Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications

Article *in* Molecular Informatics · June 2011

DOI: 10.1002/minf.201100005

CITATIONS
63

READS
254

4 authors:

Miguel Vazquez
Barcelona Supercomputing Center
80 PUBLICATIONS **2,927** CITATIONS
SEE PROFILE

Martin Krallinger
Barcelona Supercomputing Center
127 PUBLICATIONS **3,339** CITATIONS
SEE PROFILE

Florian Leitner
Universidad Politécnica de Madrid
30 PUBLICATIONS **1,546** CITATIONS
SEE PROFILE

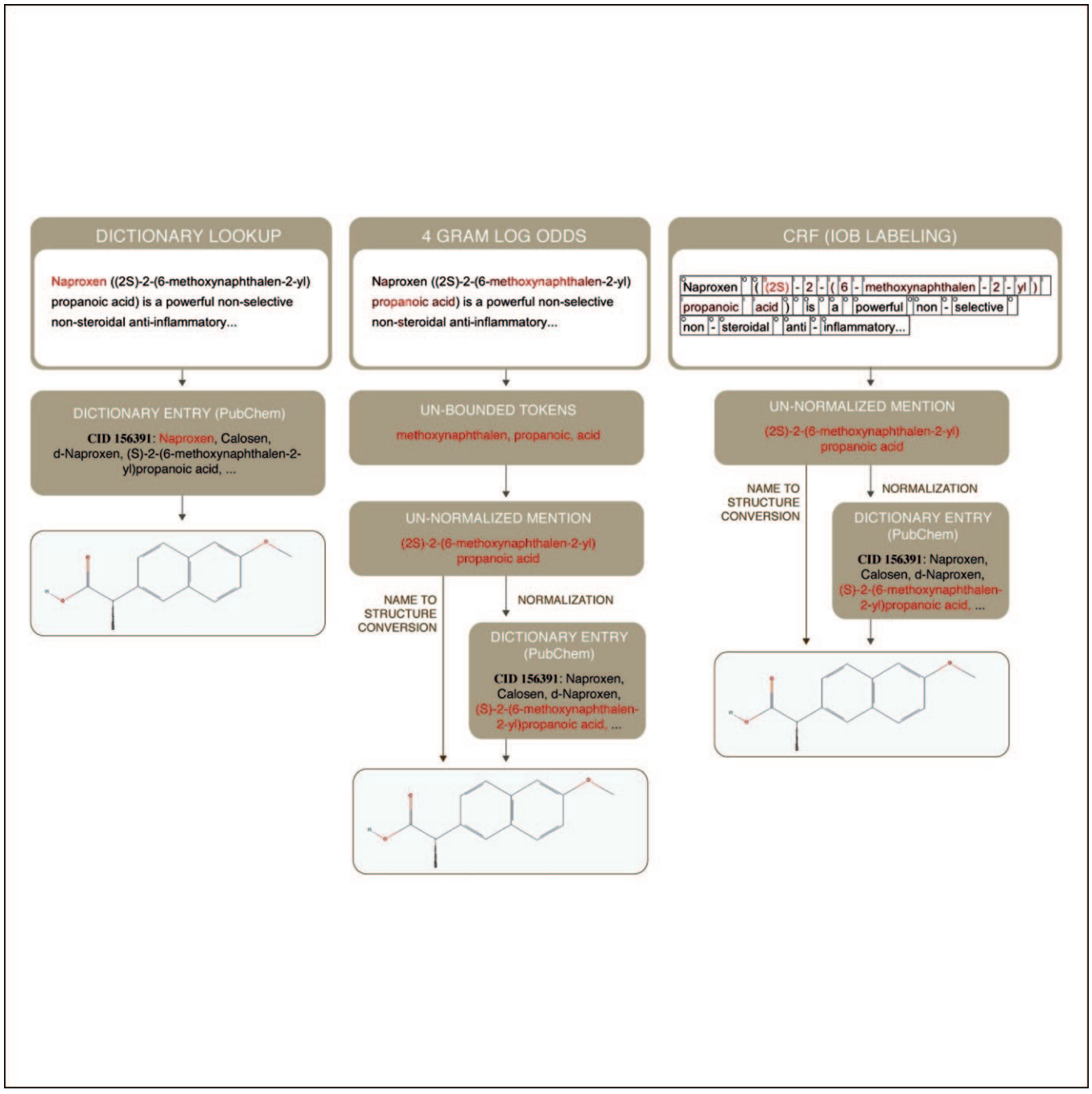
Alfonso Valencia
Centro Nacional de Investigaciones Oncológicas
667 PUBLICATIONS **52,860** CITATIONS
SEE PROFILE

Review

DOI: 10.1002/minf.201100005

Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications

Miguel Vazquez,[a] Martin Krallinger,[a] Florian Leitner,[a] and Alfonso Valencia*[a]



molecular informatics

Abstract: Providing prior knowledge about biological properties of chemicals, such as kinetic values, protein targets, or toxic effects, can facilitate many aspects of drug development. Chemical information is rapidly accumulating in all sorts of free text documents like patents, industry reports, or scientific articles, which has motivated the development of specifically tailored text mining applications. Despite the potential gains, chemical text mining still faces significant challenges. One of the most salient is the recognition of chemical entities mentioned in text. To help practitioners contribute to this area, a good portion of this review is devoted to this issue, and presents the basic concepts and principles underlying the main strategies. The technical details are introduced and accompanied by relevant bibliographic references. Other tasks discussed are retrieving relevant articles, identifying relationships between chemicals and other entities, or determining the chemical structures of chemicals mentioned in text. This review also introduces a number of published applications that can be used to build pipelines in topics like drug side effects, toxicity, and protein-disease-compound network analysis. We conclude the review with an outlook on how we expect the field to evolve, discussing its possibilities and its current limitations.

Keywords: Text mining · Named entity recognition · Information extraction · Chemical compounds · Drugs

1 Introduction

Modern drug research integrates information from very heterogeneous research areas such as chemistry, toxicology, molecular biology or physiology. The explosion of data in these diverse fields poses a great opportunity and a formidable challenge. Bioinformatics and Cheminformatics are the disciplines in charge of producing the necessary computational resources and databases on which the entire research pipeline is based on. Databases hold structured information about entities such as drugs, proteins/genes and diseases, together with the description of their relationships, and the experimental information that supports them; their organized structure makes them easy to query and facilitates the computational analysis of the results. However, compiling information in databases, a process usually requiring manual extraction directly from scientific publications, reports, or patent files, is significantly limited due to the cost and effort it requires; and it is obvious that these difficulties will only increase as the amount of scientific publications continues to rapidly grow. It is difficult to imagine that in a reasonably near future all relevant information will be contained within a database. Even if it would be possible to have all the basic facts recorded in databases, scientists will still constantly need to retrieve the essential details that complement those basic facts; for example, the specific experimental conditions in which a given protein-drug interaction was detected. Therefore, Information Retrieval will remain a requirement for expert users, hopefully assisted by text mining tools. Furthermore, beyond the exposition of facts, publications and reports contain additional information about the interpretation of the experiments and elaborate reasoning about the combined analysis of results; the results of the intellectual work are even more difficult to capture and reproduce in structured repositories, and remain a challenge for information extraction systems. This complete scenario in the interface between textual sources and databases is the central motivation behind the development of text mining applications. The focus of this article is on applications to identify drugs and chemical compounds in text and characterize their biomedical context.[1]

In the domain of drug development, the extraction of information on drugs and compounds is commonly done in an effort to establish relationships with other entities, in particular proteins/genes and disease/symptoms. Relationships extracted from text mining could constitute valuable information that directly aids the interpretation of the biological context and the understanding of the underlying molecular mechanisms. Furthermore, in the very inter-related molecular systems, relationships are often essential even for the interpretation of the properties of the individual components. Extracting relationships between entities from text is typically approached as a two-step process: (1) detecting the mentions to the entities in text, and (2) inferring their relationships based on their distribution in the documents.

The initial step towards identifying the mentions to entities in text, known in text-mining as *Named Entity Recognition* (NER), is fundamental, and it largely determines the success of the applications that use it. NER is complicated by the many ways in which these entities may be mentioned and by the complexity of human language. A number of lexical resources have been developed to capture the different ways in which an entity can appear mentioned in text. Such resources are already effectively used by text mining applications in the molecular biology domain for the detection of protein/gene names. Examples of applications using NER are the discovery of cancer-associated genes,[2,3] the extraction of physical protein interactions[4] or the semi-automatic construction of specialized data repositories.[5]

[a] *M. Vazquez, M. Krallinger, F. Leitner, A. Valencia*
Centro Nacional de Investigaciones Oncológicas
Biología Computacional y Estructural, Madrid, Spain
**e-mail: valencia@cnio.es*

It is fair to say that part of the success of the lexical resources in the molecular biology domain is related with their ready availability, Equivalent resources for chemical compounds, however, tend to be more restricted, and in many cases released under a commercial licence, which partly explains why the recognition of drugs and chemical compounds has received less attention than the recognition of entities such as genes and proteins.[6] This trend has began to change with the recent appearance of free and openly accessible chemical databases such as PubChem[7] and Drugbank,[8] which, together with the pressing need for non-commercial literature mining systems operating in the chemical/drug field, has promoted numerous efforts over the last few years. Given its critical importance, this review will attempt to cover in detail the efforts towards developing NER in the chemical domain. The discussion will also extend to other entities relevant to drugs and chemical compounds, such as genes and proteins, adverse events, diseases or particular numeric values.

The step following name recognition in a text-mining application is usually the identification of relationships between entities; a common task in what is known as *Information Extraction*. In this article we first describe the main technical approaches, including those based on machine learning and *Natural Language Processing* (NLP). Further on we critically discuss some relevant current implementa-

Miguel Vazquez holds a post-doctoral position at the Structural Biology and Biocomputing group of the Spanish National Cancer Research Center (CNIO); after completing a PhD on bioinformatics at a computer science department of the Universidad Complutense de Madrid, where he worked as a teaching assistant for five years. He has done stays at Minnesota, Barcelona, and Austin. In Austin he first learned about text mining about 10 years ago, and it has been one of his main research interests ever since.

Florian Leitner is completing his PhD at the Spanish National Cancer Research (CNIO) in Spain with Alfonso Valencia, specializing on text mining. He completed his Masters in Molecular Biology, working on Computational and Structural Biology, with Frank Eisenhaber at the IMP in Vienna, Austria. He previously worked for Rebecca Wade at the EML in Heidelberg, Germany, and for Markus Jaritz at the Novartis Research Campus in Vienna.

Martin Krallinger is currently working at the Structural Biology and Biocomputing group of the Spanish National Cancer Research Center (CNIO). He is completing his Ph.D. under Alfonso Valencia specializing in biomedical text mining. He has served as co-organizer of the Second BioCreAtIvE Challenge Workshop and also for the Workshop on Text Mining for the BioCuration Workflow at the 3rd International Biocuration conference.

Alfonso Valencia is Director of the Program in Structural Biology and Biocomputation at CNIO (the Spanish National Cancer Research Center). He received formal training in population genetics and biophysics from the Universidad Complutense de Madrid and was awarded his PhD in 1988 from the Universidad Autónoma de Madrid. From 1989–1994 he was a Postdoctoral Fellow at the laboratory of C. Sander at the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. In 1994 Alfonso Valencia set up the Protein Design Group at the Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC) in Madrid where he was appointed as Research Professor in 2005. He is a member of European Molecular Biology Organization, founder, former Vice President and current member of the board of directors of the International Society for Computational Biology. He is a founding organizer of the annual European Computational Biology Conferences and serves on the Scientific Advisory Board of the Swiss Institute for Bioinformatics and Biozentrum, U. Basel, as well as the Steering Committee of the European Science Foundation Programme on Functional Genomics (2006–2011) and of the ELIXIR EU infrastructure in Bioinformatics initiative. Prof. Valencia is co-organizer of the BioCreative challenges, co-executive editor of Bioinformatics, serves on the Editorial Board of EMBO Reports, and is Director of the Spanish National Bioinformatics Institute (INB), a platform of the Spanish Health Institute (ISCIII).

tions, their possibilities and limitations. Finally, in the outlook section, we present some of the more promising trends and future developments in the application of text mining to the chemical and pharmaceutical domain.

2 Named Entity Recognition (NER): a Close Look Into Drugs and Chemical Compound Mentions

2.1 The Many Names of a Drug

Chemicals may be referenced in documents in several ways: using *systematic nomenclatures*, *common* or *trade names*, database identifiers, *InChI strings* or even *raster images*. Different types of names or references have different *word morphologies* (characteristic features in how these names are formed). Chemical name morphology directly affects how we approach the task of finding chemicals mentioned in text, and even enables determination of their structure. Thus, an overview on the chemical nomenclatures will help our discussion.

2.1.1 Systematic Nomenclatures

Systematic nomenclatures follow precise rules on how these names are formed (*grammars*) that describe the compound in terms of its structure. These grammars, when strictly followed, should allow unambiguous determination of the chemical structure from their systematic names. The International Union of Pure and Applied Chemistry (IUPAC) (http://www.iupac.org/)[9] has been in charge of maintaining the rules of chemical nomenclature since 1892, although there are other guidelines such as the Chemical Service (CAS) (http://www.cas.org/)[10] index names. Note that, even though names should un-ambiguously determine the structures, the IUPAC guidelines allow for some variability on how names are formed, so several different correct names may describe the same structure.[11] This fact coupled with the variability derived from orthographic and spelling variations, complicates the practical use of the grammars.[12] Fortunately, the chemical name grammar shows strong regularities. The basic building blocks of chemical names, known as the basic name segments (or terminal symbols in grammar jargon), are very distinct from normal English. Take for example "benzo" or "methyl": any token in text containing such substrings is in great likelihood referring to a chemical entity.

2.1.2 Non-Systematic Nomenclatures

Systematic nomenclatures are cumbersome to read and write so common names or abbreviations are frequently used instead. For example, the term "*Aspirin*" if often favored over the IUPAC name "*2-acetyloxybenzoic acid*". These common or trivial names are catalogued and linked to their structure in resources such as PubChem. This is also the case of drug names. Some of the common names may still show regularities, e.g., drug names following the WHOINN dictionary of stems for non-proprietary names. But these naming conventions tend not to be as rigorous as the systematic nomenclatures, and thus do not allow reliable identification of the chemical structures from their grammars. It is however possible to match common names against dictionaries of names, whereas systematic names are hard to catalogue comprehensively into a dictionary due to their variability.

In some cases, names are constructed with a mixture of systematic and non-systematic portions, for example "*2-hydroxy-toluene*" and "*2-methyl-phenol*" are semi-systematic variants for "*1-hydroxy-2-methyl-benzene*". Even if semi-systematic names present some morphological regularity similar to fully systematic names, they are difficult to translate to the corresponding structures.

To get an estimate of the relative use of the different types of mentions, Kolárik et al.[13] examined a manually annotated collection or corpus of 100 MEDLINE abstracts. Out of the 1206 mentions found in these abstracts, 391 are IUPAC and IUPAC-like (systematic and semi-systematic) names, while 414 are common names, 161 are abbreviations, and the rest are parts of IUPAC names (49), chemical family names (99) and other formulas and 49 unambiguous representations (including *sum formulas*, *SMILES*, *InChI*). Note, however, that the distribution of names in abstracts may be very different from the distribution in the full text of the articles or other types of documents such as patents. In fact, another study, that in this case uses automated approach, compares the distribution of different names in abstracts, full text and patents, suggesting a much larger prevalence of IUPAC names in patents than in MEDLINE entries.[14]

2.2 Methods to Detect Chemical Compound Mentions

Currently, many of the academic text mining applications focus in the identification of common names, such as names of marketed drugs, using simple approaches like dictionary matching (see below). Less frequent is finding systematic names, which often require more involved approaches than simple dictionary matching. However, current research in chemical NER is now increasingly placing emphasis precisely on these systematic nomenclatures. The field is still at an early stage, and thus the number of free open-source solutions is rather limited.

We classify NER approaches in three categories: *dictionary based*, *morphology based*, and *context based*. Some of the methods presented have been used before for the identification of mentions to other types of entities; in particular genes and proteins, for which there is a significant body of work. The characteristics of chemical names, however, require particular adaptations to these methods, especially names following systematic nomenclatures, typically long multiword terms with large spelling variability and

subject to frequent misuse, truncation or misspecification. In fact, the length of the chemical name can vary greatly: from short acronyms to long multiword expressions spanning over several lines. These problems are only aggravated by the lack of representative publicly available and manually labeled compound mentions that could be used for training and evaluating systems.

The following section will describe the different approaches to chemical NER and discuss their most important issues.

2.2.1 Matching Text to Dictionary Entries

We use the term "dictionary matching" to refer to a family of techniques that find mentions in text by comparing it with a dictionary, or catalogue, of known names. The process of identifying the dictionary names in text is called matching or lookup. Implementing a dictionary matching technique consists in producing a good quality dictionary and implementing a *matching method*.

Devising a dictionary that is both good and comprehensive is perhaps the most critical aspect. Dictionaries can be manually generated, but in most cases they are generated automatically from resources such as public chemical databases or *thesauri* (lists of words grouped together by similarity of meaning, e.g., synonyms) like the Unified Medical Language System (UMLS).[15] Different resources may contain different entries; some are centered on drugs, others on general chemical compounds.[16] It is also a common practice to build dictionaries by merging several resources.

There are several issues related to working with dictionaries, not least of which is the practical problem of their size, specially when including systematic chemical names in addition to drug and trivial names. These dictionaries can reach several million entries, one or two orders of magnitude larger that the typical dictionary used to capture gene names. For instance, the jochem joint chemical dictionary (http://www.biosemantics.org/index.php?page=Jochem) contains close to 2 million synonyms, while the number for a typical gene name dictionary would be in the tens of thousands. Additionally, dictionaries may require extensive manual curation to maintain them and to remove redundant entries, e.g. names corresponding to common words or other problematic entries. In many cases heuristics and statistical properties are used to assist the curation process.[17,18] Dictionaries are effective in identifying names when these are correctly written; otherwise, it is necessary to enhance them by including the potential spelling and orthographic variations. Another option is to use *Regular Expressions* instead of exact string matching to capture the variability directly in the matching process. Another possibility to deal with the variability in chemical names is to use string comparison metrics such as edit distances like the *Levenshtein distance*,[19] a procedure that calculates the similarity of two character sequences by counting the number of characters that need to be changed in one sequence to transform it into the other. This metrics can determine that the mention '8-(*p*-sulphophenyl)theophylline' found in text and the dictionary entry '8-(*p*-sulfophenyl)theophylline' differ only in the character 'f' been written as 'ph', and thus, conclude that both may actually be equivalent. Implementing a matching strategy using the Levenshtein distance, for instance, may be computationally infeasible for a large dictionary, since each dictionary entry needs to be compared with all possible substrings of the text. It may, however, prove useful for exhaustive detection of just one or a few particular compounds. Furthermore, heuristics can be used to improve its performance; for example by first trying to find chemical-like tokens in the text, such as 'theophylline' in our previous example (see Section 2.2.2 for a discussion on different strategies on how to do this), and restrict the matching to only dictionary entries containing those tokens and to just the immediate context of the token in text.

Dictionaries are available for different types of entities. Pharmspresso[20] or PolySearch[21] compile dictionaries for many different entities which can be downloaded from their respective online sites. The entities they support vary, but both include drugs, genes/proteins and diseases. Another notable dictionary for drugs and chemical compounds is the Jochem dictionary,[16] which is automatically produced by merging several lexical resources such as UMLS, ChEBI, MeSH terms, PubChem or DrugBank. A corpus of 100 annotated MEDLINE abstracts was used to evaluate the performance of each of the different sources used in Jochem dictionary.[13] This evaluation also includes another freely available dictionary based on ChEBI and used in the open-source named chemical recognitions system OSCAR3.[22] This system implements several named chemical recognition strategies, including one using dictionary matching. In recent work the Jochem dictionary was compared to the manually curated ChemSpider dictionary.[23] The results of this study suggest that ChemSpider has fewer but higher quality entries resulting in better *precision* but lower *recall* (coverage) than Jochem.

2.2.2 Exploiting Morphological Regularities of Systematic Names

Systematic nomenclatures are in fact, as described earlier, grammars that use a finite set of terminal symbols which roughly correspond to the chemical name segments (e.g. "benzo" and "methyl"). Note that basic name segments may span over several grammar terminal symbols, such as the term 'tetraphenylene', where the terminal symbols may include 'tetra', 'phenyl', and 'ene'. Most of these name segments are very characteristic of chemical names. These sequences of characters have a higher probability of appearing inside a chemical name than inside background English words, thus improving our chances of detecting them. Several approaches use a dictionary of chemical name segments and try to find them in text. This dictionary can be

built automatically by segmenting chemical names with manually crafted heuristic rules,[24] or by visually introducing frequent substrings.[25] There are also official dictionaries of name segments such as the Registry File Basic Name Segment Dictionary (http://www.cas.org/ASSETS/EFF8CA9AA8324FC5A2F0428765287059/regbasicname.pdf), which can be used to divide a chemical name into basic name segments (see Table 1). The basic name segment dictionary was used in the segmentation approach presented by Wilbur et al.[26] In that case, to help capture generic and trade names, the segment dictionary is expanded to include a list of other biologically significant segments. Note that, even though these techniques make use of dictionaries, we do not strictly consider them dictionary-matching strategies, since they do not deal with the full chemical names.

Table 1. Examples of chemical name segments in the Registry File Basic Name Segment Dictionary. A segmenting algorithm matching the leftmost longest segment in a greedy way would make errors, so the registry includes the correct segmentation for those cases; like PENTOX, which could incorrectly be segmented as PENTO X.

Segment
HEPT
PHENYL
PENTOX→PENT OX

Text mining often requires segmenting text into smaller portions, such as words. This process is called *tokenization* or *segmentation*. Due to the nature of systematic names, it may be convenient to establish these segments not only relying on word boundaries, but also taking into account hyphens, parenthesis, transitions from a letter to a digit character, etc. (see Table 2). Tokens derived from text segmentation that resemble basic chemical name segments are easier to recognize since they correlate better with the underlying chemical entities and have statistical distributions that differentiate them of other terms.

There are other methods that exploit the statistical properties of chemical names and do not require explicitly building dictionaries of name segments. For instance, *n-grams*, which are sequences of n consecutive characters appearing in a token (e.g. "methyl" has three 4-grams: "meth", "ethy", "thyl"), may also have different frequencies in chemical names and in background English words. This was exploited in an approach introduced by Wilbur et al. In this case, a 4-gram *sliding window* is passed over the tokens derived from chemical names and over the tokens derived from the background text, taking counts to establish the conditional frequencies (see Table 3). A simple statistical model (i.e. Naïve Bayes model) is used to decide if the 4-grams derived from a new portion of text belong to a chemical name or to background text. Naïve Bayes assumes independence between features; since in this case they are extracted from overlapping n-grams this assumption is clearly violated. Vasserman[27] corrected it by taking non-overlapping n-grams, selecting them based on their discriminative ability. The appropriate number of characters for the n-grams is estimated to be 4 in both these approaches. Vasserman also introduces a technique that interpolates n-grams from different lengths, an approach that clearly out-performs the others in his evaluation corpus.

Table 2. Example tokenization of a chemical name by non-word characters and digit-letter transitions. Tokenization helps the named entity recognition systems by offering a smaller and more granular unit of analysis. In the case of chemical names, some of these tokens may resemble basic name segments and thus benefit from their statistical properties.

Tokens in "2-(hydroxymethyl)-6methoxyoxane-3,4,5-triol"
2
Hydroxymethyl
6
Methoxyoxane
3,4,5
Triol

Table 3. Example of 4-gram segmentation for "*2-(hydroxymethyl)-6methoxyoxane-3,4,5-triol*". The second column shows an estimate of the log odds between the 4-gram appearing in chemical names and in background English text (titles and abstracts from one thousand PubMed ids randomly extracted from GeneRIF entries of human genes). This supports the idea that statistics along can help pinpoint the location of chemical names. Only the top 5 4-grams are shown. Note the overlap between 4-grams in rows 1 and 2, and 3,4, and 5; when using these 4-grams as features in a Naïve Bayes classifier Vasserman took care of considering only a non-overlapping subset.

4-grams	Log odds
thox	7.55
hoxy	7.55
xyme	5.27
ethy	5.27
oxym	5.12

Wren proposed exploiting the conditional frequencies of character-to-character transitions, which are also different between chemical names and background text.[28] For example, character transitions like 'x' to 'y' appear in chemical tokens like 'etho**xy**butyl' but rarely in background English text. He compared two systems based on Markov models, (Markov models use transition probabilities between states, or, in our case, between characters[29]), one using chemical names and one using background text. At the end of the process each token is classified as chemical name or not, based on the Markov model under which it has the highest likelihood.

The approaches proposed in this section suffer from two important limitations. Since they work at a token, n-gram, or even character level, chemical names end up segmented into different pieces, and the correct mention boundaries

must be determined by some other procedure, for example using heuristic rules (see Kemp and Lynch for an example of such an heuristic[25]). The other limitation is that the context of the mention, for example the surrounding tokens, is not used to enhance the process. Both these limitations are addressed by the context aware systems described below.

As a final note, chemical identifiers such as SMILES and InChI strings are not being discussed here, since they present such strong morphological regularities that their identification is rather trivial.

2.2.3 Context Aware Systems

Approaches that make use of the context of a mention can be divided into machine learning models, which employ statistical properties, and *Natural Language Processing* (NLP) techniques or manual rules based on our knowledge of how natural language is formed.

Machine learning approaches work by examining example cases and inferring from them general patterns, and they thus require training data from which to learn from. The training data typically consist of portions of text in which the chemical entities have been labelled. A common labelling scheme uses the BIO (or IOB) format: Instead of labelling each token as being part of a mention of a chemical or not, the (B)egining, (I)nner, and (O)utside labels are used to tag the tokens as follows: (B) is used on the first token starting a mention, (I) any further token in the mention, and (O) for all other tokens not part of a mention. From these labels it is easy to determine the boundaries of chemical name mentions. Some examples of available training corpora are the 100 MEDLINE abstracts previously mentioned,[13] or the corpus developed as collaboration between the European Patent Office and the ChEBI team (http://chebi.cvs.sourceforge.net /viewvc/chebi/chapati/patentsGoldStandard/).[30]

A very popular technique introduced for NER in 2001,[31] and extensively used for these tasks ever since, is Conditional Random Fields (CRFs). CRFs, in their most common form, linear chain CRFs, are an attractive model because they consider the label for one token as conditional on the label of the previous token, a collection of features extracted from the token itself, and other features describing the surrounding tokens. To each token in the string a collection of features is assigned, constituting what the CRF learns from. These features may be very different in nature; for instance, morphological features like the presence of dashes and parenthesis, or whether the token is included in a dictionary (e.g. a dictionary of name segments). In fact, morphological features have been shown to be the most discriminative features in the identification of gene and protein names,[32] and are central to many of the machine learning approaches for systematic chemical names as well. CRFs have been widely used for the detection of gene and protein mentions, and, more recently, chemical mentions in IUPAC and IUPAC-like form[33] or in chemical formulas.[34] These CRFs very often follow the IOB scheme. Software packages that can be used to produce CRF systems include Mallet[35] and CRF++[36] Each type of entity to be detected by CRF requires a particular configuration, especially in the feature functions used to process the tokens into the algorithm inputs. This configuration may in fact differ greatly for gene and chemical mentions.

Using a corpus of manually annotated patent documents, dictionary approaches obtain better results in matching portions of chemical names, but are outperformed by CRFs when complete chemical mentions are considered.[30] An alternative model to CRF, called Maximum Entropy Markov Model (MEMM), is employed in OSCAR3 as a complementary NER strategy to dictionary matching.[37]

The obvious difficulty with these approaches is the need of *training data*, the quality of which significantly determines the success of the approach. Indeed, in most cases domain experts are required to produce the training datasets manually. This is a cumbersome and laborious process, even if *semi-automatic approaches* are used to assist the human experts.[38]

A different family of techniques is based on the use of linguistic analyses of the text, such as *syntactic analysis* (see the literature[39] for a comprehensive overview). Here, the regularities of the language constructs are used to derive rules (also called *frames*, see the literature[40] for an initial description of this type of systems in biology) that can complement those derived from statistical analysis with orthogonal information. An example of an entity recognition system that employs these techniques is the MetaMap program.[41] This application incorporates NLP techniques to identify terms from the UMLS thesaurus in text. Segura-Bedmar et al.[42] classified terms extracted by MetaMap as mentions to drugs or not by matching them against the WHOINN dictionary of nonproprietary names. A follow-up study by the same group[43] implemented a technique to track mentions to chemicals in *anaphoric expressions* (expressions referring to entities mentioned elsewhere in the text, -typically in a previous sentence- instead of explicitly naming them again).

We can find an example of a system using manual rules in Narayanaswamy et al.[44] for the recognition of biological entities, including names and part of names of chemical compounds. This system has a list of (functional) words that help to determine the location of chemical entities in text, for example words such as 'Drug' or 'Steroid' that can also help to classify the mentioned entities.

2.3 Determining Structures for Chemicals in Documents

One of the most useful representations of a chemical compound is via its *chemical structure*, which is also amenable to computational analysis. In many applications it is very important to be able to assign the structures to the chemicals detected in the text. There are three methods for extracting structures from documents: (1) finding mentions of

chemicals and identifying the chemical in a database containing structures, such as PubChem; or (2) if the chemical is represented as a systematic name, perform a *name to structure conversion*; and finally, (3) directly extract structures from raster images.

2.3.1 Linking Chemical Mentions to Database IDs

Very often in Information Extraction finding mentions to entities is just the first step; it is usually required that these mentions are then associated with the actual concepts they make reference to, thus connecting them to other available information for them. This task is known as *Mention Normalization*. For example, the mention of "*methylglucoside*" would be followed by the identification of the corresponding PubChem identifier *(CID:2108)*.

For common names, looking for the closest match in a dictionary may be a good strategy; dictionary entries may be associated with only one identifier, making this mapping step straightforward. In general, however, normalization of entities may be complicated due to differences in spelling or ambiguity. In the case of gene and protein mentions, normalization is further complicated by the ambiguous use of names referring to general classes (i.e. protein family names) instead of names of specific entities. In this context normalization is significantly harder than detection. Several strategies have been devised to approach these problems, many of them in the context of the BioCreative competitions.[4,45] These strategies include comparing the context where the mention was found to the text in database records describing each gene and protein. These strategies could also be applied for chemical mentions, although we are not aware of any publication so far that actually studies this, at least using non-trivial approaches. Matching of IUPAC names to dictionary entries is specially complicated, not only due to the aforementioned coverage problems of chemical dictionaries, but also because it hard to tell when similar mentions are spelling variations for the same name, a very frequent scenario,[46] or are in fact different chemicals.

2.3.2 Name to Structure Conversion

The use of systematic names should, in theory, allow for unambiguous identification of chemicals and their structure. This is true for SMILES and InChI representations; and also for IUPAC names, when the recommendations are rigorously followed. Unfortunately SMILES and InChI strings are not typically used in free text, nor are IUPAC conventions followed strictly in practice (see the literature[47] for an evaluation of the correctness of mentions used in published text compared to automatic naming software).

Cooke et al.[48] introduced the IUPAC nomenclature as a chemical 'language' and described it in terms of Chomsky grammar types as a *context-free grammar*. Further chapters of the same issue[49,50] attempt at describing the terminal symbols and the production rules necessary to define the IUPAC grammar. The IUPAC rules are followed in the literature[51] to identify chemical names in text, derive their structure, and classify them in classes.

For the Name = Struct program,[52] a commercial product, the 'grammar approach' is abandoned in favor of rules that follow IUPAC recommendations; allowances are however made for considerations on common usage in order to deal with ambiguity, misspellings, or even creative use of the nomenclature.

It is notable that the open-source package OPSIN, part of the OSCAR3 chemistry suite, also performs name to structure conversion. ACD/Labs offers tools under commercial licenses for name to structure conversion, which allow for batch processing.

2.3.3 Optical Chemical Structure Recognition

Structural information of compounds is often depicted by means of two-dimensional diagrams. They are commonly included in journal articles and patents as raster images.[53]

Optical Chemical Structure Recognition (OCSR) attempts to use pattern recognition to extract the structural information from these images so they can be used in cheminformatics tools. This process typically involves the following tasks:[54] (1) extracting the images from the flow of the documents and segment atoms and bonds into individual molecules; (2) determine which elements in the image are graphics (bond lines) and which are text (atom symbols); (3) identify the positions of elements, directions and length of bond lines, and determine the type of bond; and finally, (4) determine the connectivity between atoms and the number of bonds between them. Some systems such as ChemReader[54] crosscheck the resulting structure against physical rules such as bond lengths and angles to verify its correctness.

Several tools exist that can perform OCSR are reviewed in the literature,[54] namely Kekule,[55] CLiDE,[53] chemOCR[56] and the open-source package OSRA,[57] which are compared against their own development ChemReader.[54]

3 Beyond NER

While identifying mentions to chemicals and deriving their chemical structure may have interest in its own, it is often desirable to establish associations between the chemicals and other entities in the text. These other entities could be other chemicals, but they could also be entities of other nature such as proteins, adverse events or, in a more ample sense, even dosages and affinities.

These entities may be associated in many ways. Consider, for example, trying to extract mutations in CYP proteins associated with adverse events after administration of a chemical. In this case we would like to associate 4 entities: a mutation, a CYP protein, a drug or chemical compound and an adverse event. Additionally we might even want to

add to this list the dosage level and the percentage of incident, if available. The first step would be to find mentions to all this entities in the text. We have discussed amply the subject of finding mentions to chemicals; the Subsection 3.1 will introduce the problem of identifying mentions to other types of entities. After the mentions have been identified they must be grouped in sets of entities that are described to be associated in the text (chemicals and their target proteins, diseases and the main proteins involved, etc). This step, relationship extraction, has still seen little development in this area of application; Subsection 3.2 will introduce some of the basic concepts and present a few drug research related efforts.

3.1 Recognition of Mentions to Other Types of Entities

Although identification of chemical entities is the central topic of this review, other entities are no less important, and identifying them is essential for a complete interpretation of the textual information. Among these other entities are description of *adverse drug events*, *diseases* and *toxicology end-points*, or, in the biomedical domain, genes and proteins, which are required for the interpretation of the *mechanism of drug action* (e.g., *pharmacodynamics* and *pharmacokinetics*).

Adverse event, disease and toxicology end-point identification is typically approached using dictionary lookup strategies. These strategies use lexical resources such as the Medical Dictionary for Regulatory Activities (MedDRA),[58] The Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) or, more generally, UMLS.[15]

Gene and protein NER and normalization is a subject of study that has received ample attention and enjoyed significant advances thanks to community challenges such as the BioCreative[6] Most of the successful systems benefit from the wealth of training data available, and many use machine-learning approaches such as CRFs, which are especially popular in this domain. Some examples of CRF-based systems for the identification of gene and proteins that are very well known within the community are ABNER[59] and BANNER.[60]

Some entities that show strong regularities like *Single Nucleotide Polymorphisms* (SNP) ids (e.g. *rs10991377*), amino acid substitutions (e.g. *N445T*) or numeric values like percentages or probability values can be extracted using regular expressions. Applications like Pharmspresso[20] use a mixture approach of dictionary and regular expressions. In the case of entities such as genes and proteins regular expressions are derived from dictionary entries; for entities like SNPs, the application uses custom-built sets of regular expressions.

3.2 Extracting Relationships

The simplest and most common approach to Relationship Extraction is to use the *co-occurrence* of the detected entities. The rationale of this approach is that if two entities are mentioned in the same portion of text there is a good chance they share an association and are directly related. Co-occurrence may be defined at different levels, e.g. full document, text sections, paragraphs or sentences.

Document level co-occurrence is also known as *co-publication*. It has been used to find connections between chemical compounds and gene/proteins;[61,62] or drugs, proteins and diseases.[63,64] Co-publication is prone to finding many spurious associations, which grants high coverage but low precision. A common approach to deal with spurious associations is to use article counts as a surrogate measure of confidence on the derived associations. Article counts may have the drawback of emphasizing well-known relationships while obscuring others that might be more interesting due to their novelty.

On the other hand, *sentence level co-occurrence* achieves a higher precision by restricting the scope of the co-occurrence. Sentence co-occurrence has been used in several applications that link a large variety of entities such as drugs, proteins, and diseases, and also cell types, mutations, organisms, etc.[20,21,65,66] An added benefit of this approach is that the sentence where the co-occurrence occurs serves as supporting evidence that can be quickly assessed.

A simple enhancement of this scheme, sometimes called *tri-co-occurrence*, inspects additional words in the sentence that inform about the kind of associations, for example the word "phosphorylates" in the sentence "*kinase A* was found to phosphorylate *protein B*". This approach has been followed to automatically derive typed relationships between proteins.[67,68]

Despite its simplicity, the co-occurrence approach has been successfully applied to engross the list of annotations in resources such as STITCH and STRING,[69,70] or to determine metabolic and pharmacokinetic pathways.[71,72]

NLP techniques offer more sophisticated means of Relationship Extraction based on lexical and syntactic analyses. These techniques can be used, for instance, in *template-matching* or *rule-based strategies*, which have been successfully applied to mine relationships for entities such as proteins and genes,[73–75] pharmacogenomics entities,[76] or drugs and cytochrome proteins.[77,78] They have even been used for the extraction of numerical pharmacokinetic values in,[79] where syntactic analysis was limited to just determining the subjects of phrases which where then used in a template matching strategy.

4 Applications

One of the more interesting applications of text mining in this domain is the inference of the *mechanisms of action* based on the network of drugs and their phenotype effects. For example, the similarities of the side effects associated to a set of drugs have been used to propose relations between drugs and protein target, assuming that drugs that share a significant number of side effects will act

through the same targets. The feasibility of this approach was established in the literature,[80] based on side effects of commercial drugs. This work lead to a resource called SIDER that compiles side effects for approximately 900 drugs.[81]

An alternative strategy is the use of information on reported adverse events instead of the one on side effects. Information on Adverse Events can be obtained from the adverse event reporting systems (AERS), where health professionals report these adverse events directly, or perhaps can be extracted from patient generated content, organized in patient forums and blogs.[82] It has been pointed out that text mining over patient generated content suffers from the more colloquial use of language, but, in some settings, they may still provide information that would complement the one provided by health professionals.[83] Additionally, some studies have suggested the feasibility of using patient reported Adverse Events in pharmacovigilance.[84]

Another interesting area of application of text mining is the extraction of toxicological information. The ADME/Tox domain includes linking drugs to liver toxicities;[85] extracting pharmacokinetic values;[79] or associations between drugs and cytochrome proteins.[77,78]

On the biological domain the most demanded applications are those involved with linking drugs to diseases via their protein targets. For example, text mining derived drug-protein and protein-disease associations are part of the Pfizer discovery pipeline, as reported in a recent article.[68] Some online tools that can be used for this purpose are Pharmspresso.[20] and Polysearch.,[21] which can detect associations between various entities. Both tools use sentence level co-occurrence, and consider several other entities in addition to drugs, gene/proteins, and diseases, such as cell types, cellular locations, or mutations. Other related text mining systems are: CoPub,[86] IDMap,[61] CAESAR,[87] and ENDEAVOUR.[88] CoPub links drugs to diseases through proteins using co-publication, but uses the MeSH annotations of the articles/abstracts instead of the direct detection of the entities (NER). IDMap, is a downloadable tool with a graphical interface devised to investigate the properties of chemicals and their protein targets; it uses co-publication to determine chemical-target associations. CAESAR links genes to human traits, while ENDEAVOUR is a gene-prioritizing tool that also incorporates text-mining derived information.

Text mining derived associations have been used in generating network information valuable for -omics analysis of chemicals in a systems biology context: protein–protein interaction networks,[89] metabolic networks,[71,90–92] regulatory networks,[75] and pharmacokinetic pathways.[72] Complete pathways, regulatory networks or interaction modules generally cannot be extracted from a single article. Generating such networks automatically requires the concatenation of events from multiple documents that share the presence of common individual bioentities. This makes the procedure especially cumbersome, as it is currently virtually impossible for a text mining application to determine whether the contextual conditions of an event derived from one article are sufficiently similar to the ones described by another article. Moreover, some of the associations (e.g. gene regulation, enzyme-substrate-product relations) rely on the identification of the directionality of the relationship. Therefore most of the automatically generated networks assume some sort of generalization that does not always satisfactorily describe the constraints of real biological networks.

Additionally, a number of relevant efforts have been made to extract the parameters corresponding to the biochemical reactions. Rojas et al.[93] used the regularities in the terminology used to refer to the parameters in the literature, and Hakenberg et al. applied machine learning approaches to classify whether documents describe experimentally obtained parameters for kinetic models.[94] A more fine-grained extraction pipeline for kinetic information exploiting dictionary and rule based methods to recognize parameters (KM, Ki, kcat, pH, temperatures) and entities (enzyme names, EC numbers, ligands, organisms, localisations) important for reconstructing enzymatic pathways from PubMed abstract has recently been proposed by Heinen and colleagues.[95]

Text mining is also used to aid in the creation and annotation of the drug/chemical repositories and databases.[38] The comparative toxicogenomics database (CTD), which contains information about diseases, genes/proteins, and cell functions related to chemicals, has reported the use of text mining to improve curation efficiency.[96] These text mining tools range from *Information Retrieval* systems to systems that enhance reading experience. Information Retrieval systems are used for the identification of pharmacogenomics articles,[97] or to help patent searches by annotating them with MeSH terms.[98] Some even allow structure-based queries over patent records, like SureChem, an online site that uses chemical NER and name to structure conversion. Tools that directly enhance the documents in the browser are also known as user scripts,[99] like Reflect,[100] which offers a browser plug-in that can detect named chemicals and genes/proteins, and link them to other databases and bioinformatics systems. SENT[101] extracts topics discussed in a corpus of literature (for example articles describing a list of genes specified by the user) as a succinct list of representative terms. The associated online interface integrates a literature examination tool that ranks that corpus of articles by relevance to those particular topics.

Finally, in the context of personalized medicine one of the current limitations for adapting medical treatments to particular patient characteristics is the difficulty in associating genetic alterations detected in the patient to potential drug treatments. Part of this information may be in medical records and medical trials (patient information, drug treatments, genotyping or genomic information), which are particularly difficult to access and process. One of the applica-

Table 4. Selected applications.

Name	Comments
OSCAR3 and Reflect[19,95]	Finds mentions to drugs and compounds
OPSIN (OSCAR3) and name = struct	Name to structure conversion
SureChem (commercial)	Structure based searching for automatically annotated patent documents
QueryChem[100]	Find compounds similar to a query structure and combines their names with user defined terms on a Google query
FACTA+, Polysearch, Pharmspresso, CoPub, STITCH and Litminer[63,18,17,62,101,102]	Co-occurrence based discovery for drugs, compounds, and other entities
ABNER, BANNER and MetaMap[38,57,58]	Identification of gene mentions (ABNER,BANNER), diseases, adverse events and other UMLS terms (MetaMap).

tions of text mining in this area is associated to the curation the PharmaGKB database,[102] where genetic variations are linked to drug responses.[103,104]

5 Other Related Text Mining Approaches

Even if we have focused on the basic process of NER and Information Extraction, there are a number of other text mining applications that are relevant in the domain of biological chemistry and drug development. In particular, those related with Information Retrieval and document categorization.[105]

Information Retrieval systems are designed to rank large collections of documents, typically scientific articles, according to some criteria defined by the user. The most prominent of such systems is the interface to the search for similar abstracts in PubMed, based on similarities between distributions of words.[106] These systems function as part of the PubMed facilities. PubMed receives around 60 million user queries each month,[107] and allows queries specifying fields such as 'substance name' or 'pharmacological action'. These queries can specify terms appearing in different fields of the article entry (e.g. title, abstract, authors) or make used of the MeSH terms, a hierarchical vocabulary of terms used to index PubMed articles.

Many of the applications in this field are based on Machine learning algorithms trained with information derived from documents of interest to the user, Published examples are the classification of articles relevant to pharmacogenomics[97] or describing drug-drug interactions.[108]

In addition to scientific articles that are extensively processed by molecular biology text mining application, other sources of documents relevant to drug research are patents and competitive intelligence resources. Indeed, according to Roberts et al. *competitive intelligence* resources and patents second scientific articles in the interest for users with queries that are mainly on drugs, diseases, and genes.[109]

The evaluation of chemical Information Retrieval engines has been carried out in a community challenge called TREC-CHEM, an initiative centered in determining how Information Retrieval methods adapt to patents that contain chemical names and formulas.[110]

To close the circle, applying NER techniques can be used to enhance Information Retrieval and document categorization systems. For example, systems like Semedico[111] improve retrieval by automatically tagging documents with terms of different type, including compounds and drugs. Linking text to chemicals also allows for the formulation of complex structure-based queries with which to find compounds containing a particular substructure (such as pharmaco or toxicophores) or computing structure similarity metrics to a query molecule. Applications like SureChem offer this functionality to search for patents where chemical compounds are previously identified using NER.

6 Outlook

Chemical databases and the associated literature extraction tools have been mainly proprietary, which makes them fundamentally different from their counterparts in the molecular biology field, where the availability of a huge body of freely available databases and applications has contributed to a substantial improvement of the related research areas. Fortunately, the trend has begun to change after substantial efforts to share knowledge and provide open access databases of chemical interest. Some of the new openly accessible resources include databases such as DrugBank, ChEMBL, ChemProt and CTD, together with the development of important lexical resources like PubChem and Jochem, and chemical ontologies like ChEBI (Chemical entities of biological interest) and the Chemical Information Ontology (CHEMINF).

Specially promising will be the three-way integration of chemical information, biological- genomics- experimental data and pharmacological and medical information, essential for a better understanding of the relevant properties of chemical entities in comprehensive biological context. The combination of information organized in knowledge bases and Information Extraction techniques is essential to sustain this type of efforts. An initial example of the combination of information across domains and data sources could be the reconstruction of metabolic pathways.[112,113] In this case it is necessary to merge biological and chemical information; combining the knowledge on enzyme activity derived from the description of the function of proteins with the one on the specific chemical compounds subject to the reactions and transformations, with the obvious complications of connecting the substrates and products of consecutive reactions.

Furthermore, identifying numerical values, such as those describing the affinity of a compound-protein interaction, may be of as much importance as identifying the involved entities themselves. We have found little work on the extraction of numerical values from text for the pharmacological domain; one example is the work done by Wang et al.[79] It seems that mining numerical values would require some specific considerations. For example, numerical values are often found in tables; extracting text from tables can be particularly difficult, especially when processing documents in unstructured formats such as PDF. Numerical values may need also to be extracted along with their units, and these must be understood by the system so that '3 wk' can be recognize as a time span, and as been the same as '21 days'. These numerical values are often extracted manually to be included in databases such as ChEMBLdb (public: https://www.ebi.ac.uk/chembldb/index.php) or GVKBio (commercial: http://www.gvkbio.com/). Text mining methods to either extract them automatically, or to assist manual curation would be desirable.

Text-mining assisted construction of pathway knowledge bases based on manual validation of automatically extracted facts has been a promising strategy.[5,114] Despite progress made by these initial steps, there is still a long way ahead to integrate pathway reconstruction with information about reaction parameters and physiological information, in order to enable the modeling of the dynamic properties of corresponding systems. Text mining will undoubtedly be important not only for the extraction of the basic facts and entities, but for the description of the experimental conditions that will constrain the corresponding models.

Integration of text mining and high-throughput data analysis together with biochemical and signalling pathway databases can also be seen as a novel mechanism for target discovery and biomarker identification.[115] Current bottlenecks still reside in the lack of open access systems for efficiently handling chemical-relevant documents, together with the limited availability of Optical Character Recognition (OCR) systems optimized for chemical literature that can handle scanned PDF files, the common document format of patents.

A fundamental building block of chemical text mining systems consists in the detection of chemical compound mentions. A strategy to improve NER for particular tasks is to use community and shared task evaluation efforts, as previously explored for the recognition of genes/protein mentions in case of the BioCreative assessment[6]). Setting up text mining evaluation efforts like BioCreative for chemistry, a kind of ChemCreative could help to promote both the implementation of new, cutting edge technologies, as well as independently determining the state of the art in recognition of compound mentions using a common *Gold Standard* evaluation data set. One step in this direction was taken by CALBC (Collaborative Annotation of a Large Biomedical Corpus) with the aim of integrating automatically generated annotations of multiple systems for large text collections, covering also the annotation chemical compounds.[116]

Following the model of the molecular biology domain there is also considerable interest in assisting community annotation efforts through text-mining applications,[117] building semi-automatically manually validated databases and online community portals, an endeavour that requires a critical mass of active contributors. The PubChem database, for instance, allows deposition of new user-provided compounds after a series of validation steps to ensure correctness of the compound and avoid redundancy, providing links to the depositor in the data source field of the PubChem record, e.g. to the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) in case of PubChem compound SID:14709858.

A different level of community collaboration is the association of academic groups, small and medium enterprises, and pharmaceutical companies (eTOX consortium, see http://www.etoxproject.eu/) to enable the mining of pharmaceutical industry legacy toxicoloy reports in the search of information to complement in silico toxicology predictions and assisting in off-target detection efforts.

Acknowledgements

We would like to thank *eTOX* (Grant Agreement n°115002, funded by the Innovative Medicines Initiative Joint Undertaking), the Eurocancercoms 7th framework programme (FP7) project and the ISCIII combiomed network for funding. We thank *Obdulia Rabal* for useful feedback and discussions during the preparation of the manuscript.

References

[1] M. Krallinger, A. Valencia, L. Hirschman, *Genome Biol.* **2008**, *9 Suppl 2*, S8.
[2] J. Liu, M. Ghanim, L. Xue, C. D. Brown, I. Iossifov, C. Angeletti, S. Hua, N. Nègre, M. Ludwig, T. Stricker, H. A. Al-Ahmadie, M. Tretiakova, R. L. Camp, M. Perera-Alberto, D. L. Rimm, T. Xu, A. Rzhetsky, K. P. White, *Science* **2009**, *323(5918)*, 1218–1222.
[3] M. Maqungo, M. Kaur, S. K. Kwofie, A. Radovanovic, U. Schaefer, S. Schmeier, E. Oppon, A. Christoffels, V. B. Bajic, *Nucleic Acids Res.* **2011**, *39(Database issue)*, D980–985.
[4] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, A. Valencia, *Genome Biol.* **2008**, *9 Suppl 2*, S4.
[5] M. Krallinger, A. M. Rojas, A. Valencia, *Ann. N.Y. Acad. Sci.* **2009**, *1158*, 14–28.
[6] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, A. Valencia, *Genome Biol.* **2008**, *9 Suppl. 2*, S1.
[7] Q. Li, T. Cheng, Y. Wang, S. H. Bryant, *Drug Discov. Today* **2010**, *15(23–24)*, 1052–1057.
[8] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, D. S. Wishart, *Nucleic Acids Res.* **2011**, *39(Database issue)*, D1035–1041.
[9] A. Wilkinson, A. D. McNaught, *IUPAC Compendium of Chemical Terminology*, Vol. 2, IUPAC, **1997**.

[10] Chemical Service, *Index Guide*, Appendix IV, Chemical Substance Index Names, Chemical Service: Columbus, OH, **1997**.
[11] E. Garfield, *J. Chem. Doc.* **1962**, *2(3)*, 177 – 179.
[12] P. Murray-Rust, J. B. O. Mitchell, H. S. Rzepa, *BMC Bioinformatics* **2005**, *6*, 180.
[13] C. Kolárik, R. Klinger, C. M. Friedrich, M. Hofmann-Apitius, J. Fluck, *Workshop on Building and Evaluating Resources for Biomedical Text Mining*, in *Language Resources and Evaluation Conference*, 6th ed. **2008**.
[14] B. Müller, R. Klinger, Harsha, *Proc. 1st IRF Conf.*, Springer, Heidelberg, **2010**.
[15] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. van Mulligen, J. Kleinjans, J. A. Kors, *Bioinformatics* **2009**, *25(22)*, 2983 – 2991.
[16] A. T. McCray, O. Bodenreider, J. D. Malley, A. C. Browne, *Proc. AMIA Symp.* **2001**, 448 – 452.
[17] W. J. Rogers, A. R. Aronson, *Technical Report* **2008**, *http://skr.nlm.nih.gov/papers/references/filtering07.pdf*.
[18] M. Blume, *Proc. Intern. Conf. Intelligence Anal.* **2005**.
[19] Y. Garten, R. B. Altman, *BMC Bioinformatics* **2009**, *10 Suppl. 2*, S6.
[20] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, D. S. Wishart, *Nucleic Acids Res.* **2008**, *36 (Web Server issue)*, W399 – 405.
[21] C. R. Batchelor, P. T. Corbett, *Proc. 45th Ann. Meeting of the ACL on Interactive Poster and Demonstration Sessions*, **2007**, 45 – 48.
[22] K. M. Hettne, A. J. Williams, E. M. van Mulligen, J. Kleinjans, V. Tkachenko, J. A. Kors, *J. Cheminform.* **2010**, *2(1)*, 3.
[23] D. R. Heym, H. Siegel, M. C. Steensland, H. V. Vo, *J. Chem. Inf. Comput. Sci.* **1976**, *16(3)*, 171 – 176.
[24] N. Kemp, M. Lynch, *J. Chem. Inf. Comput. Sci.* **1998**, *38(4)*, 544 – 551.
[25] W. J. Wilbur, G. F. Hazard Jr, G. Divita, J. G. Mork, A. R. Aronson, A. C. Browne, *Proc. AMIA Symp.* **1999**, 176 – 180.
[26] A. Vasserman, *Proc. Student Res. Workshop at HLT-NAACL 2004*, Stroudsburg, PA, USA, Association for Computational Linguistics, **2004**, 7 – 12.
[27] J. D. Wren, *BMC Bioinformatics.* **2006**, *7 Suppl 2*, S3.
[28] L. R. Rabiner, *Proc. IEEE.* **1989**, *77(2)*, 257 – 286.
[29] T. Grego, P. Pęzik, F. M. Couto, D. Rebholz-Schuhmann, *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, Vol. 5518, (Eds: S. Omatu, M. P. Rocha, J. Bravo, F. Fernández, E. Corchado, A. Bustillo, J. M. Corchado), Springer, Berlin **2009**, 942 – 949.
[30] J. D. Lafferty, A. McCallum, F. C. N. Pereira, *Proc. 18th Int. Conf. Machine Learning*, **2001**, 282 – 289.
[31] C. M. Friedrich, T. Revillion, M. Hofmann, J. Fluck, *Proc. 2nd Int. Symp. Semantic Mining in Biomedicine (SMBM 2006)*, Vol. 7, **2006**, 85 – 89.
[32] R. Klinger, C. Kolárik, J. Fluck, M. Hofmann-Apitius, C. M. Friedrich, *Bioinformatics* **2008**, *24 (13)*, i268 – 276.
[33] B. Sun, Q. Tan, P. Mitra, C. L. Giles, *Proc. 16th Intern. Conf. World Wide Web*, ACM, New York, **2007**, 251 – 260.
[34] A. Mccallum, *MALLET: A Machine Learning for Language Toolkit*, (http://mallet.cs.umass.edu), **2002**.
[35] T. Kudo, *CRF++: Yet another CRF toolkit*, (http://crfpp.sourceforge.net/), **2007**.
[36] P. Corbett, A. Copestake, *BMC Bioinformatics* **2008**, *9 Suppl 11*, S4.
[37] B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, X. Wang, *Pac. Symp. Biocomput.* **2008**, 556 – 567.
[38] K. B. Cohen, L. Hunter, *Artificial Intelligence Methods And Tools For Systems Biology*, Vol. 5 (Eds: W. Dubitzky, F. Azuaje), Springer Netherlands, Dordrecht, **2004**, 147 – 173.
[39] C. Blaschke, A. Valencia, *Intelligent Systems, IEEE.* **2002**, *17(2)*, 14 – 20.
[40] A. R. Aronson, *Proc. AMIA Symp.* **2001**, 17 – 21.
[41] I. Segura-Bedmar, P. Martínez, M. Segura-Bedmar, *Drug Discov. Today.* **2008**, *13(17–18)*, 816 – 823.
[42] I. Segura-Bedmar, M. Crespo, C. de Pablo-Sánchez, P. Martínez, *BMC Bioinformatics* **2010**, *11 Suppl 2*, S1.
[43] M. Narayanaswamy, K. E. Ravikumar, K. Vijay-Shanker, *Pac. Symp. Biocomput.* **2003**, 427 – 438.
[44] F. Leitner, A. Chatr-aryamontri, S. A. Mardis, A. Ceol, M. Krallinger, L. Licata, L. Hirschman, G. Cesareni, A. Valencia, *Nat. Biotechnol.* **2010**, *28(9)*, 897 – 899.
[45] A. Pirkola, *Extracting variant forms of chemical names for information retrieval*, (http://informationr.net/ir/13–3/paper347.html), **2008**.
[46] G. A. Eller, *Molecules* **2006**, *11(11)*, 915 – 928.
[47] D. I. Cooke-Fox, G. H. Kirby, J. D. Rayner, *J. Chem. Inf. Model.* **1989**, *29 (2)*, 101 – 105.
[48] D. I. Cooke-Fox, G. H. Kirby, J. D. Rayner, *J. Chem. Inf. Model.* **1989**, *29 (2)*, 106 – 112.
[49] D. I. Cooke-Fox, G. H. Kirby, J. D. Rayner, *J. Chem. Inf. Model.* **1989**, *29 (2)*, 112 – 118.
[50] S. Anstein, G. Kremer, U. Reyle, *Proc. 5th Language Resources and Evaluation Conference*, **2006**, 1095 – 1098.
[51] J. Brecher, *J. Chem. Inf. Comp. Sci.* **1999**, *39(6)*, 943 – 950.
[52] A. T. Valko, A. P. Johnson, *J. Chem. Inf. Model.* **2009**, *49(4)*, 780 – 787.
[53] J. Park, G. R. Rosania, K. A. Shedden, M. Nguyen, N. Lyu, K. Saitou, *Chem. Cent. J.* **2009**, *3*, 4.
[54] J. R. McDaniel, J. R. Balmuth, *J. Chem. Inf. Comp. Sci.* **1992**, *32(4)*, 373 – 378.
[55] M.-E. Algorri, M. Zimmermann, C. M. Friedrich, S. Akle, M. Hofmann-Apitius, *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2007**, *2007*, 4609 – 4612.
[56] I. V. Filippov, M. C. Nicklaus, *J. Chem. Inf. Model.* **2009**, *49(3)*, 740 – 743.
[57] E. G. Brown, L. Wood, S. Wood, *Drug Saf.* **1999**, *20(2)*, 109 – 117.
[58] O. Bodenreider, *Nucleic Acids Res.* **2004**, *32(Database issue)*, D267 – 270.
[59] B. Settles, *Bioinformatics* **2005**, *21(14)*, 3191 – 3192.
[60] R. Leaman, G. Gonzalez, *Pac. Symp. Biocomput.* **2008**, 652 – 663.
[61] S. Ha, Y.-J. Seo, M.-S. Kwon, B.-H. Chang, C.-K. Han, J.-H. Yoon, *Bioinformatics* **2008**, *24(11)*, 1413 – 1415.
[62] S. Zhu, Y. Okuno, G. Tsujimoto, H. Mamitsuka, *Bioinformatics* **2005**, *21 Suppl 2*, ii245 – 251.
[63] N. C. Baker, B. M. Hemminger, *J. Biomed. Inform.* **2010**, *43(4)*, 510 – 519.
[64] R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg, W. Alkema, *PLoS Comput. Biol.* **2010**, *6(9)*.
[65] Y. Tsuruoka, J. Tsujii, S. Ananiadou, *Bioinformatics* **2008**, *24 (21)*, 2559 – 2560.
[66] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, P. Stoehr, *Bioinformatics* **2007**, *23(2)*, e237 – 244.
[67] R. Hoffmann, A. Valencia, *Bioinformatics* **2005**, *21 Suppl 2*, ii252 – 258.
[68] S. J. Campbell, A. Gaulton, J. Marshall, D. Bichko, S. Martin, C. Brouwer, L. Harland, *Drug Discov. Today.* **2010**, *15(1–2)*, 3 – 15.
[69] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, P. Bork, *Nucleic Acids Res.* **2008**, *36 (Database issue)*, D684 – 688.

[70] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krüger, B. Snel, P. Bork, *Nucleic Acids Res.* **2007**, *35 (Database issue)*, D358 – 362.
[71] C. Nobata, P. D. Dobson, S. A. Iqbal, P. Mendes, J. Tsujii, D. B. Kell, S. Ananiadou, *Metabolomics* **2010**, *7(1)*, 94 – 101.
[72] L. Tari, S. Anwar, S. Liang, J. Hakenberg, C. Baral, *Pac. Symp. Biocomput.* **2010**, 465 – 476.
[73] C. Blaschke, M. A. Andrade, C. Ouzounis, A. Valencia, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1999**, 60 – 67.
[74] K. Fundel, R. Küffner, R. Zimmer, *Bioinformatics* **2007**, *23(3)*, 365 – 371.
[75] J. Saric, L. J. Jensen, R. Ouzounova, I. Rojas, P. Bork, *Bioinformatics* **2006**, *22(6)*, 645 – 650.
[76] A. Coulet, N. H. Shah, Y. Garten, M. Musen, R. B. Altman, *J. Biomed. Inform.* **2010**, *43 (6)*, 1009 – 1019.
[77] D. Jiao, D. J. Wild, *J. Chem. Inf. Model.* **2009**, *49(2)*, 263 – 269.
[78] C. Feng, F. Yamashita, M. Hashida, *J. Chem. Inf. Model.* **2007**, *47(6)*, 2449 – 2455.
[79] Z. Wang, S. Kim, S. K. Quinney, Y. Guo, S. D. Hall, L. M. Rocha, L. Li, *J. Biomed. Inform.* **2009**, *42(4)*, 726 – 735.
[80] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, P. Bork, *Science* **2008**, *321(5886)*, 263 – 266.
[81] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, P. Bork, *Mol. Syst. Biol.* **2010**, *6*, 343.
[82] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, *Proc. 2010 Workshop on Biomedical Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, **2010**, 117 – 125.
[83] A. Blenkinsopp, P. Wilkie, M. Wang, P. A. Routledge, *Br. J. Clin. Pharmacol.* **2007**, *63(2)*, 148 – 156.
[84] N. Jarernsiripornkul, W. Kakaew, W. Loalukkana, J. Krska, *Pharmacoepidemiol. Drug Safety* **2009**, *18 (3)*, 240 – 245.
[85] D. Fourches, J. C. Barnes, N. C. Day, P. Bradley, J. Z. Reed, A. Tropsha, *Chem. Res. Toxicol.* **2010**, *23(1)*, 171 – 183.
[86] R. Frijters, B. Heupers, P. van Beek, M. Bouwhuis, R. van Schaik, J. de Vlieg, J. Polman, W. Alkema, *Nucleic Acids Res.* **2008**, *36(Web Server issue)*, W406 – 410.
[87] K. J. Gaulton, K. L. Mohlke, T. J. Vision, *Bioinformatics* **2007**, *23(9)*, 1132 – 1140.
[88] L.-C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts, Y. Moreau, *Nucleic Acids Res.* **2008**, *36(Web Server issue)*, W377 – 384.
[89] R. Saetre, K. Yoshida, M. Miwa, T. Matsuzaki, Y. Kano, J. Tsujii, *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2010**, *7(3)*, 442 – 453.
[90] C. Knox, S. Shrivastava, P. Stothard, R. Eisner, D. S. Wishart, *Pac. Symp. Biocomput.* **2007**, 145 – 156.
[91] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, I. Forsythe, *Nucleic Acids Res.* **2009**, *37(Database issue)*, D603 – 610.
[92] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, B. Ø. Palsson, *Proc. Natl. Acad. Sci. USA* **2007**, *104(6)*, 1777 – 1782.
[93] I. Rojas, M. Golebiewski, R. Kania, O. Krebs, S. Mir, A. Weidemann, U. Wittig, *In Silico Biol.* **2007**, *7(2 Suppl)*, S37 – 44.
[94] J. Hakenberg, S. Schmeier, A. Kowald, E. Klipp, U. Leser, *OMICS* **2004**, *8(2)*, 131 – 152.
[95] S. Heinen, B. Thielen, D. Schomburg, *BMC Bioinformatics* **2010**, *11*, 375.
[96] T. C. Wiegers, A. P. Davis, K. B. Cohen, L. Hirschman, C. J. Mattingly, *BMC Bioinformatics* **2009**, *10*, 326.
[97] D. L. Rubin, C. F. Thorn, T. E. Klein, R. B. Altman, *J. Am. Med. Inform. Assoc.* **2005**, *12(2)*, 121 – 129.
[98] T. D. Griffin, S. K. Boyer, I. G. Councill, *Advances in Computational Biology*, Vol. 680, (Ed: H. R. Arabnia), Springer, New York, **2010**, 737 – 744.
[99] E. L. Willighagen, N. M. O'Boyle, H. Gopalakrishnan, D. Jiao, R. Guha, C. Steinbeck, D. J. Wild, *BMC Bioinformatics* **2007**, *8*, 487.
[100] E. Pafilis, S. I. O'Donoghue, L. J. Jensen, H. Horn, M. Kuhn, N. P. Brown, R. Schneider, *Nat. Biotechnol.* **2009**, *27(6)*, 508 – 510.
[101] M. Vazquez, P. Carmona-Saez, R. Nogales-Cadenas, M. Chagoyen, F. Tirado, J. M. Carazo, A. Pascual-Montano, *Nucleic Acids Res.* **2009**, *37(Web Server issue)*, W153 – 159.
[102] T. E. Klein, J. T. Chang, M. K. Cho, K. L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D. E. Oliver, *Pharmacogen. J.* **2001**, *1 (3)*, 167 – 170.
[103] Y. Garten, N. P. Tatonetti, R. B. Altman, *Pac. Symp. Biocomput.* **2010**, 305 – 314.
[104] Y. Garten, R. B. Altman, *Pharmacogenomics* **2010**, *11 (4)*, 515 – 518.
[105] H. Shatkay, R. Feldman, *J. Comput. Biol.* **2003**, *10 (6)*, 821 – 855.
[106] W. J. Wilbur, L. Coffee, *Inform. Process. Manag.* **1994**, *30(2)*, 253 – 266.
[107] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, J. Ye, *Nucleic Acids Res.* **2011**, *39(Database issue)*, D38 – 51.
[108] S. Duda, C. Aliferis, R. Miller, A. Statnikov, K. Johnson, *AMIA Ann. Symp. Proc.* **2005**, 216 – 220.
[109] P. M. Roberts, W. S. Hayes, *Pac. Symp. Biocomput.* **2008**, 592 – 603.
[110] H. Gurulingappa, R. Klinger, Martin, *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Valetta, Malta **2010**.
[111] J. Wermter, K. Tomanek, U. Hahn, *Bioinformatics* **2009**, *25 (6)*, 815 – 821.
[112] M. Krallinger, F. Leitner, A. Valencia, *Methods Mol. Biol.* **2010**, *593*, 341 – 382.
[113] R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke, A. Valencia, *Sci. STKE* **2005**, *2005(283)*, pe21.
[114] A. Chang, M. Scheer, A. Grote, I. Schomburg, D. Schomburg, *Nucleic Acids Res.* **2009**, *37(Database issue)*, D588 – 592.
[115] Y. Yang, S. J. Adelstein, A. I. Kassis, *Drug Discov. Today* **2009**, *14(3–4)*, 147 – 154.
[116] D. Rebholz-Schuhmann, A. J. J. Yepes, E. M. Van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, U. Hahn, *J. Bioinform. Comput. Biol.* **2010**, *8(1)*, 163 – 179.
[117] R. Hoffmann, *Nat. Genet.* **2008**, *40(9)*, 1047 – 1051.

Received: January 10, 2011
Accepted: June 7, 2011
Published online: July 12, 2011