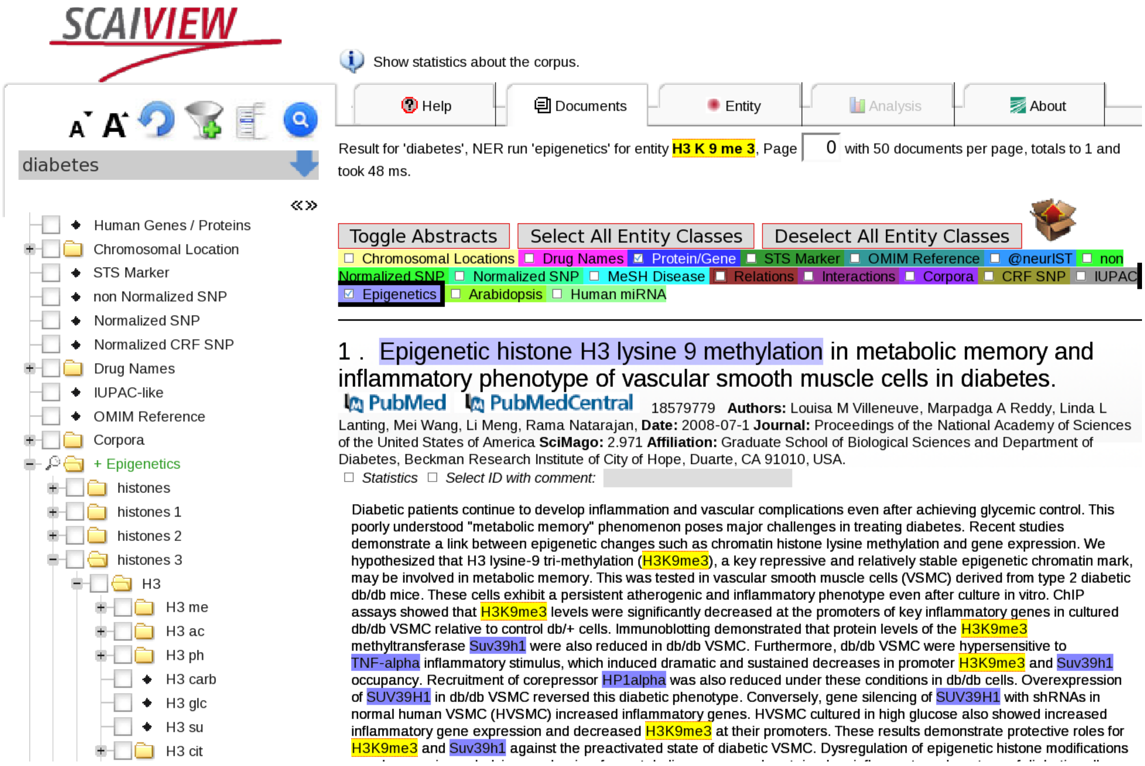
Figure 5.9: Screenshot of SCAIView showing histone modification entities that are related to *diabetes*. The disease term was used as query.

Figure 5.9 at page 181 illustrates the obtained result.

Histone modifications related to *diabetes* that emerged are almost of the methylation type of the amino acids lysine and arginine at histone 3 on several N-terminal protein sequence positions. As epigenetic chromatin marks can be related to gene repression or activation, their changes at certain chromatin positions in certain cell types might result in a stable expression pattern change of several genes. For more specific information the articles sorted according to the modification types have to be studied. In the document view proteins and genes, drugs, etc. can be highlighted, providing a fast overview on further described biomedical entities. As an example, the publication of [Villeneuve et al., 2008] is shown in Figure 5.10. Here, the proteins TNF-alpha, a cytokine involved in systemic inflammation, HP1 alpha, a fundamental protein in heterochromatin formation, and SUV39H1, a H3K9me3 methyltransferase, are highlighted.

The application example illustrates how the result of the histone modification identification in text in combination with the generated concept hierarchy can be utilized in advanced document retrieval systems like SCAIView.

Co-occurrence Network Construction based on Named Entity Identification A question which can be answered with the developed system is which diseases, genes, and proteins have been predominantely investigated in relation to histone modifications. Scientific articles, such as those provided by MEDLINE, embody an answer in form of discussed research results, reviews, etc. However, it is not easy to obtain, because information on diseases


SCAIVIEW
diabetes
Show statistics about the corpus.
Help
Documents
Entity
Analysis
About
Toggle Abstracts
Select All Entity Classes
Deselect All Entity Classes
□ Chromosomal Locations □ Drug Names ☑ Protein/Gene □ STS Marker □ OMIM Reference □ @neurIST □ non Normalized SNP □ Normalized SNP □ MeSH Disease □ Relations □ Interactions □ Corpora □ CRF SNP □ IUPAC ☑ Epigenetics □ Arabidopsis □ Human miRNA
Human Genes / Proteins
Chromosomal Location
STS Marker
non Normalized SNP
Normalized SNP
Normalized CRF SNP
Drug Names
IUPAC-like
OMIM Reference
Corpora
+ Epigenetics
histones
histones 1
histones 2
histones 3
H3
H3 me
H3 ac
H3 ph
H3 carb
H3 glc
H3 su
H3 cit
PubMed
PubMedCentral
□ Statistics □ Select ID with comment:

Result for 'diabetes', NER run 'epigenetics' for entity H3 K 9 me 3, Page 0 with 50 documents per page, totals to 1 and took 48 ms.

1 . Epigenetic histone H3 lysine 9 methylation in metabolic memory and inflammatory phenotype of vascular smooth muscle cells in diabetes.

18579779 **Authors:** Louisa M Villeneuve, Marpadga A Reddy, Linda L Lanting, Mei Wang, Li Meng, Rama Natarajan, **Date:** 2008-07-1 **Journal:** Proceedings of the National Academy of Sciences of the United States of America **SciMago:** 2.971 **Affiliation:** Graduate School of Biological Sciences and Department of Diabetes, Beckman Research Institute of City of Hope, Duarte, CA 91010, USA.

Diabetic patients continue to develop inflammation and vascular complications even after achieving glycemic control. This poorly understood "metabolic memory" phenomenon poses major challenges in treating diabetes. Recent studies demonstrate a link between epigenetic changes such as chromatin histone lysine methylation and gene expression. We hypothesized that H3 lysine-9 tri-methylation (H3K9me3), a key repressive and relatively stable epigenetic chromatin mark, may be involved in metabolic memory. This was tested in vascular smooth muscle cells (VSMC) derived from type 2 diabetic db/db mice. These cells exhibit a persistent atherogenic and inflammatory phenotype even after culture in vitro. ChIP assays showed that H3K9me3 levels were significantly decreased at the promoters of key inflammatory genes in cultured db/db VSMC relative to control db/+ cells. Immunoblotting demonstrated that protein levels of the H3K9me3 methyltransferase Suv39h1 were also reduced in db/db VSMC. Furthermore, db/db VSMC were hypersensitive to TNF-alpha inflammatory stimulus, which induced dramatic and sustained decreases in promoter H3K9me3 and Suv39h1 occupancy. Recruitment of corepressor HP1alpha was also reduced under these conditions in db/db cells. Overexpression of SUV39H1 in db/db VSMC reversed this diabetic phenotype. Conversely, gene silencing of SUV39H1 with shRNAs in normal human VSMC (HVSMC) increased inflammatory genes. HVSMC cultured in high glucose also showed increased inflammatory gene expression and decreased H3K9me3 at their promoters. These results demonstrate protective roles for H3K9me3 and Suv39h1 against the preactivated state of diabetic VSMC. Dysregulation of epigenetic histone modifications

Figure 5.10: Selected example MEDLINE article provided by the document view of SCAIView. The respective histone modification concept H3K9me3, proteins and genes are highlighted.

and proteins related to epigenetic mechanisms is scattered over a large pool of natural language data. The analysis of this huge text collection can be supported by Named Entity Recognition whose results are leveraged to build a co-occurrence network for visualizing the extracted data. It is based on the hypothesis that entities which co-occur in the same article or sentence are functionally linked. Potentially interrelated entities from articles of diverse subfields are collected, pair-wise occurrences determined and included into a graphical representation. As this is a simple technique, it has been widely used for analyzing textual data [Rodriguez-Esteban, 2009]. Such a network allows to identify those entities – network nodes – that posses a high connectivity to other entities (also called hubs [Barabási and Oltvai, 2004]).

The co-occurrence network construction was performed by following procedure: histone modifications, human proteins or genes, and MeSH terms of the MeSH subhierarchy C as well as parts of G and F, that were recognized and identified with ProMiner, were utilized to generate three independent article-entity type indexes. The three obtained entity type indexes were utilized to build up network connections of histone modifications with proteins/genes or diseases if at least one or more of every entity type co-occurred in the same article. This constraint should decrease the probability that a histone modification is related to both a protein and a disease only by chance. The number of co-occurrences between a histone modification h and a protein/gene or a disease term (denoted as pd) was determined, which is called $h(pd)_{Tri}$ here. For normalization the frequency of all co-occurrences between a histone modification and a protein/gene or disease, $h(pd)_{Di}$, was taken. Here only two entity types were needed to be found in the same article as constraint. The found entities were utilized to establish a histone modification protein/gene, disease co-occurrence network, whereas the computed ratio R (cf. Formula 5.1) was taken as one of two constraints for filtering and generating a subnetwork.

$$R = \frac{h(pd)_{Tri}}{h(pd)_{Di}} \tag{5.1}$$

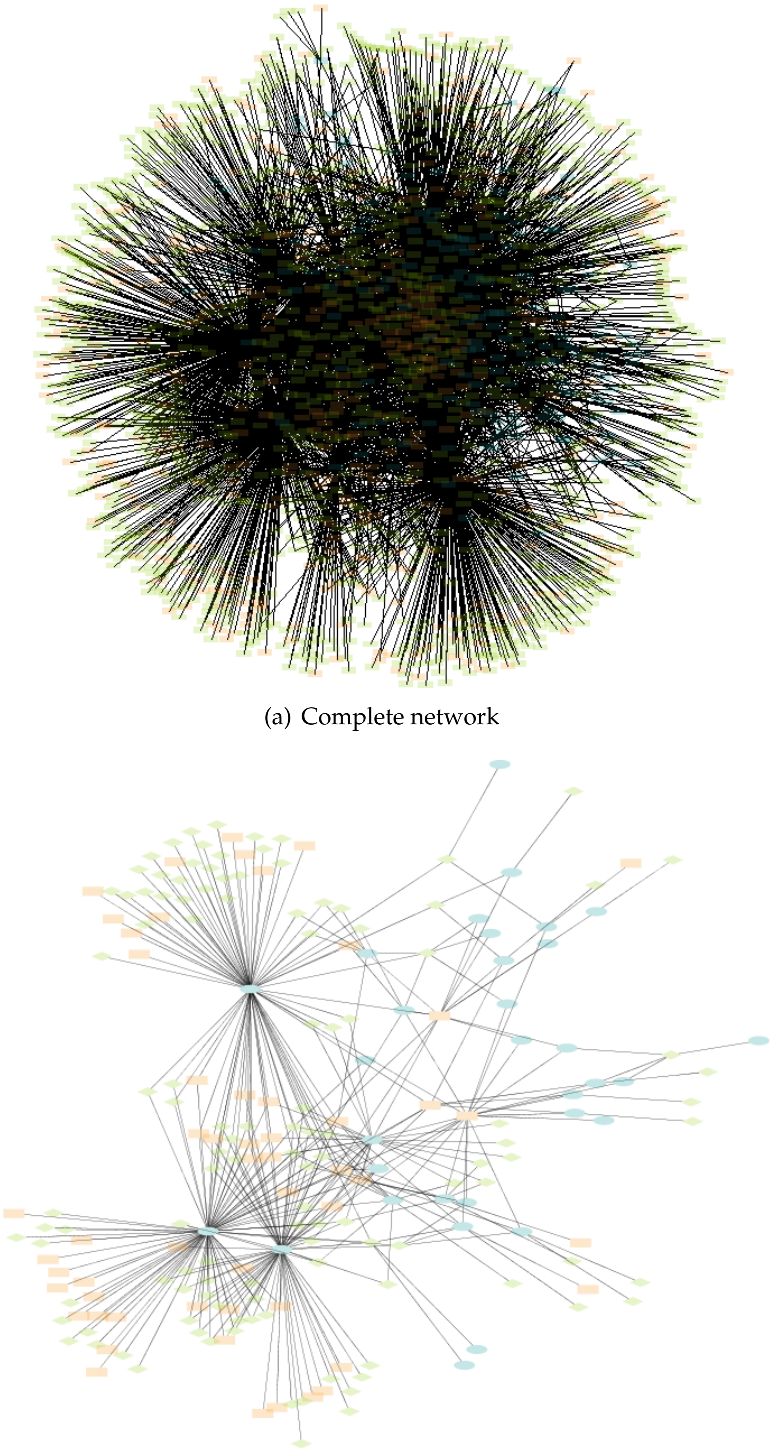
The basic network is shown in Figure 5.11 (a) and the filtered subnetwork in (b) (cf. next page). The latter excludes those entity relations with $R < 0.3$ and $h(pd)_{Tri} < 4$ for obtaining a sufficiently clear network.

The analysis of subnetwork revealed that all entities are interrelated. Acetylated histones 4 and 3 as well as phosphorylated histone H2A.X are related to many proteins/genes and diseases, thus constituting the hubs in the network. A list of the most highly connected entities of the network are shown in Table 5.15. Short descriptions of its molecular or biological roles are given as well.

As can be seen, the highly connected entities of the subnetwork illustrate that most textual data are related to research in cancer. Histone modifications with elevated connectivity are less specified. They are involved in the transcription activation of genes in general or mechanisms responsive to DNA damage. The high connectivity of cancer types and molecular mechanisms that point to the damage of DNA suggest a complex epigenetic mechanism. According to the general view, the found genes and proteins are linked to the development of cancer. However, for identifying its role respective to the chromatin modification machinery more detailed literature studies have to be conducted and further information, like pathway or protein interaction data should be added.

For identifying the most prevalent disease classes of the network, the MeSH hierarchy superclasses of every found MeSH disease term was obtained and the frequency of respective superclass occurrences was determined. The list of the most often emerging superclass terms are shown in Table 5.16. It further revealed that the most prevalent disease of the network belongs to the class of neoplams. Both findings, the prevalent histone modification related to gene expression regulation and the main found disease, reflect the fact that epigenetic mechanisms are involved in the development of cancer [Esteller, 2007]. On the other side it might also reflect that most research related to epigenetic mechanisms and disease was done in the field of cancer. Nevertheless, the collection of MeSH disease terms also demonstrates that epigenetic phenomena are related to many different disease classes and might thus be involved in the development of many diseases.

Subsequently, an enrichmet analysis of Gene Ontology terms that are related to the network proteins/genes was performed to obtain an insight into the their molecular function and related biological processes. Therefore, the obtained protein/gene list was analyzed by GOEAST [Zheng and Wang, 2008]. They use a hypergeometric test to identify significantly, species-specifically enriched GO terms among a given list of genes. As background they utilize GO terms of probes available on species-specific microarrays from diverse platforms of



(a) Complete network

(b) Subnetwork

Figure 5.11: Network of co-ocurring histone modifications, proteins or genes, and diseases extracted from MEDLINE articles generated with Cytoscape [Shannon et al., 2003]. Histone modifications are depicted as red ellipses, human proteins or genes as green diamonds, and diseases as red rectangles. In (a) all relations are shown, (b) presents the same network that was filtered and shows only those co-occurring entities with $R > 0.3$ and $h(pd)_{Tri} > 3$.

Relations to other entities	Entity	Molecular and biological role
85	H4ac	Gene activation (A)
81	H2A.Xph	In chromatin micro-environment of DNA double-strand breaks caused by ionizing or UV irradiation (B)
71	H3ac	Gene activation (A)
28	H3ph	Gene activation (A)
20	Neoplasms	
14	DNA Damage	
10	H1ph	Chromatin decondensation (K)
9	H3S10ph	Gene activation (A)
9	Mammary Carcinomas, Human	
8	Leukemia	
7	H4K20me3	Heterochromatin formation and gene silencing (C)
7	H3K9me	Heterochromatin formation and gene silencing (C)
7	H2Aph	In chromatin micro-environment of DNA double-strand breaks caused by ionizing or UV irradiation (D)
7	TP53	Transcription factor, inductor of apoptosis, tumor suppressor, works together with CDKN1A (E)
6	Prostatic neoplasms	
6	HIST4H4	Encodes histone H4 protein
6	H2AFX	Encodes histone H2A protein variant
6	Genome Instability	
6	H2A.XS139ph	In chromatin micro-environment of DNA double-strand breaks caused by ionizing or UV irradiation (D)
5	H3K27me3	Heterochromatin formation and gene silencing (C)
5	MAPK14	MAP kinase, responsive to stress stimuli, involved in cell differentiation & apoptosis (I), phosphorylates TP53 (F)
5	Hereditary Retinoblastoma	
5	DNMT1	Maintenance DNA methyltransferase, co-operating with histone modifications in gene silencing (G)
5	CDKN1A	Kinase, cell cycle regulator, tumor suppressor (E)
5	CASP3	Caspase, important role in apoptosis (H)

Table 5.15: Number of co-occurrence relations between histone modification types and proteins, genes or diseases of the subnetwork. The biological role of histone modifications and proteins are given. References: (A) [Lo et al., 2004, Zippo et al., 2009], (B) [Rogakou et al., 1998, Hanasoge and Ljungman, 2007], (C) [Richards and Elgin, 2002, Kourmouli et al., 2004], (D) [Rogakou et al., 1998, Hanasoge and Ljungman, 2007], (E) [el Deiry et al., 1993], (F) [Lafarga et al., 2007], (G) [Robertson, 2002], (H) [Soung et al., 2004], (I) [Bulavin et al., 2001], (K) [Sarg et al., 2006]

Frequency	MeSH term superclass
45	C04 Neoplasm
16	C06 Digestive System Diseases
12	C23 Pathological Conditions
9	C08 Respiratory Tract Diseases
8	G05 Genetic Phenomena
7	C15 Hemic and Lymphatic Diseases
7	C13 Female Urogenital Diseases and Pregnancy Complications
6	C20 Immune System Diseases
6	C18 Nutritional and Metabolic Diseases
5	C10 Nervous System Diseases
3	C16 Congenital, Hereditary, and Neonatial Disease and Abnormalities
3	C12 Male Urogenital Diseases
2	C21 Disorders of Environmental Origin
2	C19 Endocrine System Diseases
2	C17 Skin and Connectivity Tissue Diseases
2	C14 Cardiovascular Diseases
2	C11 Eye Disease
1	C01 Bacterial Infections and Mycoses

Table 5.16: Frequency of the most occurring MeSH superclasses that are related to found MeSH terms in the generated co-occurrence network.

Affymetrix[5], Illumina[6], and Agilent[7]. GOEAST corrects the raw P-values with the Benjamini-Yekutieli method [Benjamini and Yekutieli, 2001] by default. The obtained GO terms were filtered to exclude the Molecular component part of the Gene Ontology and the two highest levels of Molecular function and Biological process GO parts because of their low information content. For lack of space only the first 38 most significant GO terms and respective p-values are provided in Table 5.17. The extended GO term list can be found in the Appendix B.1.

The enriched GO terms contain information about molecular mechanisms which confirm the picture of the most present disease and genetic mechanisms – cancer and DNA damage – in the network. Most of the over represented terms are related to cell proliferation regulation and stress responses. Additionally, a high amount of significantly enriched GO terms belong to metabolic process regulation which points to the known fact that an altered metabolism is a further essential hallmark of cancer cells [Kroemer and Pouyssegur, 2008].

The chosen example demonstrates the value of Named Entity Recognition and Identification. It shows how co-occurrence relations of entity mentions found in text can help to obtain a literature-wide view on present research in a selected field, like epigenetics. The

[5] http://www.affymetrix.com/estore/

[6] http://www.illumina.com/

[7] http://www.home.agilent.com/agilent/home.jspx?cc=US&lc=eng

GO term	p-value
regulation of cell cycle	8.62e-46
negative regulation of cellular process	9.72e-37
positive regulation of cellular process	1.06e-36
cellular response to stimulus	4.09e-36
regulation of cellular process	3.24e-35
regulation of metabolic process	7.23e-35
regulation of apoptosis	1.20e-33
response to stress	1.60e-33
regulation of programmed cell death	1.84e-33
regulation of cell death	2.70e-33
protein binding	6.04e-33
cellular response to stress	1.80e-32
DNA metabolic process	1.89e-31
response to DNA damage stimulus	7.16e-31
positive regulation of metabolic process	9.54e-31
positive regulation of cellular metabolic process	1.07e-30
cellular macromolecule metabolic process	3.65e-30
regulation of cellular metabolic process	3.81e-30
cell cycle	3.99e-30
regulation of macromolecule metabolic process	2.16e-29
regulation of cell proliferation	4.59e-29
positive regulation of macromolecule metabolic process	6.84e-29
regulation of nitrogen compound metabolic process	1.09e-27
cell cycle process	3.46e-27
cell cycle checkpoint	4.33e-27
positive regulation of nitrogen compound metabolic process	5.14e-27
response to abiotic stimulus	8.63e-27
response to stimulus	1.06e-26
negative regulation of cellular metabolic process	1.07e-26
regulation of DNA metabolic process	1.78e-26
chromosome organization	1.89e-26
negative regulation of metabolic process	3.80e-26
regulation of primary metabolic process	5.76e-26
organelle organization	8.60e-26
response to radiation	2.14e-25
response to chemical stimulus	2.45e-25
regulation of mitotic cell cycle	4.76e-25
DNA damage response, signal transduction	5.18e-25

Table 5.17: Most significant Gene Ontology terms of the subontologies Molecular function and Biological process related to proteins and genes in network. Respective p-values obtained by an enrichment analysis via GOEAST are given. (Due to space only 38 terms are given. The longer list can be found in the Appendix B.1)

combination with further related information, like the annotation classes of proteins and its analyses, allows for getting a fast and more detailed insight into the respective area. However, a drawback of this technique is the omission of relation types between entities which carry more detailed information and the disregard of negation of relations between entities, which could lead to false positive findings. Nevertheless, co-occurrence networks can be used for getting an overview on a subfield or as starting point for further investigations, e.g. for identifying new research topics. The integration with further data, like pathway information, can lead to the generation of new hypotheses which drive research in new directions.

5.6 Summary and Discussion

The goal of the third task was to develop a system for recognizing histone modification descriptions in text and to map different representations to one canonical term form. Since there was no terminology available ready to be used in a dictionary-based approach, the state-of-the-art machine learning based NER method Conditional Random Fields has been chosen. Therefore, three corpora were annotated from which one was utilized as training corpus to obtain an initial CRF model. This was tested on the other two evaluation corpora yielding good recognition results. To test if the performance can be increased by selective extension of the training corpus it was extended by the active learning technique. With the extended final model a further improvement of the approach could be obtained leading to a high performance with a precision of 0.93 and recall of 0.81.

As authors seldom consider the devised *Brno* nomenclature for histone modification descriptions in scientific text, a procedure was developed and implemented for mapping different term variants and synonyms of individual histone modification concepts to defined term representations. The canonicalization was realized by transforming the information provided by the recognized terms to canonical term forms which follow the definition of the *Brno* nomenclature. Therefore, a program was implemented which consist of consecutively arranged rule sets that check the terms for defined information units. Additionally, the workflow removes false positive recognized terms. The performance, measured by the fraction of correctly transferred terms, was shown to be high. 96 % and 98 % of the two manually transformed term sets were automatedly transformed correctly. The mapping of different term variants to one representative concept denomination is of high value for the improvement of retrieval systems which is discussed below. However, long terms exhibiting wrong boundaries that were recognized by the CRF model provide a limitation of the canonicalization process. This is caused by the fact that the transformation rule sets were developed from histone modification terms of the annotated training and testing corpora which possess correct term boundaries. Two procedures could lead to an improvement:

a) The annotation of more training data used for the generation of the CRF model which could hence lead to a decrease of the recognition of terms with wrong boundaries and

b) A further adaption of the canonicalization process.

The latter one might require less efforts than the first one, because it is not clear how much training data is needed to reduce the number of terms with wrong boundaries.

A further advantage of the term mapping procedure is the possibility to generate a synonym dictionary in a straight forward way. Therefore, terms recognized with the CRF served as synonym source for the generation of the term list to be included into a dictionary-based NER method like ProMiner. This experiment was conducted because it was shown that the CRF-based approach has the weakness of not finding every modification term in the document collection, although it was given in the training data. In contrast, dictionary-based approaches are expected to find the given terms in every case. Hence, $\text{ProMiner}_{\text{Hmod}}$ and the CRF approach were run on complete MEDLINE and the number of entities obtained by the subsequent canonicalization procedure were compared. It could be demonstrated that the dictionary-based approach leads to more canonicalized histone modification entities than the CRF-based method. Hence, for obtaining more reliable results the $\text{ProMiner}_{\text{Hmod}}$ should be preferred when a term source is available or when a term list can be generated. However, it has the disadvantage that newly coined terms are not covered by the present dictionary version. Thus, the dictionary generation procedure has to be repeated from time to time.

The results obtained by the procedures described above have been included into the knowledge discovery system SCAIView. To allow semantic search at different levels of granularity a histone modification concept hierarchy was added to SCAIView as well. It was assembled from diverse resources, as the investigation of considerable ontologies and thesauri like Gene Ontology and MeSH revealed that there is no single comprehensive resource available. The concept hierarchy implies 462 concepts at 7 levels of granularity. To automatedly support its prospective extension with new modification concepts appearing in text, an expansion procedure was developed. It proposes new concepts for manual selection and inclusion into the present hierarchy. They are based on terms extracted from text, transformed to standard term forms and sorted by its frequency.

By means of several application scenarios the value of the newly developed systems was demonstrated. In the first application the occurrence of histone modification concepts in MEDLINE and the arrangement of typical information units implied in the recognized terms have been investigated. These were histone, modification, position, and amino acid displayed as information unit patterns. It revealed that most publications in MEDLINE deal with the acetylation, phosphorylation and methylation of histones in general. This is followed by methylations of lysines at several histone tail positions. This information content is also reflected by the frequency of corresponding information unit patterns of the terms.

The second utilization dealt with the retrieval of documents covering information on selected histone modification concepts. Therefore, the number of retrieved articles from MEDLINE obtained by a PubMed search or the collection with the new approaches were compared with each other. It was shown that a PubMed query could by far not retrieve that many articles with the query term following the nomenclature as was identified with the newly developed system. It thus allocates more comprehensive information to a researcher than PubMed. Furthermore, for the first time a historical analysis of the publication rate per year for every histone modification type over the last five decades is possible. It illustrated the high increase of publication for some general histone modification descriptions, like *Hac* and *Hph* and most of the more specific types in the period from 1995 to 2000.

The application of the histone modification hierarchy and recognized concepts in a knowledge discovery system like SCAIView was illustrated by a further scenario. It demonstrates the improved navigation, retrieval, and analysis of documents from a large text corpus.

In a fourth scenario the generation of a co-occurrence network consisting of histone modifications, proteins/genes and diseases is described. It demonstrates the high relation of epigenetic mechanisms to cancer-related diseases and DNA damage. However, the network analysis also revealed that histone modifications are connected to many diverse diseases classes. The co-occurrence network show that information visualized this way can be used for getting an overview on a subfield. Futhermore, it can be utilized as a starting point for further investigations. Together with the integration of additional data to this network, such as pathway information, it can lead to a better understanding of the linkage of epigenetic mechanisms with cellular processes and the generation of new hypotheses which drive research in new directions.

In summary, the developed approaches support the improvement of information retrieval in the field of epigenetics. Furthermore, they build the basis for further information extraction tasks that harness the wealth of information contained in scientific articles. It can be used to generate networks including histone modifications, proteins/genes and diseases for finding new interrelations between them that were hidden in text so far. Additionally, it can support the assembly of structured information resources, like databases, that are based on scientific findings on histone modifications residing in text.

Chapter 6

Conclusion and Perspectives

6.1 Conclusion

In this thesis a framework was developed that supports the aggregation of function annotation information on chemical entities from structured and unstructured resources. It comprises two challenges: the recognition and identification of chemical entities in text combined with its mapping to a unique representation and the extraction of related function information.

The recognition of chemical named entities (NER) and its normalization/identification (NEI) constitute a basic step to make unstructured natural language text applicable for more complex information extraction and successive data mining tasks. Therefore, a dictionary-based method was chosen for the chemical named entity recognition since it easily allows for entity identification, which is the mapping of different denominations onto one standard representation. It required the investigation of the terminology that is utilized for the description chemical molecules and the analysis of respective available terminology resources to build up a chemical entity dictionary. The terminology was collected from 7 data resources which provide a different grade of comprehensiveness in the number of chemical entities and related synonyms. In order to map recognized names of chemical entities to structural representations and reduce ambiguity in the dictionary, according chemical entities residing in separate resources were merged. Therefore, InChI and CAS identifiers as well as synonyms were utilized for terminology joining, whereof the challenging synonym merging is discussed in detail in this work. Furthermore, it was shown that the curation – the processing of the synonyms – highly improves the performance of the utilized dictionary-based system. Names of chemical entities recognizable via the dictionary approach $ProMiner_{Chem}$ in text are thus normalizable to single entities as well as structural representations which enables semantic search in successive information retrieval approaches. The performance of the resulting $ProMiner_{Chem}$ is comparable to the dictionary-based approach Peregrine which was developed by Schuemie et al. [2007] in parallel to this work. It includes a termlist that was generated from a similar set of terminology resources and thus allows to draw the conclusion that dictionary-based methods are limited by an upper performance bound. It emerged that $ProMiner_{Chem}$ is well suited for the recognition of the class of trivial names, for which it achieved the highest recall, whereas in contrast, the system has a low performance in finding systematic names in text.

To combine function annotation information from literature with those residing in structured sources, such as chemical content databases a new framework has been established that enables the excerption of information linked with chemicals from text. Therefore, the linguistic conception of hypernymic phrases was utilized to extract phrases which describe hierarchical relationships between chemical entities and function or property terms. The framework includes the recognition of chemical named entities as a first step to locate respective text corpora and to finally filter the obtained results. It could be shown that function annotation information could be harvested from such phrases in form of terms. They embody a high percentage of new pharmacological class information not yet applied as function annotation of chemical entities in comparison to two chosen reference resources; the database DrugBank and the drug classification system ATC. Concluding, this developed framework helps to find new annotation information for chemical entities on the basis of literature data in an automatic way. Furthermore, it supports a faster extension of function annotation of chemical entities in databases, class hierarchies and ontologies. As the procedure was developed generic, further entity classes can be annotated in a similar way in the future through exchanging the chemical named entity dictionary used for the corpus definition and phrase filtering. Succeedingly, the developed framework was successfully applied on a large chemical entity set for which respective function annotation terms were extracted from text.

As pharmacological classification systems used for function annotation of chemical entities, such as ATC, are incomplete, an automated support of its extension can help to improve these sources and make them more valuable for pharmacological and biomedical research. Therefore, it was investigated whether the pharmacological function concepts that were harvested for the set of chemical entities from text are ample to characterize chemical entities in order to compare or classify them. Therefore, a novel feature vector was assembled that specifies chemical entities through pharmacological concepts found in text. It differs from existing approached fingerprints that are almost exclusively based on structural information and/or physicochemical properties. It was shown that this new descriptor can be utilized to successively predict classes of the ATC system for chemical compounds not covered by this scheme. The ATC class prediction approach was evaluated on a subset of the ATC system and disclosed a high performance. Thus, it is supposed that the expansion of the workflow on complete ATC supports its extension by novel chemical instances and would lower the number of chemical entities that are missing in the classification system on principle. This would broaden its scope and would make ATC a more valuable resource for network studies that are conducted in pharmacological research for example. Concluding, analog to the work of Campillos et al. [2008] the thesis shows that pharmacological concepts extracted from textual data provide a valuable source for the comparison and classification of chemical compounds in order to find new functions for present chemical entities.

As the highly topical and almost all context information on histone modifications resides in scientific articles in the form of natural language, text is an important resource for building hypotheses in epigenetic research. Making text accessible to automated information extraction and retrieval thus helps to find information on histone modifications in a more elaborate way. For the first time this thesis approached the recognition and identification of histone modification descriptions in text. Therefore, text itself served as terminology resource from which histone modification descriptions were extracted via the Machine Learning method Conditional Random Fields (CRF). The recognized and extracted histone modification term variants were succeedingly mapped to standard representations following nomenclature rules through a novel developed term canonicalization approach. This new procedure supports the generation of a new histone modification dictionary applicable in ProMiner. It has been chosen as no other comprehensive terminology resource than text was available. It was demonstrated that the generation of dictionaries on the basis of extracted terminology from literature is a successful way, especially for new named entity types related to newly emerging topics. Furthermore, it would also allow for the completion of already existing term collections, such as Gene Ontology. The comparison of the recognition results on complete MEDLINE revealed that the dictionary-based approach $ProMiner_{Hmod}$ outperformed the CRF approach. It found more histone modification descriptions and also new spelling variants not yet included in the dictionary. In order to structure articles on histone modifications by different grade of histone modification information a concept hierarchy was established. Every item in the hierarchy is related to a standard term through which the hierarchy is connectable to entities found in text. Its integration into a knowledge management system, such as SCAIView, helps to conduct semantic search for information related to histone modifications on selectable levels of the histone modification hierarchy. As SCAIView now newly includes chemical entities, histone modifications, and many other yet established biomedically relevant concepts and entities identified in text, e.g. proteins, genes, cell types, diseases, etc., it is possible to conduct complex semantic text queries. Thus, scientific articles can be retrieved and sifted through in a new way to generate novel hypotheses in the field of biomedicine.

Concluding, this work investigated techniques to make the enormous amount of natural language data available in the chemical and biomedical domain amenable for automated methods. They thus support the improvement of document retrieval, entity annotation as well as hierarchy extension and ontology generation and thus the finding of new hypotheses, planning of experiments, etc. which are important to make progress in biomedical and pharmaceutical research.

6.2 Perspectives for Future Research

The presented thesis revealed several aspects and issues that should be pursued in the future for improving the performance of the approaches and for tackling unsolved challenges. Additionally, the developed approaches should be considered for further applications.

The investigation of the chemical named entity recognition with $ProMiner_{Chem}$ showed that this method is not able to recognize the entire chemical named entity space, because of the limitation of the generated chemical dictionary. For improving the recall of the chemical named entity recognition further resources should be considered for the extension of the chemical name dictionary. Additionally, the combination of $ProMiner_{Chem}$ with a complementary machine learning approach, like the IUPAC-tagger, would harness the advantages of both approaches which hence would increase the overall recall. However, it implicates the requirement to solve the normalization problem of chemical named entities recognized with the machine learning approach.

A further challenge that has to be tackled in the future is the resolution of co-references. Authors often introduce numbers or article-specific abbreviations that refer to systematic or semi-systematic chemical compound descriptions or specific side groups in scientific articles or patents. Therewith authors avoid the repetitive writing of long complicated chemical names which are usually defined once in a text document. In addition, abbreviations can be connected with further information residing in tables, e.g. certain properties, or linked with pictures providing the complete chemical structure or a Markush structure (definition cf. Appendix A.1). As this demonstrates, information on chemical compounds is not necessarily exclusively embodied within the textual part of documents. The extraction and combination of textual and picture information would thus provide improved capabilities to comprehensively navigate through the large amount of documents on chemical compounds.

Another issue is the facilitation to make textual documents amenable for chemical structure search methods. Therefore, identified chemical entities in textual documents which are connected to a machine-readable and structure-searchable format would be made accessible for structure search engines. This would broaden the document search capabilities and provide a new dimension for querying and accessing textual information.

In consideration of the chemical's property information extraction from text performed with the Hearst phrase pattern matching, it was shown that the obtained concepts imply new information not applied by the ATC classification scheme. They thus should be used for the extension of pharmacological drug classification schemes or ontologies by new classes or concepts. This closes information gaps and is of high importance for approaches that are based on them, such as network studies.

Methodically, the inclusion of further relation types could lead to the extraction of additional concepts. Thus, they would augment the space of pharmacological concepts newly available for qualitative function annotation of chemical compounds. The application of the approach on complete articles could furthermore result in a higher number of extracted concepts compared to abstracts. Such an extended approach would enhance the prediction of pharmacologically classes on chemical compounds and hence the extension of compound classification schemes by enriching the feature vector with further concepts. Beyond, the implication of additional pharmacological function concepts from other sources than text would enlarge the feature space to be considered and could help to handle textual data sparseness.

Altogether, pharmacological classification schemes extended by further compounds and/or pharmacological classes would allow more comprehensive analyses of information on drugs. When applied in network approaches, like performed by Nacher and Schwartz [2008], chemical entities can be set into a broader context which supports the finding of possibly new relations between compounds and diseases, pharmacological effects, adverse events, etc. to support drug repositioning for instance or to get new insights into the mechanism of action. Proteins – components of pathways – could for instance be a bridge between chemical compounds and histone modifications. Therewith, new insights can be obtained in the future how environmental factors, like drugs and nutraceuticals, or organisms internal compounds, like metabolites or hormones, influence the regulation of epigenetic mechanisms. The recognition of histone modification descriptions in text lays the basis for such studies. As described in the introduction, up to now text is the most comprehensive resource on histone modifications available as it embodies context information not yet contained in databases. Thus, the rapid growing number of publications in epigenetics might contain hidden information which has not been put into an overall context yet. The automated exploitation of text for research in epigenetics could help to fill open questions on mechanism which regulate histone modification insertion and its erasure. Furthermore, text contains descriptions on affected expression states of specific genes, its chromosomal positions, and studied cell types, diseases etc. Hence, the implication of textual information in epigenomic research can help to learn new aspects about the etiology of diseases and to stimulate the discovery of new agents, which modulate the epigenome in a therapeutic advantageous manner. On the other side, unexpected environmental toxic and pharmacological agents might target the class of chromatin modifying proteins or influence signaling pathways, thus affecting the long-term expression programming of many genes in diverse tissues. Therefore, research in drug development has to contemplate potential hazards to the epigenome in the future.

Appendix A

Definitions and Data Resources

A.1 Definitions

Tautomer Tautomers are constitution isomers of organic compounds that result by the relocation of a proton. In solutions where tautomerization is possible, a chemical equilibrium of the tautomers will be reached, whereas the ratio of the tautomers depends on several factors, including temperature, solvent, and pH-value [Sykes, 1988]. Figure A.1 shows examples of basic tautomeric variants.

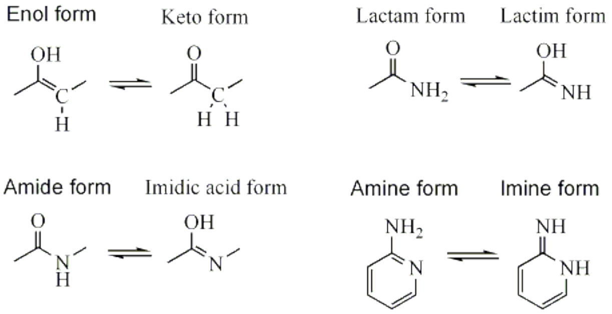


Figure A.1: The figure illustrates four basic different tautomery types. (It was adapted from [A1]).

Markush structure A Markush structure is a generalized formula or description for a related set of chemical compounds. It is named after Eugene Markush (1888–1968), an American manufacturer of dyes and pharmaceuticals. Markush structures, often provided by chemical patents, are used to describe compounds comprising substituents at several positions. Hence, often many thousands of possible compounds are defined in this way [M1, M2].

A.2 Data Resources

A.2.1 Data Sources on Chemical Compounds

- Commercial Databases
  - **CrossFire Beilstein database**[1] is a large repository for information of over 10 Million organical compounds. Beside structural information stored entities are associated with chemical and physical properties, bioactivity data, literature references as well as their environmental fates and reactions.
  - **CAS REGISTRY$^{\text{SM}}$** [2] provided by CAS, is one of the largest databases of chemical substance providing information back to the beginning of the late 19th century. It contains more than 33 million organic and inorganic substances which are related to calculated properties, like physico-chemical information and experimental property data.
  - **The World Drug Index**[3] is an authoritative index for marketed and development drugs. It contains chemical and biomedical data as well as synonyms for over 80,000 marketed and development drugs. Each record has a chemical structure and is classified by drug activity, mechanism of action, treatment, manufacturer, and medical information.
- Freely available Databases
  - **Kyoto Encyclopedia of Genes and Genomes (KEGG)**[4] is a composite database that integrates genomic, chemical, and systemic function information. According to Goto et al. [1998] is the intention of KEGG to computerize all molecular components and the network of molecular interactions to describe, utilize and predict function aspects of living systems. It comprises two subdatabases containing chemical structures of most known metabolic compounds *KEGG COMPOUND* and all approved drugs in the US and Japan *KEGG DRUG* [Kanehisa and Goto, 2000, Kanehisa et al., 2004, 2008].
  - **PubChem**[5] is part of the NIH Roadmap for Medical

    Research[6]. It focuses on the chemical, structural and biological properties of small molecules, particularly their application as diagnostic and therapeutic agents. PubChem consists of three linked databases – *PubChem Substance*, *PubChem Compound*, and *PubChem BioAssay*.
  - **DrugBank**[7] is a database about pharmaceuticals and nutraceuticals, that combines detailed chemical, pharmacological and pharmaceutical information, with drug target information [Wishart et al., 2006]. It was developed to facilitate

[1] http://www.info.crossfiredatabases.com/

[2] http://www.cas.org/expertise/cascontent/registry/regsys.html

[3] http://www.daylight.com/products/wdi.html

[4] http://www.genome.jp/kegg/

[5] http://pubchem.ncbi.nlm.nih.gov/

[6] http://nihroadmap.nih.gov/

[7] www.drugbank.ca

in silico drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction.

- **Human Metabolome Database (HMDB)**[8] is a database containing detailed information about metabolites found in the human body, like hormones, disease-associated metabolites, essential nutrients and signaling molecules as well as ubiquitous food additives and some common drugs [Wishart et al., 2007]. The focus lies on quantitative, analytic or molecular-scale information, metabolite associated enzymes or transporters and disease-related properties.

- Thesauri and Ontologies
  - **Medical Subject Headings (MeSH)**[9] is a controlled vocabulary thesaurus from the National Library of Medicine (NLM)[10]. It is used by NLM for indexing articles from the MEDLINE/PubMed database as well as a catalog database for other media of the library. The terms are organized in a hierarchy to which synonyms as well as inflectional term variants are assigned. A subset of the MeSH thesaurus (version 2007 MeSH) covering the chemical category of the MeSH hierarchy (tree concepts with an identifier starting with a 'D') was extracted to give one dictionary of MeSH (referenced further as MeSH-T). Furthermore, NLM provides a compound list with over 175,000 entries containing synonyms like trivial and brand names, IUPAC and abbreviations. This was used to generate another dictionary, referenced further as MeSH-C.
  - **Chemical Entities of Biological Interest (ChEBI)**[11] developed at European Bioinformatics Institute (EBI) contains 12 000 molecular entities, groups and classes. It catalogs small molecules, i.e. enzyme substrates and products, atoms, ions, ion pairs, radicals and other small chemical entities, which are related to the ChEBI ontology. The objective of ChEBI is to bridge the gap between proteins and small molecules as well as the correct and to ensure the consistent utilization of unambiguous biochemical terminology throughout the molecular biology databases at the EBI [Brooksbank et al., 2005, Degtyarenko et al., 2008].

A.2.2 Pharmacological Classification Schemes on Chemical Compounds

Anatomical Therapeutic Chemical (ATC) classification system The ATC/Defined Daily Dose (DDD) system[12] was developed in the 70the by the Norwegian Medicinal Depot for drug utilisation studies. Since 1996, the World Health Organization (WHO) recommends the ATC/DDD system as an international standard which is updated every year [Ronning, 2001]. The ATC classification system divides drugs into different groups according to the organ or organ system on which they act and their chemical, pharmacological and therapeutic properties. Drugs are classified in groups at following five levels, whereas the same substance may be assigned to different ATC codes:

[8] http://www.hmdb.ca/

[9] http://www.nlm.nih.gov/mesh/

[10] http://www.nlm.nih.gov/

[11] http://www.ebi.ac.uk/chebi/

[12] http://www.whocc.no/atcddd/

- **1st level:** Anatomical main group (e.g.: A *'Alimentary tract and metabolism'*)
- **2nd level:** Pharmacological/therapeutic main group (e.g.: A10 *'Drugs used in diabetes'*)
- **3rd level:** Chemical/pharmacological/therapeutic subgroup (e.g.: A10B *'Oral blood glucose lowering drugs'*)
- **4th level:** Chemical/pharmacological/therapeutic subgroup (e.g.: A10BA *'Biguanides'*)
- **5th level:** Subgroup for chemical substance (e.g.: A10BA02 *'Metformin'*)

The importance of the ATC classification is the possibility of the international comparability, the monitoring of drug utilization to study long term trends and consumption from various aspects [Skrbo et al., 1999].

Appendix B

Extended Results

B.1 Extended Result List of the Gene Ontology Term Enrichment Study from Section 5.5.3

Table B.1: Most significant Gene Ontology terms of the subontologies Molecular function and Biological process related to proteins and genes in network. Respective p-values obtained by an enrichment analysis via GOEAST are given.

GO term	p-value
regulation of cell cycle	8.62e-46
negative regulation of cellular process	9.72e-37
positive regulation of cellular process	1.06e-36
cellular response to stimulus	4.09e-36
regulation of cellular process	3.24e-35
regulation of metabolic process	7.23e-35
regulation of apoptosis	1.20e-33
response to stress	1.60e-33
regulation of programmed cell death	1.84e-33
regulation of cell death	2.70e-33
protein binding	6.04e-33
cellular response to stress	1.80e-32
DNA metabolic process	1.89e-31
response to DNA damage stimulus	7.16e-31
positive regulation of metabolic process	9.54e-31
positive regulation of cellular metabolic process	1.07e-30
cellular macromolecule metabolic process	3.65e-30
regulation of cellular metabolic process	3.81e-30
cell cycle	3.99e-30
regulation of macromolecule metabolic process	2.16e-29
regulation of cell proliferation	4.59e-29
positive regulation of macromolecule metabolic process	6.84e-29
regulation of nitrogen compound metabolic process	1.09e-27

GO term	p-value
cell cycle process	3.46e-27
cell cycle checkpoint	4.33e-27
positive regulation of nitrogen compound metabolic process	5.14e-27
response to abiotic stimulus	8.63e-27
response to stimulus	1.06e-26
negative regulation of cellular metabolic process	1.07e-26
regulation of DNA metabolic process	1.78e-26
chromosome organization	1.89e-26
negative regulation of metabolic process	3.80e-26
regulation of primary metabolic process	5.76e-26
organelle organization	8.60e-26
response to radiation	2.14e-25
response to chemical stimulus	2.45e-25
regulation of mitotic cell cycle	4.76e-25
DNA damage response, signal transduction	5.18e-25
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.85e-24
regulation of cellular biosynthetic process	3.07e-24
negative regulation of macromolecule metabolic process	3.71e-24
regulation of biosynthetic process	4.73e-24
DNA damage checkpoint	5.46e-24
regulation of macromolecule biosynthetic process	8.79e-24
DNA integrity checkpoint	2.10e-23
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	5.60e-23
negative regulation of cellular biosynthetic process	6.31e-23
DNA repair	8.45e-23
negative regulation of programmed cell death	1.24e-22
negative regulation of biosynthetic process	1.25e-22
negative regulation of cell death	1.94e-22
negative regulation of macromolecule biosynthetic process	3.09e-22
cell cycle phase	6.89e-22
negative regulation of nitrogen compound metabolic process	7.38e-22
positive regulation of cellular biosynthetic process	1.40e-21
regulation of molecular function	1.60e-21
positive regulation of biosynthetic process	2.54e-21
negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6.91e-21
positive regulation of gene expression	8.55e-21
nucleic acid metabolic process	8.87e-21

GO term	p-value
positive regulation of macromolecule biosynthetic process	2.34e-20
negative regulation of cell cycle	2.96e-20
regulation of developmental process	3.21e-20
primary metabolic process	4.86e-20
regulation of gene expression	5.86e-20
signal transduction	9.31e-20
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.18e-19
positive regulation of apoptosis	2.29e-19
positive regulation of transcription	2.43e-19
positive regulation of programmed cell death	2.73e-19
positive regulation of cell death	3.68e-19
response to organic substance	8.38e-19
metabolic process	9.94e-19
intracellular signaling pathway	1.85e-18
regulation of gene-specific transcription	2.64e-18
response to UV	3.74e-18
double-strand break repair	4.86e-18
regulation of transcription from RNA polymerase II promoter	5.12e-18
signal transmission	6.81e-18
interphase of mitotic cell cycle	9.15e-18
regulation of cell differentiation	9.18e-18
regulation of multicellular organismal process	9.45e-18
positive regulation of RNA metabolic process	1.08e-17
transcription activator activity	1.21e-17
cellular nitrogen compound metabolic process	1.22e-17
DNA binding	1.33e-17
response to ionizing radiation	2.05e-17
interphase	2.08e-17
cell death	3.80e-17
death	4.47e-17
regulation of biological quality	4.59e-17
regulation of DNA replication	6.27e-17
regulation of transcription	6.49e-17
positive regulation of cell proliferation	6.92e-17
positive regulation of transcription, DNA-dependent	9.02e-17
negative regulation of cell proliferation	1.07e-16
regulation of cell cycle process	1.23e-16
nitrogen compound metabolic process	1.28e-16
regulation of catalytic activity	2.12e-16
post-translational protein modification	3.88e-16

GO term	p-value
regulation of RNA metabolic process	8.52e-16
DNA replication	1.64e-15
response to drug	1.72e-15
positive regulation of cellular protein metabolic process	1.79e-15
regulation of transcription, DNA-dependent	1.83e-15
response to steroid hormone stimulus	1.85e-15
transcription factor binding	1.94e-15
programmed cell death	2.10e-15
binding	2.20e-15
regulation of phosphorylation	2.43e-15
anti-apoptosis	3.06e-15
positive regulation of transcription from RNA polymerase II promoter	4.14e-15
response to hormone stimulus	4.93e-15
positive regulation of protein metabolic process	5.15e-15
regulation of binding	5.71e-15
regulation of cyclin-dependent protein kinase activity	6.75e-15
regulation of phosphorus metabolic process	7.10e-15
negative regulation of DNA metabolic process	7.52e-15
negative regulation of DNA replication	1.02e-14
regulation of cellular protein metabolic process	1.85e-14
mitotic cell cycle checkpoint	2.95e-14
macromolecule modification	3.28e-14
response to gamma radiation	3.37e-14
positive regulation of developmental process	4.20e-14
regulation of protein metabolic process	4.29e-14
chromatin modification	5.01e-14
response to light stimulus	6.06e-14
response to endogenous stimulus	6.31e-14
signaling pathway	8.82e-14
negative regulation of gene expression	1.05e-13
developmental process	1.18e-13
regulation of growth	1.22e-13
signaling	1.39e-13
transcription regulator activity	1.42e-13
structure-specific DNA binding	1.76e-13
telomere maintenance	1.78e-13
negative regulation of transcription from RNA polymerase II promoter	1.84e-13
protein modification process	1.91e-13
organ development	2.21e-13
mitotic cell cycle	2.43e-13

GO term	p-value
telomere organization	2.56e-13
positive regulation of molecular function	3.62e-13
negative regulation of transcription	7.61e-13
negative regulation of transcription, DNA-dependent	7.97e-13
induction of apoptosis	8.06e-13
cell cycle arrest	8.29e-13
induction of programmed cell death	8.50e-13

Bibliography

URL http://en.wikipedia.org/wiki/Tautomer.

URL http://www.answers.com/topic/markush-structure.

URL http://www.nature.com/nrd/journal/v1/n3/glossary/nrd745_glossary.html.

P. Agarwal and D. B. Searls. Literature mining in support of drug discovery. *Brief Bioinform*, 9(6):479–492, Nov 2008.

C. B. Ahlers, M. Fiszman, D. Demner-Fushman, F.-M. Lang, and T. C. Rindflesch. Extracting semantic predications from medline citations for pharmacogenomics. *Pac Symp Biocomput*, pages 209–220, 2007.

S. Anstein, G. Kremer, and U. Reyle. Identifying and classifying terms in the life sciences: The case of chemical terminology. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, editors, *Proc. of the Fifth Language Resources and Evaluation Conference*, pages 1095–1098, Genoa. Italy, 2006.

M. D. Anway, A. S. Cupp, M. Uzumcu, and M. K. Skinner. Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science*, 308(5727):1466–1469, Jun 2005.

S. Argamon-Engelson and I. Dagan. Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, 11:335–360, 1999.

A. R. Aronson. Comparison of lvg and metamap functionality. http://skr.nlm.nih.gov/papers/references/LVG-MetaMap.comparison.pdf, 1994.

A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp*, pages 17–21, 2001.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25 (1):25–29, May 2000.

S. M. B. Yildiz. Motivating ontology-driven information extraction. In *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), Bangalore, India*, 2007.

E. Badia, J. Oliva, P. Balaguer, and V. Cavaillès. Tamoxifen resistance and epigenetic modifications in breast cancer cell lines. *Current medicinal chemistry*, 14(28):3035–3045, 2007. ISSN 0929-8673.

A. Bairoch and R. Apweiler. The swiss-prot protein sequence data bank and its supplement trembl. *Nucleic Acids Res*, 25(1):31–36, Jan 1997.

J. Bajorath. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J Chem Inf Comput Sci*, 41(2):233–245, 2001.

J. Bajorath. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.*, 1: 882 – 894, 2002.

D. L. Banville. Mining chemical structural information from the drug literature. *Drug Discov Today*, 11(1-2):35–42, Jan 2006.

A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, Feb 2004.

J. M. Barnard and G. M. Downs. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.*, 37:141 – – 142, 1997.

W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–i48, Jul 2007.

T. Bekhuis. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomed Digit Libr*, 3:2, 2006.

J. C. n. . J. P. Ben Wellner. Adaptive string similarity metrics for biomedical reference resolution. In *In Proceedings of BioLink 2005*, pages 9–16, 2005.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

S. L. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard. An operational definition of epigenetics. *Genes Dev*, 23(7):781–783, Apr 2009.

C. Biemann. Ontology learning from text: A survey of methods. *LDV-Forum*, 20(2):75–93, 2005.

C. Blaschke, E. A. Leon, M. Krallinger, and A. Valencia. Evaluation of biocreative assessment of task 2. *BMC Bioinformatics*, 6 Suppl 1:S16, 2005.

D. Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics*, pages 977–981, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

D. Bradley. Dealing with a data dilemma. *Nature Reviews Drug Discovery*, 7:632–633, 2008.

J. Brecher. Name=struct: A practical approach to the sorry state of real-life chemical nomenclature. *J. Chem. Inf. Comput. Sci.*, 39(6):943–950, 1999.

E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 152–155, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

E. Brill. Some advances in transformation-based part of speech tagging. In *In Proceedings of the twelfth national conference on artificial intelligence*, pages 722–727, 1994.

C. Brooksbank, G. Cameron, and J. Thornton. The european bioinformatics institute's data resources: towards systems biology. *Nucleic Acids Res*, 33(Database issue):D46–D53, Jan 2005.

R. D. Brown and Y. C. Martin. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.*, 36:572 – 584, 1996.

D. V. Bulavin, Y. Higashimoto, I. J. Popoff, W. A. Gaarde, V. Basrur, O. Potapova, E. Appella, and A. J. Fornace. Initiation of a g2/m checkpoint after ultraviolet radiation requires p38 kinase. *Nature*, 411(6833):102–107, May 2001. doi: 10.1038/35075107.

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2:121–167, June 1998. ISSN 1384-5810.

E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler. The gene ontology annotation (goa) project: implementation of go in swiss-prot, trembl, and interpro. *Genome Res*, 13(4):662–672, Apr 2003.

M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, Jul 2008.

C. Cardie. Empirical methods in information extraction. *AI magazine*, 18:65–79, 1997.

J. T. Chang and R. B. Altman. Extracting and characterizing gene-drug relationships from the literature. *Pharmacogenetics*, 14(9):577–586, September 2004. ISSN 0960-314X.

J.-L. Chen, D. McLeod, and D. O'Leary. Schema evolution for object-based accounting database systems. In *ISOOMS '94: Proceedings of the International Symposium on Object-Oriented Methodologies and Systems*, pages 40–52, London, UK, 1994. Springer-Verlag. ISBN 3-540-58451-X.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In A. Joshi and M. Palmer, editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, 1996. Morgan Kaufmann Publishers.

H.-W. Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. *Pac Symp Biocomput*, pages 4–15, 2006.

P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab. Learning taxonomic relations from heterogeneous sources of evidence. In P. Buitelaar, B. Magnini, and P. Cimiano, editors, *Ontology Learning from Text: Methods, Applications, Evaluation*, Frontiers in AI. IOS Verlag, 2005.

B. L. Claus and D. J. Underwood. Discovery informatics: its evolving role in drug discovery. *Drug Discov Today*, 7(18):957–966, Sep 2002.

B. K. Cohen and L. Hunter. Natural language processing and systems biology. In *Artificial Intelligence Methods and Tools for Systems Biology*, pages 147–173. Springer, 2004.

B. K. Cohen, L. Fox, P. V. Ogren, and L. Hunter. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 38–45, Detroit, June 2005a. Association for Computational Linguistics.

K. B. Cohen, L. Fox, P. V. Ogren, and L. Hunter. Empirical data on corpus design and usage in biomedical natural language processing. *AMIA Annu Symp Proc*, pages 156–160, 2005b.

W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string distance metrics for name-matching tasks, 2003.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Mach. Learn.*, 15(2):201–221, 1994. ISSN 0885-6125.

P. Corbett and P. Murray-Rust. High-throughput identification of chemistry in life science texts. pages 107–118. 2006.

P. Corbett, C. Batchelor, and S. Teufel. Annotation of chemical named entities. In *BioNLP 2007: Biological, translational, and clinical language processing*, pages 57–64, Prague, June 2007.

E. Cornacchia, J. Golbus, J. Maybaum, J. Strahler, S. Hanash, and B. Richardson. Hydralazine and procainamide inhibit t cell dna methylation and induce autoreactivity. *J Immunol*, 140 (7):2197–2200, Apr 1988.

T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. *Proc Int Conf Intell Syst Mol Biol*, pages 77–86, 1999.

M. P. Crosland. *Historical Studies in the Language of Chemistry*. Courier Dover Publications, 2004.

G. L. Cuthbert, S. Daujat, A. W. Snowden, H. Erdjument-Bromage, T. Hagiwara, M. Yamada, R. Schneider, P. D. Gregory, P. Tempst, A. J. Bannister, and T. Kouzarides. Histone deimination antagonizes arginine methylation. *Cell*, 118(5):545–553, Sep 2004.

I. Dagan and K. Church. Termight: identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing*, pages 34–40, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer. Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J. Chem. Inf. Comput. Sci.*, 32(3):244–255, 1992.

W. Dang, K. K. Steffen, R. Perry, J. A. Dorsey, F. B. Johnson, A. Shilatifard, M. Kaeberlein, B. K. Kennedy, and S. L. Berger. Histone h4 lysine 16 acetylation regulates cellular lifespan. *Nature*, 459:802–807, 2009.

K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, 36(Database issue):D344–D350, Jan 2008.

G. Divita, A. C. Browne, and R. Loane. dtagger: a pos tagger. *AMIA Annu Symp Proc*, pages 200–203, 2006. ISSN 1559-4076.

A. Doms and M. Schroeder. Gopubmed: exploring pubmed with the gene ontology. *Nucleic Acids Res*, 33(Web Server issue):W783–W786, Jul 2005.

M. Dunkel, S. Günther, J. Ahmed, B. Wittig, and R. Preissner. Superpred drug classification and target prediction. *Nucleic Acids Res*, 36(Web Server issue):W55–W59, Jul 2008.

J. W. Edmunds, L. C. Mahadevan, and A. L. Clayton. Dynamic histone h3 methylation during gene induction: Hypb/setd2 mediates all h3k36 trimethylation. *EMBO J*, 27(2): 406–420, Jan 2008.

S. Egorov, A. Yuryev, and N. Daraselia. A simple and practical dictionary-based approach for identification of proteins in medline abstracts. *J Am Med Inform Assoc*, 11(3):174–178, 2004. ISSN 1067-5027.

W. S. el Deiry, T. Tokino, V. E. Velculescu, D. B. Levy, R. Parsons, J. M. Trent, D. Lin, W. E. Mercer, K. W. Kinzler, and B. Vogelstein. Waf1, a potential mediator of p53 tumor suppression. *Cell*, 75(4):817–825, Nov 1993.

P. E. B. J. Elena M. Zamora. Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 2. semantic phase. *J. Chem. Inf. Comput. Sci.*, 24(3):181–188, 1984.

G. A. Eller. Improving the quality of published chemical names with nomenclature software. *Molecules*, 11(11):915–928, 2006. ISSN 1420-3049.

S. P. Engelson and I. Dagan. Minimizing manual annotation cost in supervised training from corpora. In *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 319–326. Springer, 1996.

M. Esteller. Cancer epigenomics: Dna methylomes and histone-modification maps. *Nat Rev Genet*, 8(4):286–298, Apr 2007. doi: 10.1038/nrg2005.

Y.-C. Fang, H.-C. Huang, and H.-F. Juan. Meinfotext: associated gene methylation and cancer information from text mining. *BMC Bioinformatics*, 9:22, 2008.

A. P. Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447 (7143):433–440, May 2007.

R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95*, pages 112–117. AAAI Press, 1995.

A. H. Fielding. *Cluster and Classification Techniques for the Biosciences*. Cambridge University Press, New York, NY, USA, 2007. ISBN 0521618002.

M. Fiszman, T. C. Rindflesch, and H. Kilicoglu. Interpreting hypernymic propositions in an online medical encyclopedia. *AMIA Annu Symp Proc*, page 840, 2003a.

M. Fiszman, T. C. Rindflesch, and H. Kilicoglu. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. *AMIA Annu Symp Proc*, pages 239–243, 2003b.

J. Fluck, H.-T. Mevissen, H. Dach, M. Oster, and M. Hofmann-Apitius. Prominer: Recognition of human gene and protein names using regularly updated dictionaries. In *2nd BioCreAtIvE Challenge Workshop 2006, Critical Assessment of Information Extraction in Molecular Biology, Madrid Spain*, 2006.

M. F. Fraga, E. Ballestar, M. F. Paz, S. Ropero, F. Setien, M. L. Ballestar, D. H.-S. ner, J. C. Cigudosa, M. Urioste, J. Benitez, M. Boix-Chornet, A. Sanchez-Aguilera, C. Ling, E. Carlsson, P. Poulsen, A. Vaag, Z. Stephan, T. D. Spector, Y.-Z. Wu, C. Plass, and M. Esteller. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A*, 102(30):10604–10609, Jul 2005.

K. T. Frantzi. Incorporating context information for the extraction of terms. In *ACL-35: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 501–503, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

K. T. Frantzi, S. Ananiadou, and J.-i. Tsujii. The c-value/nc-value method of automatic recognition for multi-word terms. In *ECDL '98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 585–604, London, UK, 1998. Springer-Verlag. ISBN 3-540-65101-2.

K. Fundel. *Text Mining and Gene Expression Analysis. Towards Combined Interpretation of High Throughput Data*. PhD thesis, Ludwig-Maximilians-Universität München, 2007.

K. Fundel and R. Zimmer. Gene and protein nomenclature in public databases. *BMC Bioinformatics*, 7:372, 2006.

K. Fundel, D. Güttler, R. Zimmer, and J. Apostolakis. A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics*, 6 Suppl 1:S15, 2005.

T. Gattermayer. Scaiview : Annotation and visualization system for knowledge discovery. Master's thesis, Bonn-Aachen International Center for Information Technology, 2007.

K. Gendler, T. Paulsen, and C. Napoli. Chromdb: the chromatin database. *Nucleic Acids Res*, 36(Database issue):D298–D302, Jan 2008.

U. R. Gerhard Kremer, Stefanie Anstein. Analysing and classifying names of chemical compounds with chemorph. In J. F. Sophia Ananiadou, editor, *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine*, April 2006.

F. Giunchiglia, U. Kharkevich, and I. Zaihrayeu. Concept search: Semantics enabled syntactic search. In S. Bloehdorn, M. Grobelnik, P. Mika, and D. T. Tran, editors, *Proceedings of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008) , June 2, 2008, Tenerife, Spain*, volume 334 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

A. Golbraikh and A. Tropsha. Beware of q2! *J Mol Graph Model*, 20(4):269–276, Jan 2002.

S. Goto, T. Nishioka, and M. Kanehisa. Ligand: chemical database for enzyme reactions. *Bioinformatics*, 14(7):591–599, 1998.

M. Göttlicher, S. Minucci, P. Zhu, O. H. Krämer, A. Schimpf, S. Giavara, J. P. Sleeman, F. L. Coco, C. Nervi, P. G. Pelicci, and T. Heinzel. Valproic acid defines a novel class of hdac inhibitors inducing differentiation of transformed cells. *EMBO J*, 20(24):6969–6978, Dec 2001.

J. Gräff and I. M. Mansuy. Epigenetic codes in cognition and behaviour. *Behav Brain Res*, 192 (1):70–87, Sep 2008.

D. Grau and G. Serbedzija. Innovative strategies for drug repurposing. Drug Discovery and Development, May 2005.

R. Grishman and B. Sundheim. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, pages 466–471, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

N. Guarino. Formal ontology and information systems. In *Proceedings of the First International Conference on Formal Ontologies in Information Systems (FOIS-98), June 6-8, 1998, Trento, Italy*, 1998.

H. Gurulingappa. Concept-based semi-automatic classification of drugs. Master's thesis, Bonn-Aachen Internation Center for Information Technology, 2008.

H. Gurulingappa, C. Kolářik, M. Hofmann-Apitius, and J. Fluck. Concept-based semi-automatic classification of drugs. *Journal of Chemical Information and Modeling*, 49:1986–1992, 2009.

R. Hale. Text mining: getting more value from literature resources. *Drug Discov Today*, 10(6): 377–379, Mar 2005.

S. Hanasoge and M. Ljungman. H2ax phosphorylation after uv irradiation is triggered by dna repair intermediates and is mediated by the atr kinase. *Carcinogenesis*, 28(11): 2298–2304, Nov 2007. doi: 10.1093/carcin/bgm157.

D. Hanisch. *New Analysis Methods for Gene Expression Data via Construction and Incorporation of Biological Networks*. PhD thesis, Ludwig-Maximilians-Universität München, 2005.

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

D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 (Suppl 1)(S14), 2005.

M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

M. A. Hearst. Untangling text data mining. In *University of Maryland*, pages 3–10, 1999.

Z. Herceg. Epigenetics and cancer: towards an evaluation of the impact of environmental and dietary factors. *Mutagenesis*, 22(2):91–103, Mar 2007.

K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. Hendriksen, B. J. Schijvenaars, E. M. van Mulligen, J. Kleinjans, and J. A. Kors. A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, Sep 2009.

R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nat Genet*, 36(7): 664, Jul 2004.

M. Hofmann-Apitius, J. Fluck, L. Furlong, O. Fornes, C. Kolárik, S. Hanser, M. Boeker, S. Schulz, F. Sanz, R. Klinger, T. Mevissen, T. Gattermayer, B. Oliva, and C. M. Friedrich. Knowledge environments representing molecular entities for the virtual physiological human. *Philos Transact A Math Phys Eng Sci*, 366(1878):3091–3110, Sep 2008.

A. Hotho, A. Nürnberger, and G. Paaß. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20(1):19–62, MAY 2005. ISSN 0175-1336.

L. Hunter, Z. Lu, J. Firby, W. A. Baumgartner, H. L. Johnson, P. V. Ogren, and K. B. Cohen. Opendmap: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, 9:78, 2008.

A. R. Isles and L. S. Wilkinson. Epigenetics: what is it and why is it important to mental disease? *Br Med Bull*, 85:35–45, 2008.

C. Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 341–348, Morristown, NJ, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3.

T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, Aug 2001.

J. Jiang and C. Zhai. An empirical study of tokenization strategies for biomedical information retrieval. *Inf. Retr.*, 10(4-5):341–363, 2007. ISSN 1386-4564.

V. Jijkoun, M. A. Khalid, M. Marx, and M. de Rijke. Named entity normalization in user generated content. In *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 23–30, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-196-5.

G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. pages 338–345, 1995.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, Jun 2007.

S. O. Jónsdóttir, F. S. Jørgensen, and S. Brunak. Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics*, 21(10):2145–2160, 2005. ISSN 1367-4803.

J. Justeson and S. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, pages 9–27, 1995.

M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–D280, Jan 2004.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–D484, Jan 2008.

J. W. Katrin Tomanek and U. Hahn. Sentence and token splitting based on conditional random fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 49–57, 2007.

G. B. Kauffman. The making of modern chemistry. *Science*, 338:699 – 700, 1989.

R. J. Kavlock, G. Ankley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, E. Hubal, R. Judson, J. Rabinowitz, A. Richard, R. W. Setzer, I. Shah, D. Villeneuve, and E. Weber. Computational toxicology–a state of the science mini review. *Toxicol Sci*, 103(1):14–27, May 2008.

D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002. ISSN 1077-2626.

N. Kemp and M. Lynch. Extraction of information from the text of chemical patents. 1. identification of specific chemical names. *J. Chem. Inf. Comput. Sci.*, 38:544–551, 1998.

A. Kimura, K. Matsubara, and M. Horikoshi. A decade of histone acetylation: marking eukaryotic chromosomes with specific codes. *J Biochem*, 138(6):647–662, Dec 2005.

R. Klinger and K. Tomanek. Classical Probabilistic Models and Conditional Random Fields. Technical Report TR07-2-013, Department of Computer Science, Dortmund University of Technology, December 2007. ISSN 1864-4503.

R. Klinger, L. I. Furlong, C. M. Friedrich, H. T. Mevissen, J. Fluck, F. Sanz, and M. Hofmann-Apitius. Identifying gene specific variants in biomedical text. *Journal of Bioinformatics and Computational Biology*, 5(6):1277–1296, December 2007 2007.

R. Klinger, C. Kolářik, J. Fluck, M. Hofmann-Apitius, and C. M. Friedrich. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics*, 24(13):i268–i276, 2008. Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB).

C. M. Koch, R. M. Andrews, P. Flicek, S. C. Dillon, U. Karaoz, G. K. Clelland, S. Wilcox, D. M. Beare, J. C. Fowler, P. Couttet, K. D. James, G. C. Lefebvre, A. W. Bruce, O. M. Dovey, P. D. Ellis, P. Dhami, C. F. Langford, Z. Weng, E. Birney, N. P. Carter, D. Vetrie, and I. Dunham. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.*, 17(6):691–707, June 2007.

C. Kolářik, M. Hofmann-Apitius, M. Zimmermann, and J. Fluck. Identification of new drug classification terms in textual resources. *Bioinformatics*, 23(13):i264–i272, 2007.

C. Kolářik, R. Klinger, C. M. Friedrich, M. Hofmann-Apitius, and J. Fluck. Chemical Names: Terminological Resources and Corpora Annotation. In *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*, pages 51–58, Marrakech, Morocco, May 2008.

C. Kolářik and R. Klinger, 2008. URL `http://www.scai.fraunhofer.de/tms08.html`.

S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31(3):249–268, 2007.

N. Kourmouli, P. Jeppesen, S. Mahadevhaiah, P. Burgoyne, R. Wu, D. M. Gilbert, S. Bongiorni, G. Prantera, L. Fanti, S. Pimpinelli, W. Shi, R. Fundele, and P. B. Singh. Heterochromatin and tri-methylated lysine 20 of histone h4 in animals. *J Cell Sci*, 117(Pt 12):2491–2501, May 2004. doi: 10.1242/jcs.01238.

M. Krallinger and A. Valencia. Text-mining and information-retrieval services for molecular biology. *Genome Biol*, 6(7):224, 2005.

M. Krallinger, A. Valencia, and L. Hirschman. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol*, 9 Suppl 2:S8, 2008.

M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *J Biomed Inform*, 37(6):512–526, Dec 2004.

G. Kroemer and J. Pouyssegur. Tumor cell metabolism: cancer's achilles' heel. *Cancer Cell*, 13(6):472–482, Jun 2008. doi: 10.1016/j.ccr.2008.05.005.

H. Kubinyi. Validation and predictivity of qsar models. In E. A. Sener and I. Yalcin, editors, *QSAR & Molecular Modelling in Rational Design of Bioactive Molecules (Proceedings of the 15th European Symposium on QSAR & Molecular Modelling, Istanbul, Turkey, 2004)*, pages 30–33. CADDD Society, 2004.

T. Kudoh and Y. Matsumoto. Use of support vector learning for chunk identification. In *In Proceedings of CoNLL-2000 and LLL-2000*, pages 142–144, 2000.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

V. Lafarga, A. Cuadrado, and A. R. Nebreda. p18(hamlet) mediates different p53-dependent responses to dna-damage inducing agents. *Cell Cycle*, 6(19):2319–2322, Oct 2007.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 2001.

J. A. Latham and S. Y. R. Dent. Cross-regulation of histone modifications. *Nat Struct Mol Biol*, 14(11):1017–1024, Nov 2007.

G. Leech. Corpus annotation schemes. *Literary and linguistic computing*, 8(4):275–281, 1993.

J.-H. Lim, H. Jang, J. Lim, and S.-J. Park. Normalization of gene/protein names in biological literatures using vector-space model. *Conf Proc IEEE Eng Med Biol Soc*, 2007:390–393, 2007.

W.-S. Lo, K. W. Henry, M. F. Schwartz, and S. L. Berger. Histone modification patterns during gene activation. *Methods Enzymol*, 377:130–153, 2004. doi: 10.1016/S0076-6879(03)77007-4.

A. H. Lund and M. van Lohuizen. Epigenetics and cancer. *Genes Dev*, 18(19):2315–2335, Oct 2004.

R. Mack, S. Mukherjea, A. Soffer, N. Uramoto, E. Brown, A. Coden, J. Cooper, A. Inokuchi, B. Iyer, Y. Mass, H. Matsuzawa, and L. V. Subramaniam. Text analytics for life science using the unstructured information management architecture. *IBM Syst. J.*, 43(3):490–515, 2004. ISSN 0018-8670.

A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems and Their Applications*, 16(2):72–79, 2001.

C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0262133601.

Y. Mansour. Pessimistic decision tree pruning based on tree size. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 195–201. Morgan Kaufmann, 1997.

M. Marcus, G. Kim, M. A. Marcinkiewicz, R. Macintyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The penn treebank: Annotating predicate argument structure. In *In ARPA Human Language Technology Workshop*, pages 114–119, 1994a.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1994b.

R. Margueron, P. Trojer, and D. Reinberg. The key to development: interpreting the histone code? *Curr Opin Genet Dev*, 15(2):163–176, Apr 2005.

L. Marino-Ramírez, B. Hsu, A. D. Baxevanis, and D. Landsman. The histone database: a comprehensive resource for histones and histone fold-containing proteins. *Proteins*, 62(4): 838–842, Mar 2006.

R. Marmorstein. Protein modules that manipulate histone tails for chromatin regulation. *Nat Rev Mol Cell Biol*, 2(6):422–432, Jun 2001.

Y. C. Martin, J. L. Kofron, and L. M. Traphagen. Do structurally similar molecules have similar biological activity. *J Med Chem 2002*, 45:4350–4358, 2002.

A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 188–191, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

A. McCallum, D. Freitag, and F. C. N. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591–598, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2.

A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

R. McDonald and F. Pereira. Identifying Gene and Protein Mentions in Text Using Conditional Random Fields. *BMC Bioinformatics*, 6 (Suppl 1)(S6), May 2005.

R. T. McDonald, R. S. Winters, M. Mandel, Y. Jin, P. S. White, and F. Pereira. An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics*, 20: 3249–3251, 2004.

M. J. McGregor and P. V. Pallai. Clustering of large databases of compounds: using mdl 'keys' as structural descriptors. *J. Chem. Inf. Comput. Sci.*, 37:443 – 448, 1997.

A. McNaught. he iupac international chemical identifier: Inchl – a new standard for molecular informatics. *Chemistry International*, 28(6):12 – 14, 2006.

J. McNaught and B. W. J. *Information Extraction.*, chapter 7, pages 213–245. Artech House, Inc., 2005.

G. A. Miller. Nouns in wordnet: A lexical inheritance system. *Int. J. Lexicography*, 3(4): 245–264, January 1990.

R. Mizoguchi and M. Ikeda. Towards ontology engineering. Technical report, The Institute of Scientific and Industrial Research, Osaka University, 1996. Technical Report AI-TR-96-1.

A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H. hui Liu, R. Torres, M. Krauthammer, W. W. Lau, H. Liu, C.-N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman. Overview of biocreative ii gene normalization. *Genome Biol*, 9 Suppl 2:S3, 2008.

T. Morton and J. LaCivita. WordFreak: an Open Tool for Linguistic Annotation. In *HLT/NAACL 2003: demonstrations*, pages 17–18, 2003.

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

P. Murray-Rust, H. S. Rzepa, and B. J. Whitaker. The world-wide web as a chemical information tool. *Chem. Soc. Rev.*, 26:1–10, 1997.

P. Murray-Rust, J. B. O. Mitchell, and H. S. Rzepa. Chemistry in bioinformatics. *BMC Bioinformatics*, 6:141, 2005. doi: 10.1186/1471-2105-6-141.

J. C. Nacher and J.-M. Schwartz. A global view of drug-therapy interactions. *BMC Pharmacol*, 8:5, 2008.

Nadeau, David, Sekine, and Satoshi. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007. ISSN 0378-4169. Gives an overview about NER methods from 1991 until 2006.

M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker. A biological named entity recognizer. In *Proc. of the Pacific Symposium on Biocomputing*, pages 427–438, 2003.

G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001. ISSN 0360-0300.

C. Nedellec and A. Nazarenko. Ontologies and information extraction. *CoRR*, abs/cs/0609137, 2006.

C. J. Nelson, H. Santos-Rosa, and T. Kouzarides. Proline isomerization of histone h3 regulates lysine methylation and gene expression. *Cell*, 126(5):905–916, Sep 2006.

G. Nenadié, S. Ananiadou, and J. McNaught. Enhancing automatic term recognition through recognition of variation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 604, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

K. P. Nightingale, L. P. O'Neill, and B. M. Turner. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr Opin Genet Dev*, 16(2):125–136, Apr 2006.

T. R. O'Connor and J. J. Wyrick. Chromatindb: a database of genome-wide histone modification patterns for saccharomyces cerevisiae. *Bioinformatics*, 23(14):1828–1830, Jul 2007.

M. Ongenaert, L. V. Neste, T. D. Meyer, G. Menschaert, S. Bekaert, and W. V. Criekinge. Pubmeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res*, 36(Database issue):D842–D846, Jan 2008.

D. D. Palmer and M. A. Hearst. Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics*, 23:241–267, 1997.

J. Park and J. Kim. *Named Entity Recognition.*, chapter 9, pages 213–245. Artech House, Inc., 2005.

D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, and L. E. Weinberger. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J Med Chem*, 39(16):3049–3059, Aug 1996.

F. Pereira. Distributional clustering of english words. In *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.

C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser. Alibaba: Pubmed as a graph. *Bioinformatics*, 22(19):2444–2445, Oct 2006.

L. J. G. Post, M. Roos, M. S. Marshall, R. van Driel, and T. M. Breit. A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data. *Bioinformatics*, 23(22):3080–3087, Nov 2007.

P. O. Prada, J. R. Pauli, E. R. Ropelle, H. G. Zecchin, J. B. C. Carvalheira, L. A. Velloso, and M. J. A. Saad. Selective modulation of the cap/cbl pathway in the adipose tissue of high fat diet treated rats. *FEBS Lett*, 580(20):4889–4894, Sep 2006.

J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, 1986. ISSN 0885-6125.

L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–15, January 1986.

L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In D. Yarovsky and K. Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey, 1995. Association for Computational Linguistics.

D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr. Ebimed–text crunching to gather facts for proteins from medline. *Bioinformatics*, 23(2): e237–e244, Jan 2007.

D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through web services: calling whatizit. *Bioinformatics*, 24(2):296–298, Jan 2008.

A. S. Reddy, S. P. Pati, P. P. Kumar, H. N. Pradeep, and G. N. Sastry. Virtual screening in drug discovery – a computational perspective. *Current protein & peptide science*, 8(4):329–351, August 2007. ISSN 1389-2037.

B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–2309, Dec 2000.

U. Reyle. Understanding chemical terminology. *Terminology*, 12(1):111–136, 2006.

E. J. Richards and S. C. R. Elgin. Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell*, 108(4):489–500, Feb 2002.

F. Rikken and R. Vos. How adverse drug reactions can play a role in innovative drug research. *Pharm World Sci*, 17(6):195–200, Nov 1995.

E. Riloff. Information extraction as a stepping stone toward story understanding. pages 435–460, 1999.

T. C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform*, 36(6):462–477, Dec 2003.

P. M. Roberts and W. S. Hayes. Information needs and the role of text mining in drug development. *Pac Symp Biocomput*, pages 592–603, 2008.

K. D. Robertson. Dna methylation and chromatin - unraveling the tangled web. *Oncogene*, 21(35):5361–5379, Aug 2002. doi: 10.1038/sj.onc.1205609.

R. Rodriguez-Esteban. Biomedical text mining and its applications. *PLoS Comput Biol*, 5(12): e1000597, Dec 2009. doi: 10.1371/journal.pcbi.1000597.

E. P. Rogakou, D. R. Pilch, A. H. Orr, V. S. Ivanova, and W. M. Bonner. Dna double-stranded breaks induce histone h2ax phosphorylation on serine 139. *J Biol Chem*, 273(10):5858–5868, Mar 1998.

M. Ronning. Coding and classification in drug statistics. from national to global application. *Norwegian Journal of Epidemiology*, 11(1):37–40, 2001.

C. B. Santos-Rebouças and M. M. G. Pimentel. Implication of abnormal epigenetic patterns for human diseases. *Eur J Hum Genet*, 15(1):10–17, Jan 2007.

S. Sarawagi. Information extraction. *Found. Trends databases*, 1(3):261–377, 2008. ISSN 1931-7883.

B. Sarg, W. Helliger, H. Talasz, B. Förg, and H. H. Lindner. Histone h1 phosphorylation occurs site-specifically during interphase and mitosis: identification of a novel phosphorylation site on histone h1. *J Biol Chem*, 281(10):6573–6580, Mar 2006. doi: 10.1074/jbc.M508957200.

I. Sarkar, M. Cantor, R. Gelman, F. Hartel, and Y. Lussier. Linking biomedical language information and knowledge resources in the 21st century: Go and umls. In *Pacific Symposium on Biocomputing 2003*, pages 439–450, 2003.

Y. Sasaki, Y. Tsuruoka, J. McNaught, and S. Ananiadou. How to make the most of ne dictionaries in statistical ner. *BMC Bioinformatics*, 9 Suppl 11:S5, 2008.

A. Savary and C. Jacquemin. *Text- and Speech-Triggered Information Access*, chapter Reducing Information Variation in Text, pages 145–181. Lecture Notes in Artificial Intelligence. Springer Berlin / Heidelberg, 2003.

B. Schölkopf and A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.

M. Schuemie, R. Jelier, and J. Kors. Peregrine: Lightweight gene name normalization by dictionary lookup. In *Proceedings of the Biocreative 2 workshop 2007 April 23-25, Madrid*, pages 131–140, 2007.

C. Selassie. *Burger's Medicinal Chemistry and Drug Discovery*, chapter History of Quantitative Structure-Activity Relationships, pages 1–48. John Wiley & Sons, Inc, 2008.

A. Sen and M. Srivastava. *Regression analysis, theory, methods and applications*, volume 13. Springer Verlag GmbH, 1992.

B. Settles. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, Jul 2005. doi: 10.1093/bioinformatics/bti475.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X.

F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 134–141, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.

C. Siefkes and P. Siniakov. An overview and classification of adaptive approaches to information extraction. In *Journal on Data Semantics, LNCS 3730*, volume IV, pages 172–212. Springer, 2005.

A. Silvescu, J. Reinoso-castillo, and V. Honavar. Ontology-driven information extraction and knowledge acquisition from heterogeneous, distributed, autonomous biological data sources. In *In Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources*, 2001.

M. K. Simon. *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers, Scientists and Mathematicians*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387346570.

A. Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42, 2001.

H. Skolnik. Milestones in chemical information science. *J. Chem. Inf. Comput. Sci.*, 16:187–193, 1976.

A. Skrbo, I. Zulić, S. Hadzić, and I. D. Gaon. Anatomic-therapeutic-chemical classification of drugs. *Med Arh*, 53(3 Suppl 3):57–60, 1999.

A. Skusa, A. Rüegg, and J. Köhler. Extraction of biological interaction networks from scientific literature. *Brief Bioinform*, 6(3):263–276, Sep 2005.

N. R. Smalheiser and D. R. Swanson. Using arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed*, 57(3): 149–153, Nov 1998.

A. Smallwood, P.-O. Estève, S. Pradhan, and M. Carey. Functional cooperation between hp1 and dnmt1 mediates gene silencing. *Genes Dev*, 21(10):1169–1178, May 2007.

B. Smith. *Blackwell Guide to the Philosophy of Computing and Information*, chapter Ontology, pages 155–166. Oxford: Blackwell, 2003.

L. Smith, T. Rindflesch, and W. J. Wilbur. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321, Sep 2004.

L. Smith, L. K. Tanabe, R. J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. A. Struble, R. J. Povinelli, A. Vlachos, W. A. Baumgartner, L. Hunter, B. Carpenter, R. T.-H. Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. M. na López, J. Mata, and W. J. Wilbur. Overview of biocreative ii gene mention recognition. *Genome Biol*, 9 Suppl 2:S2, 2008.

Y. H. Soung, J. W. Lee, S. Y. Kim, W. S. Park, S. W. Nam, J. Y. Lee, N. J. Yoo, and S. H. Lee. Somatic mutations of casp3 gene in human cancers. *Hum Genet*, 115(2):112–115, Jul 2004. doi: 10.1007/s00439-004-1129-3.

I. Spasić, S. Ananiadou, J. McNaught, and A. Kumar. Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform*, 6(3):239–251, Sep 2005.

S. Staab, C. Blaschke, C. Nedellec, J. Park, B. Schatz, A. Valencia, L. Bernardi, E. Ratsch, R. Kania, J. Saric, and I. Rojas. Mining information for functional genomics. *IEEE Intelligent Systems*, 17(3):66–80, MAY 2002. Trends & Controversies.

R. Stevens, C. A. Goble, and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, 1(4):398–414, Nov 2000.

B. Sun, Q. Tan, P. Mitra, and C. L. Giles. Extraction and search of chemical formulae in text documents on the web. In *Proc. of the International World Wide Web Conference*, pages 251–260, May 2007.

C. Sutton and A. Mccallum. *Introduction to Conditional Random Fields for Relational Learning*, chapter Chap. 4, pages 93–127. MIT Press, 2006.

P. Sykes. *Reaktionsmechanismen der organischen Chemie: eine Einfuehrung*. Verl. Chemie, 1988.

M. Szyf. The dynamic epigenome and its implications in toxicology. *Toxicol Sci*, 100(1):7–23, Nov 2007.

T. Takenobu, O. Hironori, and T. Hozumi. Effectiveness of complex index terms in information retrieval, 2000.

J. Tamames and A. Valencia. The success (or not) of HUGO nomenclature. *Genome Biol*, 7(5): 402, 2006.

S. D. Taverna, H. Li, A. J. Ruthenburg, C. D. Allis, and D. J. Patel. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol*, 14(11):1025–1040, Nov 2007.

J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005. ISBN 0769523234.

A. H. Ting, K. M. McGarvey, and S. B. Baylin. The cancer epigenome–components and functional correlates. *Genes Dev*, 20(23):3215–3231, Dec 2006.

K. Tomanek and U. Hahn. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL-IJCNLP '09, pages 1039–1047, Morristown, NJ, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-46-6.

K. Tomanek, J. Wermter, and U. Hahn. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 486–495, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

J. Townsend, A. Copestake, P. Murray-Rust, S. Teufel, and C. Waudby. Language technology for processing chemistry publications. In *Proceedings of the fourth UK e-Science All Hands Meeting, 2005.*, 2005.

J. A. Townsend, S. E. Adams, C. A. Waudby, V. K. de Souza, J. M. Goodman, and P. Murray-Rust. Chemical documents: machine understanding and automated information extraction. *Org Biomol Chem*, 2(22):3294–3300, Nov 2004.

V. P. Tryndyak, L. Muskhelishvili, O. Kovalchuk, R. Rodriguez-Juarez, B. Montgomery, M. I. Churchwell, S. A. Ross, F. A. Beland, and I. P. Pogribny. Effect of long-term tamoxifen exposure on genotoxic and epigenetic changes in rat liver: implications for tamoxifen-induced hepatocarcinogenesis. *Carcinogenesis*, 27(8):1713–1720, Aug 2006.

Y. Tsuruoka and J. ichi Tsujii. Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform*, 37(6):461–470, Dec 2004.

Y. Tsuruoka and J. Tsujii. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*, pages 41–48, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. Mcnaught, S. Ananiadou, and J. i. Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics*, pages 382–392. Springer Berlin / Heidelberg, 2005.

B. M. Turner. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol*, 12(2):110–112, Feb 2005.

C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979. ISBN 0-408-70929-4.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

L. M. Villeneuve, M. A. Reddy, L. L. Lanting, M. Wang, L. Meng, and R. Natarajan. Epigenetic histone h3 lysine 9 methylation in metabolic memory and inflammatory phenotype of vascular smooth muscle cells in diabetes. *Proc Natl Acad Sci U S A*, 105(26):9047–9052, Jul 2008.

A. Vlachos. A stopping criterion for active learning. *Comput. Speech Lang.*, 22(3):295–312, 2008. ISSN 0885-2308.

H. Wallach. Efficient training of conditional random fields. Master's thesis, University of Edinburgh, 2002.

X. Wang and M. Matthews. Comparing usability of matching techniques for normalising biomedical named entities. *Pac Symp Biocomput*, pages 628–639, 2008.

X. Wang, T. Furukawa, T. Nitanda, M. Okamoto, Y. Sugimoto, S.-I. Akiyama, and M. Baba. Breast cancer resistance protein (bcrp/abcg2) induces cellular resistance to hiv-1 nucleoside reverse transcriptase inhibitors. *Mol Pharmacol*, 63(1):65–72, Jan 2003.

D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988. ISSN 0095-2338.

J. Wermter, J. Fluck, J. Stroetgen, S. Geißler, and U. Hahn. Recognizing noun phrases in biomedical text: an evaluation of lab prototypes and commercial chunkers. In U. Hahn and A. Valencia, editors, *Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine*, volume 7. SMBM, April 2005.

W. J. Wilbur, G. F. Hazard, G. Divita, J. G. Mork, A. R. Aronson, and A. C. Browne. Analysis of biomedical text for chemical names: a comparison of three methods. *Proc AMIA Symp*, pages 176–180, 1999.

P. Willett, J. M. Barnard, and G. M. Downs. Chemical similarity searching. *JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES*, 38(6):983–996, 1998.

A. G. Wilson. Epigenetic regulation of gene expression in the inflammatory response and relevance to common diseases. *J Periodontol*, 79(8 Suppl):1514–1519, Aug 2008.

C. O. Wilson. *Wilson and Gisuold's Textbook of Organic Medicinal and Pharmaceutical Chemistry*. Lippincott, Philadelphia, 8th ed edition, 1982. ISBN 0397520921.

D. A. Winkler. The role of quantitative structure-activity relationships (qsar) in biomolecular discovery. *Briefings in Bioinformatics*, 3(1):73–86, 2002.

D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, 34(Database issue):D668–D672, Jan 2006.

D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. Macinnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser. Hmdb: the human metabolome database. *Nucleic Acids Res*, 35(Database issue):D521–D526, Jan 2007.

Q. Zheng and X.-J. Wang. Goeast: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic Acids Res*, 36(Web Server issue):W358–W363, Jul 2008. doi: 10.1093/nar/gkn276.

A. Zippo, R. Serafini, M. Rocchigiani, S. Pennacchini, A. Krepelova, and S. Oliviero. Histone crosstalk between h3s10ph and h4k16ac generates a histone code that mediates transcription elongation. *Cell*, 138(6):1122–1136, Sep 2009. doi: 10.1016/j.cell.2009.07.031.

D. Zweigenbaum, Pierre & Demner-Fushman. Advanced literature-mining tools, 2008. http://www.campusvirtuel.smbh.univ-paris13.fr/claroline/document/goto/index.php?url=

P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. B. Cohen. Frontiers of biomedical text mining: current progress. *Brief Bioinform*, 8(5):358–375, Sep 2007.

Publication List

Harsha Gurulingappa, **Corinna Kolářik**, Martin Hofmann-Apitius and Juliane Fluck. Concept-Based Semi-Automatic Classification of Drugs. *Journal of Chemical Information and Modeling*, 49:1986–1992, 2009.

Corinna Kolářik, Roman Klinger and Martin Hofmann-Apitius. Identification of Histone Modifications in Biomedical Text for Supporting Epigenomic Research. *BMC Bioinformatics*, 10(Suppl 1):S28+, 2009 and *Proceedings of the Asia Pacific Bioinformatics Conference*, Beijing, China, 2009.

Corinna Kolářik and Martin Hofmann-Apitius. Linking Chemical and Biological Information with Natural Language Processing. Book chapter in: *Chemical Information Mining*, CRC Press, Chapter 7, 123–150, 2008.

M. Hofmann-Apitius, J. Fluck, L. I. Furlong, O. Fornes, **C. Kolářik**, S. Hanser, M. Boeker, S. Schulz, F. Sanz, R. Klinger, H. T. Mevissen, T. Gattermayer, B. Oliva, C. M. Friedrich. Knowledge Environments Representing Molecular Entities for the Virtual Physiological Human. *Philosophical Transactions of the Royal Society A*, 366(1878):3091–3110, 2008.

Roman Klinger, **Corinna Kolářik**, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics*, Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB), 24:268–276, 2008.

Corinna Kolářik, Roman Klinger, Christoph M. Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. Chemical Names: Terminological Resources and Corpora Annotation. In *Workshop on Building and Evaluating Resources for Biomedical Text Mining*, (6th edition of the Language Resources and Evaluation Conference), Marrakech, Morocco, 51–58, 2008.

Corinna Kolářik, Martin Hofmann-Apitius, Marc Zimmermann and Juliane Fluck. Identification of New Drug Classification Terms in Textual Resources. *Bioinformatics*,Proceedings of the ISMB/ECCB, 23:264–272, 2007.

Marc Zimmermann, Juliane Fluck, Le Thuy Bui Thi, **Corinna Kolářik**, Kai Kumpf, and Martin Hofmann. Information Extraction in the Life Sciences: Perspectives for Medicinal Chemistry, Pharmacology and Toxicology. *Current Topics in Medicinal Chemistry*, 5(8):785–96, 2005.